(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 3 720 149 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
07.10.2020 Bulletin 2020/41

(51) Int Cl.:
H04S 5/00 *(2006.01)*

(21) Application number: **19166572.8**

(22) Date of filing: **01.04.2019**

(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 MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME
Designated Validation States:
KH MA MD TN

(71) Applicant: **Nokia Technologies Oy**
02610 Espoo (FI)

(72) Inventors:
- **Laaksonen, Lasse Juhani**
  33210 Tampere (FI)
- **Luukkanen, Kalle**
  04300 Tuusula (FI)
- **Kallio, Juha**
  01610 Vantaa (FI)

(74) Representative: **Swindell & Pearson Limited**
48 Friar Gate
Derby DE1 1GY (GB)

(54) AN APPARATUS, METHOD, COMPUTER PROGRAM OR SYSTEM FOR RENDERING AUDIO DATA

(57) Certain examples of the present invention relate to rendering of audio data. Certain examples provide an apparatus 500 comprising means 501 for causing: receiving 401 first audio data 701 representative of first audio content 101; receiving 402 second audio data 702 representative of second audio content 102, wherein the second audio content 102 is derived from the first audio content 101; rendering 403 the first audio data 701 as a first virtual sound object 601 in a virtual sound scene 600 such that it is spatially rendered with a first virtual position 601o,601l within the virtual sound scene 600; rendering the second audio data 702 as a second virtual sound object 602 in the virtual sound scene 600 such that it is spatially rendered with a second virtual position 602o,602l within the virtual sound scene 600; and controlling the spatial rendering of the first and second virtual sound objects 601, 602 such that the first and second virtual positions 601o,601l;602o,602l differ.

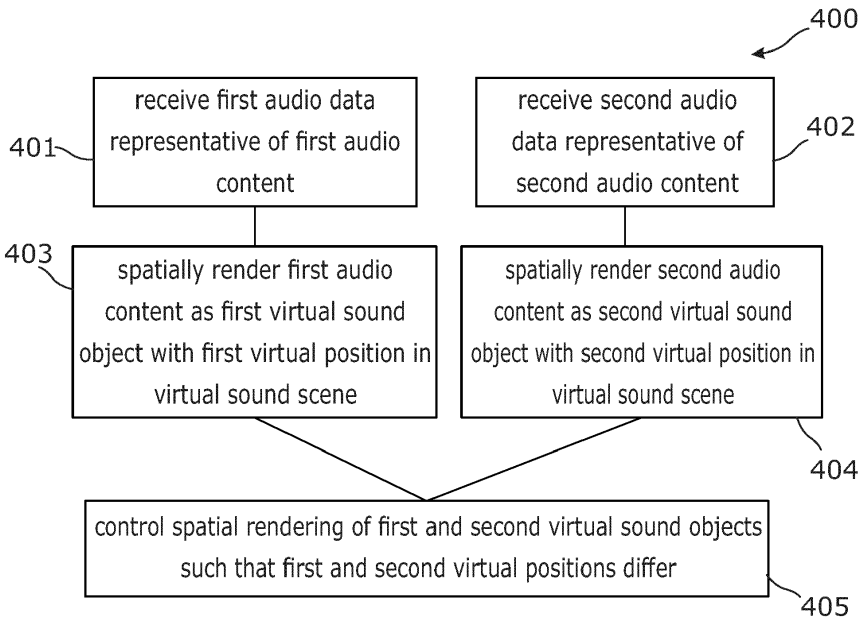

Fig. 4

EP 3 720 149 A1

Description

TECHNOLOGICAL FIELD

[0001] Examples of the present disclosure relate to apparatuses, methods, computer programs and systems for rendering audio data. Some examples, though without prejudice to the foregoing, relate to spatial rendering of speech and a real-time language translation of the same.

BACKGROUND

[0002] The rendering of audio in a virtual sound space in conventional mediated reality systems (such as, not least for example, rendering of speech in a first language and the rendering of a translation of the same in another language in an immersive aural environment of a virtual reality system) is not always optimal.

[0003] In a conventional mono voice call between a first user (who speaks a first language) and a second user (who speaks a second language), where a speech-to-speech audio translation service is provided, the first user talks to the second used in the first language. The first user's speech is transmitted to the second user who is able to hear the same. The first user's speech is also additionally routed to a data centre that carries out an automated language translation of the first user's speech to a second language. This translation of the speech is transmitted to the second user and also the first user both of whom are able to hear the same. This may then be followed by a response by the second user, with a translation of the same provided to both users. Such an experience differs from a traditional voice call in the sense that one user should not start to talk immediately following the other user's initial speech, but should await until after the translation of the same. In a traditional mono voice call, it is not possible to mix the original speech and the translated speech as this would make the speeches impossible to understand (known as the "interfering talker" problem for a mono signal/channel). This results in a "sequential voice experience" where the time consumed by such an exchange by the two users is doubled and most of the users' time is spent listening to both the speech in its original language and the translation.

[0004] This may be problematic in terms of user experience, because the natural flow of discussion between the users is constantly interrupted, and the voice call may take an unnecessarily long time (due to the time-doubling effect). This can be particularly ineffective and frustrating when the users "almost understand each other", albeit while still requiring some support for their understanding of the original speech from the translation of the same.

[0005] The listing or discussion of any prior-published document or any background in this specification should not necessarily be taken as an acknowledgement that the document or background is part of the state of the art or is common general knowledge. One or more aspects/examples of the present disclosure may or may not address one or more of the background issues.

BRIEF SUMMARY

[0006] According to various, but not necessarily all, examples of the disclosure there is provided an apparatus comprising means configured to cause:

receiving first audio data representative of first audio content;
receiving second audio data representative of second audio content, wherein the second audio content is derived from the first audio content;
rendering the first audio data as a first virtual sound object in a virtual sound scene such that it is spatially rendered with a first virtual position within the virtual sound scene;
rendering the second audio data as a second virtual sound object in the virtual sound scene such that it is spatially rendered with a second virtual position within the virtual sound scene; and controlling the spatial rendering of the first and second virtual sound objects such that the first and second virtual positions differ.

[0007] According to various, but not necessarily all, examples of the disclosure there is provided a method comprising:

receiving first audio data representative of first audio content;
receiving second audio data representative of second audio content, wherein the second audio content is derived from the first audio content;
rendering the first audio data as a first virtual sound object in a virtual sound scene such that it is spatially rendered with a first virtual position within the virtual sound scene;
rendering the second audio data as a second virtual sound object in the virtual sound scene such that it is spatially rendered with a second virtual position within the virtual sound scene; and controlling the spatial rendering of the first and second virtual sound objects such that the first and second virtual positions differ.

[0008] According to various, but not necessarily all, examples of the disclosure there is provided a chipset comprising processing circuitry configured to perform the above method.

[0009] According to various, but not necessarily all, examples of the disclosure there is provided a module, device and/or system comprising means for performing the above method.

[0010] According to various, but not necessarily all, examples of the disclosure there is provided computer program instructions for causing an apparatus to perform:

receiving first audio data representative of first audio content;
receiving second audio data representative of second audio content, wherein the second audio content is derived from the first audio content;
rendering the first audio data as a first virtual sound object in a virtual sound scene such that it is spatially rendered with a first virtual position within the virtual sound scene;
rendering the second audio data as a second virtual sound object in the virtual sound scene such that it is spatially rendered with a second virtual position within the virtual sound scene; and controlling the spatial rendering of the first and second virtual sound objects such that the first and second virtual positions differ.

[0011] According to various, but not necessarily all, examples of the disclosure there is provided an apparatus comprising:

at least one processor; and
at least one memory including computer program code;
the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus at least to perform:

receiving first audio data representative of first audio content;
receiving second audio data representative of second audio content, wherein the second audio content is derived from the first audio content;
rendering the first audio data as a first virtual sound object in a virtual sound scene such that it is spatially rendered with a first virtual position within the virtual sound scene;
rendering the second audio data as a second virtual sound object in the virtual sound scene such that it is spatially rendered with a second virtual position within the virtual sound scene; and controlling the spatial rendering of the first and second virtual sound objects such that the first and second virtual positions differ.

[0012] According to various, but not necessarily all, examples of the disclosure there is provided a non-transitory computer readable medium encoded with instructions that, when performed by at least one processor, causes at least the following to be performed:

receiving first audio data representative of first audio content;
receiving second audio data representative of second audio content, wherein the second audio content is derived from the first audio content;
rendering the first audio data as a first virtual sound object in a virtual sound scene such that it is spatially rendered with a first virtual position within the virtual sound scene;
rendering the second audio data as a second virtual sound object in the virtual sound scene such that it is spatially rendered with a second virtual position within the virtual sound scene; and controlling the spatial rendering of the first and second virtual sound objects such that the first and second virtual positions differ.

[0013] The following portion of this 'Brief Summary' section, describes various features that can be features of any of the examples described in the foregoing portion of the 'Brief Summary' section. The description of a function should additionally be considered to also disclose any means suitable for performing that function.

[0014] In some but not necessarily all examples, the first and second virtual sound objects may be simultaneously rendered in the virtual sound scene, and the spatial rendering of the first and second virtual sound objects may be controlled such that, at least whilst the first and second virtual sound objects are simultaneously spatially rendered, the first and second virtual positions differ.

[0015] In some but not necessarily all examples, the spatial rendering of the first and second virtual sound objects is controlled such that a first virtual direction of the spatially rendered first virtual sound object differs from the second virtual direction of the spatially rendered second virtual sound object.

[0016] In some but not necessarily all examples, the second audio content is a translation (for example a real time language translation) of the first audio content.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] For a better understanding of various examples of the present disclosure that are useful for understanding the detailed description and certain embodiments of the invention, reference will now be made by way of example only to the accompanying drawings in which:

FIGs. 1A and 1B schematically illustrate an example real space for use with examples of the subject matter described herein;
FIGs. 2A and 2B schematically illustrate an example virtual audio space for use with examples of the subject matter described herein;
FIGs. 3A and 3B schematically illustrate an example virtual visual space for use with examples of the subject matter described herein;
FIG. 4 schematically illustrates an example method of the subject matter described herein;
FIG. 5 schematically illustrates an example apparatus of the subject matter described herein;
FIG. 6 schematically illustrates an example implementation of the subject matter described herein;
FIG. 7 schematically illustrates a further example apparatus of the subject matter described herein;

FIG. 8 schematically illustrates an example system of the subject matter described herein;
FIGs. 9A and 9B schematically illustrate a further example implementation of the subject matter described herein;
FIGs. 10A and 10B schematically illustrate example timelines of the subject matter described herein;
FIGs. 11A-C schematically illustrate a further example implementation of the subject matter described herein;
FIGs. 12A and 12B schematically illustrate a further example implementation of the subject matter described herein;
FIG. 13A-D schematically illustrates a further example implementation of the subject matter described herein;
FIG. 14 schematically illustrates a further example implementation of the subject matter described herein; and
FIG. 15 schematically illustrates further example systems of the subject matter described herein.

[0018] The Figures are not necessarily to scale. Certain features and views of the figures can be shown schematically or exaggerated in scale in the interest of clarity and conciseness. For example, the dimensions of some elements in the figures can be exaggerated relative to other elements to aid explication. Similar reference numerals are used in the figures to designate similar features. For clarity, all reference numerals are not necessarily displayed in all figures.

DEFINITIONS

[0019]

"artificial environment" may be something that has been recorded or generated.
"virtual visual space" refers to a fully or partially *artificial environment* that may be viewed, which may be three-dimensional.
"virtual visual scene" refers to a representation of the *virtual visual* space viewed from a particular point-of-view (e.g. position) within the *virtual visual space.*
"virtual visual object" is a visible virtual object within a *virtual visual scene.*
"virtual sound space"/*"virtual audio space"* refers to a fully or partially *artificial environment* that may be listened to, which may be three-dimensional.
"virtual sound scene "/*"virtual audio scene"* refers to a representation of the *virtual sound space* listened to from a particular *point-of-view* (e.g. position) within the *virtual sound space*.
"virtual sound object" is an audible virtual object within a *virtual sound scene.*
"sound *object"* refers to a sound source that may be located within the sound space. A source sound object represents a sound source within the sound space (in contrast to a sound source associated with a virtual sound object in the virtual visual space). A recorded sound object may represent sounds recorded at a particular microphone or location. A rendered sound object represents sounds rendered from a particular location.
"sound space" refers to an arrangement of sound sources in a three-dimensional space. A *sound space* may be defined in relation to recording sounds (a recorded sound space) and in relation to rendering sounds (a rendered sound space).
"sound scene" refers to a representation of the *sound space* listened to from a particular *point-of-view (position)* within the *sound space*.
"virtual space" may mean: a *virtual* sound space, a *virtual visual space,* or a combination of a *virtual visual space* and corresponding *virtual sound space.* In some examples, the *virtual space* may extend horizontally up to 360° and may extend vertically up to 180°.
"virtual scene" may mean: a virtual visual scene, a virtual sound scene, or a combination of a virtual visual scene and corresponding virtual sound scene.
"virtual object'" is an object within a *virtual scene,* it may be an augmented virtual object (e.g. a computer-generated virtual object) or it may be an image of a real object in a *real space* that is live or recorded. It may be a *virtual sound object* and/or a *virtual visual object.*
"virtual position" is a position within a virtual space. It may be defined using a *virtual location* and/or a *virtual orientation.* It may be a movable *'point-of-view'* in *virtual visual space* and/or *virtual sound space.* The *virtual position* may be defined using a *virtual location* and/or a *virtual direction* with respect to a particular *point-of-view,* e.g. a *point-of-view* of a notional (virtual) listener and/or viewer.
"correspondence" or *"corresponding"* when used in relation to a *virtual sound space* and a *virtual visual* space means that the *virtual sound space* and *virtual visual space* are time and space aligned, that is they are the same space at the same time.
"correspondence" or "corresponding" when used in relation to a virtual sound scene and a virtual visual scene (or visual scene) means that the virtual sound space and virtual visual space (or visual scene) are corresponding and a notional (virtual) listener whose point-of-view defines the virtual sound scene and a notional (virtual) viewer whose point-of-view defines the virtual visual scene (or visual scene) are at the same location and orientation, that is they have the same point-of-view (same virtual position, i.e. same location and orientation).
"real space" (or "physical space") refers to a real environment, outside of the virtual space, which may be three-dimensional.
"real scene" refers to a representation of the *real*

space from a particular *point-of-view* (position) within the real space.

"real visual scene" refers to a visual representation of the real space viewed from a particular real point-of-view (position) within the real space.

"mediated reality", refers to a user experiencing, for example visually and/or aurally, a fully or partially artificial environment (a *virtual space*) as a *virtual scene* at least partially rendered by an apparatus to a user. The *virtual* scene is determined by a *point-of-view* (virtual position) within the *virtual space.*

"rendering or presenting the *virtual* scene means providing (e.g. audibly presenting) a virtual sound scene in a form that can be perceived by the *user, and/or* providing (e.g. displaying) a virtual visual scene in a form that can be perceived by the *user.*

"augmented reality" refers to a form of *mediated reality* in which a *user* experiences a partially artificial environment (a *virtual space*) as a *virtual scene* comprising a *real scene,* for example a real *visual scene* and *real sound scene,* of a physical real environment (*real space*) supplemented by one or more visual or audio elements rendered by an apparatus to a user. The term augmented reality implies a mixed reality or hybrid reality and does not necessarily imply the degree of virtuality (vs reality) or the degree of mediality. Augmented reality (AR) can generally be understood as providing a user with additional information or artificially generated items or content that is at least significantly overlaid upon the user's current real-world environment stimuli. In some such cases, the augmented content may at least partly replace a real-world content for the user. Additional information or content may usually be audible visual and/or haptic. Similarly to *virtual reality,* but potentially in more applications and use cases, AR may have visual-only or audio-only presentation. For example, user may move about a city and receive audio guidance relating to, e.g., navigation, location-based advertisements, and any other location-based information. Mixed reality (MR) is often considered as a more advanced form of AR where at least some virtual elements are inserted into the physical scene such that they provide the illusion that these elements are part of the real scene and behave accordingly. For audio content, or indeed audio-only use cases, many applications of AR and MR may appear difficult for the user to tell from one another. However, the difference is not only for visual content but it may be relevant also for audio. For example, MR audio rendering may take into account a local room reverberation, e.g., while AR audio rendering may not.

"virtual reality" refers to a form of *mediated reality* in which a user experiences a fully artificial environment (a *virtual visual space and/or virtual sound space*) as a *virtual scene* rendered by an apparatus to a user. Virtual Reality (VR) can generally be understood as a rendered version of visual and audio scene. The rendering is typically designed to closely mimic the visual and audio sensory stimuli of the real world in order to provide a user a natural experience that is at least significantly consistent with their movement within virtual scene according to the limits defined by the content and/or application. VR in most cases, but not necessarily all cases, requires a user to wear a head mounted display (HMD), to completely replace the user's field of view with a simulated visual presentation, and to wear headphones, to provide the user the simulated audio content similarly completely replacing the sound scene of the physical space. Some form of head tracking and general motion tracking of the user consuming VR content is typically also necessary. This allows the simulated visual and audio presentation to be updated in order to ensure that, from the user's perspective, various scene components such as items and sound sources remain consistent with the user's movements. Additional means to interact with the virtual reality simulation, such as controls or other user interfaces (UI) may be provided but are not strictly necessary for providing the experience. VR can in some use cases be visual-only or audio-only virtual reality. For example, an audio-only VR experience may relate to a new type of music listening or any other audio experience.

"extended reality" (XR) is a term that refers to all real-and-virtual combined realities/environments and human-machine interactions generated by digital technology and various wearables. It includes representative forms such as augmented reality (AR), augmented virtuality (AV), mixed reality (MR), and virtual reality (VR) and any relevant interpolations.

"virtual content" is content, additional to *real content* from a real scene, if any, that enables *mediated reality* by, for example, providing one or more augmented *virtual objects.*

"mediated reality content" is virtual content which enables a *user* to experience, for example visually and/or aurally, a fully or partially artificial environment (a *virtual space*) as a *virtual scene. Mediated reality content* could include interactive content such as a video game or non-interactive content such as motion video.

"augmented reality content" is a form of mediated reality content which enables a user to experience, for example visually and/or aurally, a partially artificial environment (a *virtual space*) as a *virtual scene. Augmented reality* content could include interactive content such as a video game or non-interactive content such as motion video.

"virtual reality content" is a form of mediated reality content which enables a user to experience, for example visually and/or aurally, a fully artificial environment (a *virtual space*) as a *virtual scene. Virtual reality content* could include interactive content such as a video game or non-interactive content such as

motion video.

"perspective-mediated" as applied to *mediated reality, augmented reality or virtual reality* means that *user* actions determine the *point-of-view* (virtual position) within the virtual space, changing the *virtual scene.*

"first person perspective-mediated" as applied to mediated reality, augmented reality or virtual reality means perspective-mediated with the additional constraint that the user's real point-of-view (location and/or orientation) determines the point-of-view (virtual position) within the virtual space of a virtual user.

"third person perspective-mediated" as applied to mediated reality, augmented reality or virtual reality means perspective-mediated with the additional constraint that the user's real point-of-view does not determine the point-of-view (virtual position) within the virtual space.

"user interactive" as applied to mediated reality, augmented reality or virtual reality means that user actions at least partially determine what happens within the virtual space.

"displaying" means providing in a form that is perceived visually (viewed) by the *user.*

"rendering" means providing in a form that is perceived by the *user,* e.g. visually (viewed) or aurally (listened to) by the user.

"virtual user" refers to a user within the virtual space, e.g. a user immersed in a mediated/virtual/augmented reality. *Virtual* user defines the *point-of-view* (virtual position - location and/or orientation) in *virtual space* used to generate a *perspective-mediated* sound *scene and/or visual scene.* A virtual user may be a notional listener and/or a notional viewer.

"notional listener" defines the *point-of-view* (virtual position - location and/or orientation) in *virtual* space used to generate a *perspective-mediated sound scene,* irrespective of whether or not a user is actually listening.

"notional viewer" defines the *point-of-view* (virtual position - location and/or orientation) in *virtual space* used to generate a *perspective-mediated visual scene,* irrespective of whether or not a *user* is actually viewing.

"three degrees of *freedom (3DoF)"* describes *mediated reality* where the *virtual position* is determined by orientation only (e.g. the three degrees of three-dimensional orientation). An example of three degrees of three-dimensional orientation is pitch, roll and yaw (i.e. just 3DoF rotational movement). In relation to *first person perspective-mediated reality 3DoF,* only the *user's* orientation determines the *virtual position.*

"six degrees of freedom (6DoF)" describes mediated reality where the *virtual position* is determined by both orientation (e.g. the three degrees of three-dimensional orientation) and location (e.g. the three degrees of three-dimensional location), i.e. 3DoF rotational and 3DoF translational movement. An example of three degrees of three-dimensional orientation is pitch, roll and yaw. An example of three degrees of three-dimensional location is a three-dimensional coordinate in a Euclidian space spanned by orthogonal axes such as left to right (**x**), front to back (**y**) and down to up (**z**) axes. In relation to *first person perspective-mediated reality 6DoF,* both the *user's* orientation and the user's location in the *real space* determine the *virtual position.* In relation to *third person perspective-mediated reality* 6DoF, the *user's* location in the *real space* does not determine the *virtual position.* The *user's* orientation in the *real space* may or may not determine the *virtual position.*

"three degrees of *freedom 'plus' (3DoF+)"* describes an example of six degrees of freedom where a change in location (e.g. the three degrees of three-dimensional location) is a change in location relative to the user that can arise from a postural change of a user's head and/or body and does not involve a translation of the user through real space by, for example, walking.

"spatial rendering" refers to a rendering technique that renders content as an object at a particular three dimensional position within a three dimensional space.

"spatial audio" is the rendering of a sound scene. "First person perspective spatial audio" or "immersive audio" is spatial audio where the user's point-of-view determines the sound scene (or "sub-sound scene"/"sub-audio scene") so that audio content selected by a current point-of-view of the user is rendered to the user. In spatial audio rendering, audio may be rendered as a sound object that has a three-dimensional position in a three-dimensional sound space. Various different spatial audio rendering techniques are available. For example, a head-related transfer function may be used for spatial audio rendering in a binaural format or amplitude panning may be used for spatial audio rendering using loudspeakers. It is possible to control not only the position of a sound object but it is also possible to control the spatial extent of a sound object by distributing the sound object across multiple different spatial channels that divide sound space into distinct sectors, such as sound scenes and sound sub-scenes.

"immersive audio" refers to the rendering of audio content to a user, wherein the audio content is selected in dependence on a current point-of-view of the user. The user therefore has the experience that they are immersed within a three-dimensional audio field/sound scene/audio scene, that may change as their point-of-view changes.

DETAILED DESCRIPTION

[0020] The Figures schematically illustrate an apparatus 500 comprising means 501 for causing:

receiving 401 first audio data 701 representative of first audio content 101;

receiving 402 second audio data 702 representative of second audio content 102, wherein the second audio content 102 is derived from the first audio content 101;

rendering 403 the first audio data 701 as a first virtual sound object 601 in a virtual sound scene 600 such that it is spatially rendered with a first virtual position 601o,601l within the virtual sound scene 600;

rendering the second audio data 702 as a second virtual sound object 602 in the virtual sound scene 600 such that it is spatially rendered with a second virtual position 602o,602l within the virtual sound scene 600; and

controlling the spatial rendering of the first and second virtual sound objects 601, 602 such that the first and second virtual positions 601o,601l;602o,602l differ.

[0021] For the purposes of illustration and not limitation, various, but not necessarily all, examples of the disclosure may provide the technical advantage of improved rendering of the first and second audio data that enables a user to distinguish and differentiate the first and second virtual sounds objects which thereby enhances the user's listening experience and better enables the user to perceive and focus on one of the first or second virtual sound objects and hence perceive the first or second audio content represented thereby. The control of the spatial rendering in examples of the disclosure takes advantage of the so-called "cocktail party effect" wherein, due to differing spatial placement of audio sources, a user is capable of concentrating on one of many audio sources regardless of their temporal overlap. Advantageously, this enables the overlapping/simultaneous/parallel rendering of the first and second audio data and may thus avoids/mitigate the issues of the simultaneous playback of more than one voice/speech in a conventional mono-audio/single channel communication/ mono voice call.

[0022] Figs. 1A, 2A and 3A illustrate an example of first person perspective mediated reality. In this context, mediated reality means the rendering of mediated reality for the purposes of achieving mediated reality for a remote user, for example augmented reality or virtual reality. It may or may not be user interactive. The mediated reality may support one or more of: 3DoF, 3DoF+ or 6DoF.

[0023] Figs. 1A, 2A and 3A illustrate, at a first time, each of: a real space 50, a virtual sound space 20 and a virtual visual space 60 respectively. There is correspondence between the virtual sound space 20 and the virtual visual space 60. A 'virtual space' may be defined as the virtual sound space 20 and/or the virtual visual space 60. In some examples, the virtual space may comprise just the virtual sound space 20. A user 51 in the real space 50 has a position defined by a (real world) location 52 and a (real world) orientation 53 (i.e. the user's real world point-of-view). The location 52 is a three-dimensional location and the orientation 53 is a three-dimensional orientation.

[0024] In an example of 3DoF mediated reality, an orientation 53 of the user 51 controls/determines a virtual orientation 73 of a virtual user 71 within a virtual space, e.g. the virtual visual space 60 and/or the virtual sound space 20. The virtual user 71 represents the user 51 within the virtual space. There is a correspondence between the orientation 53 and the virtual orientation 73 such that a change in the (real world) orientation 53 produces the same change in the virtual orientation 73. In 3DoF mediated reality, a change in the location 52 of the user 51 does not change the virtual location 72 or virtual orientation 73 of the virtual user 71.

[0025] The virtual orientation 73 of the virtual user 71, in combination with a virtual field of view 74 defines a virtual visual scene 75 of the virtual user 71 within the virtual visual space 60. The virtual visual scene 75 represents a virtual observable region within the virtual visual space 60 that the virtual user 71 can see. Such a 'virtual visual scene 75 for the virtual user 71' may correspond to a virtual visual 'sub-scene'. The virtual visual scene 75 may determine what visual content (and virtual visual spatial position of the same with respect to the virtual user's position) is rendered to the virtual user. In a similar way that the virtual visual scene 75 of the virtual user 71 may affect what visual content is rendered to the virtual user, a virtual sound scene 76 of the virtual user may affect what audio content (and virtual aural spatial position of the same with respect to the virtual user's position) is rendered to the virtual user.

[0026] The virtual orientation 73 of the virtual user 71, in combination with a virtual field of hearing (i.e. an audio equivalent/analogy to a visual field of view) may define a virtual sound scene (or audio scene) 76 of the virtual user 71 within the virtual sound space (or virtual audio space) 20. The virtual sound scene 76 represents a virtual audible region within the virtual sound space 20 that the virtual user 71 can hear. Such a 'virtual sound scene 76 for the virtual user 71' may correspond to a virtual audio 'sub-scene'. The virtual sound scene 76 may determine what audio content (and virtual spatial position/orientation of the same) is rendered to the virtual user.

[0027] A virtual visual scene 75 is that part of the virtual visual space 60 that is rendered/visually displayed to a user. A virtual sound scene 76 is that part of the virtual sound space 20 that is rendered/audibly output to a user. The virtual sound space 20 and the virtual visual space 60 correspond in that a position within the virtual sound space 20 has an equivalent position within the virtual visual space 60. In 3D0F mediated reality, a change in the location 52 of the user 51 does not change the virtual location 72 or virtual orientation 73 of the virtual user 71.

[0028] In the example of 6DoF mediated reality, the situation is as described for 3DoF and in addition it is possible to change the rendered virtual sound scene 76 and the displayed virtual visual scene 75 by movement

of a location 52 of the user 51. For example, there may be a mapping between the location 52 of the user 51 and the virtual location 72 of the virtual user 71. A change in the location 52 of the user 51 produces a corresponding change in the virtual location 72 of the virtual user 71. A change in the virtual location 72 of the virtual user 71 changes the rendered virtual sound scene 76 and also changes the rendered virtual visual scene 75.

[0029] This may be appreciated from Figs. 1B, 2B and 3B which illustrate the consequences of a change in position, i.e. a change in location 52 and orientation 53, of the user 51 on respectively the rendered virtual sound scene 76 (Fig. 2B) and the rendered virtual visual scene 75 (Fig. 3B).

[0030] Immersive or spatial audio (for 3DoF/3DoF+/6DoF) may consist, e.g., of a channel-based bed and audio objects, metadata-assisted spatial audio (MASA) and audio objects, first-order or higher-order ambisonics (FOA/HOA) and audio objects, any combination of these such as audio objects only, or any equivalent spatial audio representation.

[0031] FIG 4 schematically illustrates a flow chart of a method 400 according to an example of the present disclosure. The component blocks of FIG. 4 are functional and the functions described may or may not be performed by a single physical entity (such as an apparatus is described with reference to FIG. 5).

[0032] In block 401, first audio data representative of first audio content is received.

[0033] In block 402, second audio data representative of second audio content, wherein the second audio content is derived from the first audio content is received.

[0034] In block 403, the first audio data is rendered as a first virtual sound object in a virtual sound scene such that it is spatially rendered with a first virtual position within the virtual sound scene.

[0035] In block 404, the second audio data is rendered as a second virtual sound object in the virtual sound scene such that it is spatially rendered with a second virtual position within the virtual sound scene.

[0036] In block 405, the spatial rendering of the first and second virtual sound objects are controlled such that the first and second virtual positions differ.

[0037] The received audio data may be spatial audio or audio with associated metadata representative of virtual position at which the audio content is to be rendered. Alternatively, the received audio may not comprise or be preassociated with a virtual position at which the audio content is to be rendered (particularly for the second content which may be newly generated audio content comprising a machine translation of the first audio content). Where the audio data is spatial audio data, the initial virtual position may be determined from the audio data itself. Where the audio data comprises, or is associated with metadata representative of a virtual position at which the virtual audio content is to be rendered, the initial virtual position may be determined from the metadata. Where the audio data does not comprise, nor is associated with metadata representative of a virtual position at which the virtual audio content is to be rendered, the initial virtual position may be predetermined or determined from a user setting/user preference, e.g. the renderer may be configured such that the first virtual audio content is virtually positioned at an orientation/direction of x^0. The initial capture of the audio need not be spatial. The initial spatial placement can be based on the capture, a setting by a transmitting user, a spatialization by a service/system such as a conferencing system, or a setting by a receiving user. For example, there could be a "preferred position" for the receiving user, in which case the first virtual position of the first virtual sound object may be positioned there.

[0038] In some examples, the second audio content comprises a modified version of the first audio content. In some examples, the first audio content comprises speech (e.g. from a user in a voice call with another user) in a first language and the second audio content comprises a translation of the speech into a second language. In some examples, the second audio content is a real time language translation of the first audio content.

[0039] In some examples, the spatial rendering of first and second virtual sound objects may occur sequentially or overlapping in time such that they are rendered simultaneously (albeit possibly with a delay/lag, e.g., due to the lookahead required by the language translation).

[0040] In some examples, the first and second virtual sound objects are simultaneously rendered in the virtual sound scene, and the spatial rendering of the first and second virtual sound objects is controlled such that, at least whilst the first and second virtual sound objects are simultaneously spatially rendered, the first and second virtual positions differ. For example, each of the first and second virtual sound objects has a finite duration and, following completion of the spatial rendering of the first virtual sound object, the spatial rendering of the second virtual sound object may move, e.g., so as to correspond to the virtual position of where the first virtual sound object was spatially rendered or to the virtual position corresponding to the "preferred position" for the receiving user.

[0041] FIG. 5 schematically illustrates an apparatus 500 which is configured to receive the first and second audio data 701,702 representative of first and second audio content 101,102 respectively. The apparatus is configured to render the first audio data 701 as a first virtual sound object 601 (which is itself also representative of the first audio content 101) such that it is spatially rendered in a virtual sound scene with a first virtual position (as is schematically illustrated in FIG. 6). The apparatus is also configured to render the second audio data 702 as a second virtual sound object 602 (which is itself also representative of the second audio content 102) such that it is spatially rendered in the virtual sound scene with a second virtual position different to the first virtual position (as is schematically illustrated in FIG. 6).

[0042] FIG. 6 schematically illustrates a virtual sound scene 600. The virtual sound scene 600 is a represen-

tation of a virtual sound space as listened to from the point-of-view of a second user 802. The user's point of view corresponds to the user's position 802p - which comprises the user's location 8021 and/or the user's orientation/direction 802o. The first virtual sound object 601 is spatially rendered so as to have a perceived first virtual position 601p (which comprises the first virtual sound object's location 6011 and/or virtual orientation/direction 601o). The second virtual sound object 602 is spatially rendered so as to have a perceived second virtual position 602p (which comprises the second virtual sound object's location 6021 and/or virtual orientation/direction 602o). The second virtual position 602p is controlled such that it has a different virtual position to the first virtual position 601 p. FIG.6 schematically illustrates the virtual sound scene from a plan/elevation viewpoint, i.e. such that the illustrated separation angle θ between the first and second virtual positions relates to an azimuthal angle relative to the second user/listener 802.

[0043] In some examples, the spatial rendering of the first and second virtual sound objects 601, 602 is controlled such that the first virtual orientation 601o of the spatially rendered first virtual sound object 601 differs from the second virtual orientation 602o of the spatially rendered second virtual sound object 602.

[0044] In some examples, the spatial rendering of the first and second virtual sound objects 601,602 is controlled such that the first virtual orientation/direction 601o differs from the second virtual orientation/direction 602o by an azimuthal angle of greater than: 15°, 30°, 45°, 60°, 75°, 90°, 105°, 120°, 135°, 150°, or 165°. In some examples, the first and second virtual positions 601p,602p are controlled so as to be spatially "maximally" separated. Such a maximal separation may correspond to a maximal directional separation of the first and second virtual positions or alternatively a maximal directional separation whilst ensuring that both the first and second virtual sound objects 601,602 remain within the same hemisphere of the user's point of view (e.g. the hemisphere in front of the user).

[0045] In some examples, the spatial rendering of the first virtual sound object 601 is controlled such that, during a first time period between a start of the spatial rendering of the first virtual sound object 601 and a start of the rendering of the second virtual sound object 602, the first virtual position 601p is moved. In such a manner, the first virtual sound object 601 may be rendered in an initial virtual position 601p upon commencement of its rendering and the virtual position 601p may then move so as to "make room" for the commencement of the rendering of the second virtual sound object 602. For example, the first virtual sound object 601 may be rendered at a virtual position 601p directly in front of the user's point of view, the virtual position 601 p may then move to one side such that the second virtual sound object 602 may be rendered directly in front of the user's point of view (thereby maintaining a difference in first and second virtual positions 601p,602p). Alternatively, the virtual position 601p of the first virtual sound object 601 may be moved so that its virtual position mirrors that of the virtual position 602p of the second virtual sound object 602 relative to the user 802, i.e. such that the first and second virtual sound positions 601p,602p "mirror" each other relative to the user 802.

[0046] A determination may be made as to a first time period/delay period between a start of the rendering of the first virtual sound object 601 and a start of the rendering of the second virtual sound object 602. This may be done by receiving signalling providing an indication of such timings from the source(s) of the first and second audio data 701,702. The virtual position 601p at which the first virtual sound object 601 is spatially rendered may then be controlled so as to move during the determined first time period and prior to starting the spatial rendering of the second virtual sound object 602.

[0047] The control of the spatial rendering of the first virtual sound object 601 may be based on an upcoming commencement of the rendering of the second virtual sound object 602. In some examples, the spatial rendering of the first virtual sound object 601, not least its first virtual position 601p, is connected to and may be adapted based on the state of spatial rendering of the second virtual sound object 602 (i.e. if it has yet started) or adapted automatically based on signal activity.

[0048] In some examples. the first audio content 101 of the first virtual sound object 601 comprises speech/voice/talking in a first language, and the second audio content 102 of the second virtual sound object 602 relates to a real time language translation of the speech into a second language, such that the second audio content 102 comprises speech/voice/talking in the second language. Typically, the receipt and rendering of the translation would be behind that of the original speech (e.g. not least due to the processing of the original speech to generate the translated speech). In some examples, the transmission and/or rendering of the original speech can be delayed until the translation is ready.

[0049] In some examples, the original speech can initially be rendered in one virtual position (for example an optimal position such as substantially directly in front of the user) and it can then begin to "make room" for the translation upon the start of the rendering of the translation (so that it instead can be rendered in the optimal position).

[0050] Advantageously, this may enable the second user 802 to distinguish and differentiate the first and second virtual sounds objects 601,602 which thereby enhances the user's listening experience and better enables the user to perceive and focus on one of the first or second virtual sound objects 601, 602 and hence perceive the first or second audio content 101,102 represented thereby.

[0051] In some examples, the spatial rendering of the second virtual sound object 602 is controlled such that, during a second time period between an end of the spatial rendering of the first virtual sound object 601, the second

virtual position 602p is moved. For example, the first virtual sound object 601 may be rendered at a virtual position 601 p directly in front of the user's point of view. Whilst the first and second virtual sound objects 601, 602 are being simultaneously rendered, the second virtual sound object 602 may be rendered at a differing virtual position 602p, e.g. to the side of the user's point of view. Upon completion of the rendering of the first virtual sound object 601, the virtual position 602p of the second virtual sound object 602 may be moved to the (former) virtual position of (the now ceased rendered) first virtual sound object 601, i.e. directly in front of the user's point of view. Such movement may be smooth/gradual, i.e., such that the second virtual sound object 602 does not seem to jut disappear at one position and appear at another position. Advantageously, this may enhance the user's listening experience and better enable the user to perceive and focus on the second virtual sound object 602 and hence perceive the second audio content 102 represented thereby.

[0052] A determination may be made as to a second time period/delay period between an end of the rendering of the first virtual sound object 601 and an end of the rendering of the second virtual sound object 602. This may be done by receiving signalling providing an indication of such timings from the source(s) of the first and second audio data 701,702. (Such signalling may be provided in certain use cases for example where the second audio is computer generated, such as a real-time language translation). The virtual position 602p at which the second virtual sound object 602 is spatially rendered may then be controlled so as to move during the determined second time period where either the speed of the movement or at least one position during the movement may indicate the time left for presentation of the second virtual sound object 602. For example, there could be user-set position/direction for the second virtual sound object 602 when there is a predetermine amount of time, e.g. 5 seconds. left in the presentation of the second virtual sound object 602.

[0053] The control of the spatial rendering of the second virtual sound object 602 may be based on the completion of the rendering of the first virtual sound object 601. In some examples, the spatial rendering of the second virtual sound object 602, not least its second virtual position 602p, is connected to and may be adapted based on the state of spatial rendering of the first virtual sound object 601 (i.e. if it is still active/ongoing or has completed) or adapted automatically based on signal activity.

[0054] In some examples, the apparatus 500 is configured to receive a user input to control the spatial rendering of one of the first or second virtual sound objects 601, 602. Responsive to receipt of the user input, the spatial rendering of the one of the first or second virtual sound objects 601,602 is changed. Furthermore, responsive to the user-controlled change of the spatial rendering of the one of the first or second virtual sound objects 601,602, the spatial rendering of the other of the second or first virtual sound object 602,601 is changed. In other words, the control of the spatial rendering of the first virtual sound object 601 may be based on the user-controlled changes to the spatial rendering of the second virtual sound object 602, and vice versa.

[0055] In some examples, the spatial rendering of the first and second virtual sound objects 601,602, not least their respective first and second virtual positions 601 p,602p are connected and may be adapted based on user input. For example, responsive to a user input to reduce the volume for one of the virtual sound objects 601,602, this may cause the automatic movement of a virtual position 601p,602p of one or both of the virtual sound objects 601,602 (e.g. move the virtual position of the virtual sound object with the reduced volume away from a central position of user's field of hearing, and move the virtual position of other virtual sound object towards the central position of user's field of hearing). In some examples, responsive to a user input to increase the volume for one of the virtual sound objects, this may cause the automatic decreasing of the volume for the other virtual sound object. This may be, e.g., because it is signalled that a pair of sound objects is connected and considered alternatives with concurrent playback. For example, one virtual sound object is in a first language and the second virtual sound object is a translation in a second language corresponding to the first language voice signal.

[0056] The change in spatial rendering may comprise: one or more of:

changing a virtual position 601p,602p of one of the rendered first or
second virtual sound objects 601,602;
changing a virtual orientation/direction 601o,602o of one of the rendered first or second virtual sound objects 601,602;
changing a volume level output of one of the rendered first or second virtual sound objects 601,602;
terminating the rendering of one of the first or second virtual sound objects 601,602; or
replaying a rendering of at least a part of one of the first or second virtual sound objects 601,602.

[0057] In some examples, the apparatus 500 is configured to receive a user input to control a change in the spatial rendering of one of the first or second virtual sound objects 601,602. Following which, the apparatus 500 generates a signal indicative of the user's user-controlled change of the spatial rendering of the first or second virtual sound objects 601,602. The apparatus 500 may then transmit the signal to a further apparatus (e.g. further apparatus 803 of FIG. 8).

[0058] The further apparatus may comprise means configured to cause:

spatially rendering:

the first audio data 701 in a second virtual sound scene such that it is rendered with a third virtual position within the second virtual sound scene, and/or

the second audio data 702 in the second virtual sound scene such that it is rendered with a fourth virtual position within the second virtual sound scene;

receiving the signal from the apparatus 500; and

controlling the spatial rendering of the third and/or fourth virtual sound objects based on the received signal.

[0059] In such an example, the user's user-controlled change of the spatial rendering of the first or second virtual sound object 601, 602 may trigger a corresponding change in the spatial rendering, at the further apparatus, of the third and/or fourth virtual sound objects.

[0060] The further apparatus 803 may additionally comprise means configured to cause: transmitting, to the apparatus 500, one or more of the first and second audio data 701,702 representative of the first and second audio content 101,102; receiving the signal from the apparatus 500; and controlling the transmission of one or more of the first and second audio data 701,702 based on the received signal.

[0061] In such an example, the further apparatus 803 may be the source of the first and second audio content 101,102 for the first and second audio data 701,702. For example, the further apparatus may capture first audio content 101 and may generate second audio content 102 therefrom (or may transmit the first audio content 101 to a separate remote apparatus, e.g. server, which generates second audio content 102 therefrom). Moreover, in response to receipt of a signal indicative of a user controlled manipulation of the spatial rendering of the first and/or second virtual sound objects 601,602, the transmission of the first and/or second audio data 701,702 may be controlled. For example, if the user were to terminate the rendering of the first or second virtual sound objects 601,602, this could be signalled to the further apparatus which then ceases the transmission of the first or second audio data 701,702, thereby conserving bandwidth. Furthermore, the user's termination of the rendering of the first or second virtual sound objects 601,602 could trigger a notification/prompt/alert thereby notifying a user of the further apparatus 803 of dismissal of one of the first or second virtual sound objects 601,602 by the second user 802 of the apparatus 500. Likewise, a notification to the user of the further apparatus 803 of whether the second user 802 of the apparatus 500 has modified the rendering of the first virtual sound object 601 so as to de-emphasise the same (e.g. by virtue of a movement of its virtual position away from a front/central position, or a reduction of its volume or dismissing/terminating its rendering) or modified the rendering of the second virtual sound object 602 so as to emphasise the same (e.g. by virtue of a movement of its virtual position towards a central/frontal position or an increase in its volume) may indicate to the user of the further apparatus the degree to which the second user 802 comprehends the first language and requires a translation.

[0062] In some examples, a system is provided comprising the apparatus 500 and the further apparatus 803 as described above. The further apparatus 803 may comprise means configured to cause: determining a second time period between an end of the spatial rendering of the first virtual sound object 601 and an end of the rendering of the second virtual sound object 602; and rendering an end portion of the second virtual sound object 602, during the determined second time period after the end of the rendering of the first virtual sound object 601.

[0063] In some examples, the first audio content 101 may be generated by a first user, for example it may be the first user's voice/speech in a first language which is captured as first audio data 701 by an audio capture device of the first user. The captured first audio data 701 of the first user talking in the first language may be sent/transmitted to and received by the apparatus 500. In certain examples, the second audio data 702 is representative of a translation of the first user's speech. For example, the first audio content 101 undergoes an automatic language translation to generate second audio content 102 being an aural translation of first audio content, and wherein the second audio data is representative of the second audio content 102.

[0064] Automatic language translation can be achieved using various means. For example, an application or a service (e.g. in the cloud) may: receive the speech/spoken utterances of a user; recognize words therein; evaluate what the sentence or part of a sentence means (e.g., what an individual word most likely means in context with other words); and create the corresponding translation into a desired output language. The input and output languages may be given/pre-selected, or the input language may be recognized as part of the overall recognition task. Automatic language translation can utilize, e.g., speech-to-text (STT) and text-to-speech (TTS) techniques. At least one task in the chain may be performed by means of artificial intelligence (AI) such as deep neural networks (DNN). The automatic language service may be provided by a client device, e.g. a user's mobile communications device/smart phone, or a server.

[0065] The captured first audio data 701 representative of first audio content 101, i.e. the first user talking in the first language, may be processed and translated to the second language locally, i.e. on the first user's device, or alternatively, the first audio content may be sent to a remote device, such as a server, to undergo speech-to-speech translation to generate the second audio data 702 representative of second audio content 102 (i.e. translated speech in the second language) derived from the first audio content 101 (i.e. the original speech in the

first language). The second audio data 702 of the translation of the first user talking may be sent/transmitted to and received by the apparatus 500.

[0066] The transmission and receipt of the first or second audio data 701,702 may be via any suitable wired or wireless connection or wired or wireless communication network.

[0067] The transmission and receipt of the first or second audio data 701, 702 may utilise any suitable codec, such as speech and audio codecs. In some examples, immersive audio codecs may be used which support a multitude of operating points ranging from a low bit rate operation to transparency as well as a range of service capabilities, e.g., from mono to stereo to fully immersive audio encoding/decoding/rendering. An example of such a codec is the 3GPP IVAS (Immersive Voice and Audio Services) codec which is an extension of the 3GPP EVS codec and is intended for new immersive voice and audio services over 4G/5G. Such immersive services include, e.g., immersive voice and audio for virtual reality (VR). The multi-purpose audio codec is expected to handle the encoding, decoding and rendering of speech, music and generic audio. It is expected to support channel-based audio and scene-based audio inputs including spatial information about the sound field and sound sources. It is also expected to operate with low latency to enable conversational services as well as support high error robustness under various transmission conditions.

[0068] Where the IVAS codec is used, the first and second audio data input signals may be provided to an IVAS encoder in one of its supported formats (and in some allowed combinations of the formats). The IVAS decoder may likewise output the audio content in supported formats. In a pass-through mode the audio data could be provided in its original format after transmission (encoding/decoding).

[0069] Various, but not necessarily all, examples of the present disclosure can take the form of a method, an apparatus or a computer program. Accordingly, various, but not necessarily all, examples can be implemented in hardware, software or a combination of hardware and software. The above described functionality and method operations may be performed by an apparatus (for example such as the apparatus 500 illustrated in FIGs. 5 and 7) which include one or more components for effecting the above described functionality. It is contemplated that the functions of these components can be combined in one or more components or performed by other components of equivalent functionality.

[0070] FIG. 7 schematically illustrates a block diagram of an apparatus 500. The apparatus 500 comprises a controller 501. Implementation of the controller 501 can be as controller circuitry. Implementation of the controller 501 can be in hardware alone (for example processing circuitry comprising one or more processors and memory circuitry comprising one or more memory elements), have certain aspects in software including firmware alone or can be a combination of hardware and software (including firmware).

[0071] The controller 501 can be implemented using instructions that enable hardware functionality, for example, by using executable computer program instructions in a general-purpose or special-purpose processor that can be stored on a computer readable storage medium (disk, memory etc.) or carried by a signal carrier to be performed by such a processor.

[0072] In the illustrated example, the apparatus 500 comprises a controller 501 which is provided by a processor 502 and memory 503. Although a single processor 502 and a single memory are illustrated in other implementations there can be multiple processors and/or there can be multiple memories some or all of which can be integrated/removable and/or can provide permanent/semi-permanent/ dynamic/cached storage.

[0073] The memory 503 stores a computer program 504 comprising computer program code/ instructions 505 that control the operation of the apparatus 500 when loaded into the processor 502. The computer program code 505 provides the logic and routines that enable the apparatus 500 to perform the methods presently described.

[0074] The processor 502 is configured to read from and write to the memory 503. The processor 502 can also comprise an input interface 506 via which data and/or commands are input to the processor 502, and an output interface 507 via which data and/or commands are output by the processor 502.

[0075] The apparatus 500 therefore comprises:

at least one processor 502; and
at least one memory 503 including computer program code 505 the at least one memory 503 and the computer program code 505 configured to, with the at least one processor 502, cause the apparatus 500 at least to perform:

receiving first audio data 701 representative of first audio content 101;
receiving second audio data 702 representative of second audio content 102, wherein the second audio content 101 is derived from the first audio content 101;
rendering the first audio data 701 as a first virtual sound object 601 in a virtual sound scene 600 such that it is spatially rendered with a first virtual position 602p within the virtual sound scene 600;
rendering the second audio data 702 as a second virtual sound object 602 in the virtual sound scene 600 such that it is spatially rendered with a second virtual position 602p within the virtual sound scene 600; and
controlling the spatial rendering of the first and second virtual sound objects 601,602 such that the first and second virtual positions 601p,602p differ.

[0076] The computer program 504 can arrive at the

apparatus 500 via any suitable delivery mechanism 511. The delivery mechanism 511 can be, for example, a non-transitory computer-readable storage medium, a computer program product, a memory device, a record medium such as a compact disc read-only memory, or digital versatile disc, or an article of manufacture that tangibly embodies the computer program 504. The delivery mechanism can be a signal configured to reliably transfer the computer program 504. The apparatus 500 can receive, propagate or transmit the computer program 504 as a computer data signal. The apparatus 500 may comprise a transmitting device and a receiving device for communicating with remote devices via a communications channel (not shown).

[0077] As will be appreciated, any such computer program code 505 can be loaded onto a computer or other programmable apparatus (i.e., hardware) to produce a machine, such that the code/instructions when performed on the programmable apparatus create means for implementing the functions specified in the blocks. The computer program code 505 can also be stored in a computer-readable medium that can direct a programmable apparatus to function in a particular manner, such that the instructions stored in the computer-readable memory produce an article of manufacture including instruction means which implement the function specified in the blocks. The computer program code 505 can also be loaded onto a programmable apparatus to cause a series of operational actions to be performed on the programmable apparatus to produce a computer-implemented process such that the instructions which are performed on the programmable apparatus provide actions for implementing the functions specified in the blocks.

[0078] References to 'computer-readable storage medium', 'computer program product', 'tangibly embodied computer program' etc. or a 'controller', 'computer', 'processor' etc. should be understood to encompass not only computers having different architectures such as single /multi-processor architectures and sequential (Von Neumann)/parallel architectures but also specialized circuits such as field-programmable gate arrays (FPGA), application specific circuits (ASIC), signal processing devices and other devices. References to computer program, instructions, code etc. should be understood to encompass software for a programmable processor or firmware such as, for example, the programmable content of a hardware device whether instructions for a processor, or configuration settings for a fixed-function device, gate array or programmable logic device etc.

[0079] As used in this application, the term 'circuitry' refers to all of the following:

(a) hardware-only circuit implementations (such as implementations in only analog and/or digital circuitry) and

(b) to combinations of circuits and software (and/or firmware), such as (as applicable): (i) to a combination of processor(s) or (ii) to portions of processor(s)/software (including digital signal processor(s)), software, and memory(ies) that work together to cause an apparatus, such as a mobile phone or server, to perform various functions and

(c) to circuits, such as a microprocessor(s) or a portion of a microprocessor(s), that require software or firmware for operation, even if the software or firmware is not physically present.

[0080] This definition of 'circuitry' applies to all uses of this term in this application, including in any claims. As a further example, as used in this application, the term "circuitry" would also cover an implementation of merely a processor (or multiple processors) or portion of a processor and its (or their) accompanying software and/or firmware. The term "circuitry" would also cover, for example and if applicable to the particular claim element, a baseband integrated circuit or applications processor integrated circuit for a mobile phone or a similar integrated circuit in a server, a cellular network device, or other network device.

[0081] In the present description, the apparatus 500 described can alternatively or in addition comprise an apparatus which in some other embodiments comprises a distributed system of apparatuses, for example, a client/server apparatus system. In examples of embodiments where the apparatus 500 forms (or the method 400 is implemented as) a distributed system, each apparatus forming a component and/or part of the system provides (or implements) one or more features which collectively implement an example of the present disclosure. In some examples of embodiments, the apparatus 500 is re-configured by an entity other than its initial manufacturer to implement an example of the present disclosure by being provided with additional software, for example by a user downloading such software, which when executed causes the apparatus 500 to implement an example of the present disclosure (such implementation being either entirely by the apparatus 500 or as part of a system of apparatuses as mentioned hereinabove).

[0082] The apparatus 500 may be comprised in a device 510 and additionally comprise further components/modules 508, 509 for providing additional functionality, e.g. such as not least: data interfaces for wired/wireless data connectively, user input (e.g. buttons, microphone, touch screen ...) output devices (e.g. speakers for spatial audio rendering, display, haptic means), and sensors (not least for detecting: movement, position and orientation).

[0083] The apparatus 500, or system in which it may be embodied, can be not least for example one or more of: a client device, a server device, a user equipment device, a wireless communications device, a portable device, a handheld device, a wearable device, a head mountable device etc. The apparatus 500 can be embodied by a computing device, not least such as those mentioned above. However, in some examples, the apparatus 500 can be embodied as a chip, chip set or mod-

ule, i.e. for use in any of the foregoing.

[0084] In one example, the apparatus 500 is embodied on a hand held portable electronic device, such as a mobile telephone, wearable computing device or personal digital assistant, that can additionally provide one or more audio/text/video communication functions (e.g. telecommunication, video-communication, and/or text transmission (Short Message Service (SMS)/ Multimedia Message Service (MMS)/emailing) functions), interactive/non-interactive viewing functions (e.g. web-browsing, navigation, TV/program viewing functions), music recording/playing functions (e.g. Moving Picture Experts Group-1 Audio Layer 3 (MP3) or other format and/or (frequency modulation/amplitude modulation) radio broadcast recording/playing), downloading/sending of data functions, image capture function (e.g. using a (e.g. inbuilt) digital camera), and gaming functions.

[0085] The apparatus 500 can be provided in an electronic device, for example, mobile terminal, according to an exemplary embodiment of the present disclosure. It should be understood, however, that a mobile terminal is merely illustrative of an electronic device that would benefit from examples of implementations of the present disclosure and, therefore, should not be taken to limit the scope of the present disclosure to the same. While in certain implementation examples, the apparatus 500 can be provided in a mobile terminal, other types of electronic devices, such as, but not limited to, hand portable electronic devices, wearable computing devices, portable digital assistants (PDAs), pagers, mobile computers, desktop computers, televisions, gaming devices, laptop computers, cameras, video recorders, GPS devices and other types of electronic systems, can readily employ examples of the present disclosure. Furthermore, devices can readily employ examples of the present disclosure regardless of their intent to provide mobility.

[0086] The apparatus 500 can be provided in a module. As used here 'module' refers to a unit or apparatus 500 that excludes certain parts/components that would be added by an end manufacturer or a user.

[0087] The above described examples may find application as enabling components of: telecommunication systems; electronic systems including consumer electronic products; distributed computing systems; media systems for generating or rendering media content including audio, visual and audio visual content and mixed, mediated, virtual and/or augmented reality; personal systems including personal health systems or personal fitness systems; navigation systems; automotive systems; user interfaces also known as human machine interfaces; networks including cellular, non-cellular, and optical networks; ad-hoc networks; the internet; the internet of things; virtualized networks; and related software and services.

[0088] Although examples of the apparatus 500 have been described above in terms of comprising various components, it should be understood that the components can be embodied as or otherwise controlled by a corresponding controller 501 or circuitry such as one or more processing elements or processors 502 of the apparatus 500. In this regard, each of the components described above can be one or more of any device, means or circuitry embodied in hardware, software or a combination of hardware and software that is configured to perform the corresponding functions of the respective components as described above.

[0089] FIG. 8 schematically illustrates a high-level illustration of a system 800 and use case (namely automatic speech translation) according to an example of the present disclosure. Two users 801, 802 talk to each other via an audio/voice call. In this example, the further apparatus 803 comprises: an aural rendering module/headphones 803a for spatially rendering virtual sound objects, a mobile computing module/mobile device 803b comprising, not least a processor, a memory, a user interface and a display for enabling user control, audio pickup, and connectivity. The further apparatus 803 also comprises means 803c for performing Real Time Language Translation (RTLT) of the first user's speech translating it into a language understood by the second user 802.

[0090] Audio signals for two language tracks (the original speech and the translated speech), i.e. first and second audio data 701, 702, are received by an IVAS encoder 803d of the apparatus 803. The resulting bit stream 804 output from the IVAS encoder is transmitted over a network to an IVAS decoder/renderer 500d of the apparatus 500 of the second user 802. The IVAS decoder/renderer 500d decodes the two language tracks to (re-)create the first and second audio data 701, 702, which are aurally output via spatial rendering such that the second user 802 hears both the original speech/first audio content 101 and its translation/second audio content 102 as spatially separated first and second virtual sound objects 601, 602 as described further below with respect to FIGs. 9A and 9B.

[0091] FIG. 9A schematically illustrates a spatialization, i.e. the virtual positioning in a virtual sound scene of a second user 802, of the spatial rendering of a first virtual sound object 601 (namely the spatial rendering of the caller's voice/original voice track/first track/first audio content) which, for the purposes of illustration, is visually represented in FIG. 9A as an avatar representation of the caller/other user. FIG.9A shows the second user 802 using an apparatus 500 to have an IVAS voice call with the other user/caller (not shown). The apparatus 500 comprises a mobile device 500a (not least for: data connection/connectivity to an apparatus of the other user, user input/control and audio pickup) and headphones 500b (for the spatial audio rendering/presentation to the second user 802). The caller's voice/first audio content is provided as a directional sound source, i.e. a rendered first virtual sound object 601 having a first virtual position 601p, namely, in this example an azimuthal angle a as shown in FIG. 9B.

[0092] FIG. 9B schematically illustrates the additionally spatialization/virtual positioning of the spatial rendering

of a second virtual sound object 602 (namely a spatial rendering of a real time language translation "RTLT" of the caller's voice/RTLT voice track/second track/second audio content) which, for the purposes of illustration, is visually represented in FIG.9B as a further avatar representation of the caller/other user. In this example, second virtual sound object 602 is spatially rendered with a virtual position 602p that spatially mirrors the virtual position 601p of the first virtual sound object 601. There can be considered an angle a for the spatial rendering of the first virtual sound object 601 and an angle b for spatial rendering of the second virtual sound object 602 relative to a line running perpendicular to the line running between user's ears in a default orientation/direction/horizontal plane/horizon. In some examples, the angles a and b may be the same. In some examples, angle b may depend on angle a and some additional parameter, such as an additional parameter that may be sent by the caller, set by the receiving device, or depend on some parameter related to the RTLT transmission, e.g. not least relating to the transmission of the second audio data and/or start, duration and end times of the second audio content represented by the second audio data.

[0093] In some examples, when just the first virtual sound object 601 is being rendered alone, i.e. prior to the second virtual sound object 602 being rendered (e.g. due to delays in processing and generating the RTLT of the second audio content and generation of the second virtual sound object 602) the angle a may initially be small (e.g. 0°) i.e. such that the first virtual sound object 601 is rendered substantially directly in front of the second user 802. When the RTLT voice track/translation is introduced and is to start to be spatially rendered, the spatial rendering of the first virtual sound object 601 for the original voice track may move so as to "make room" for the spatial rendering of the second virtual sound object 602 for the RTLT voice track, i.e. the virtual position 601p of the first virtual sound object 601 may move (e.g. angle a is increased) so as to accommodate the spatial rendering of the second virtual sound object 602 and its opposing angle b of its virtual position 602p. This may enable an increased spatial separation of the spatialization of the rendering first and second virtual sound objects 601,602. This may help the second user 802 to better separate and distinguish between the first and second virtual sound objects 601,602 (i.e. the original voice and translation of the same) such that the two voices/languages can be presented, at least in part simultaneously (as discussed further below with respect to FIG. 10A).

[0094] The angles a and b can depend on various attributes and signalling. In particular, voice activity and RTLT track duration information can be used to indicate to the second user 802, without a display or additional audio prompts, the length of the current RTLT track (i.e. second audio content) that still remains. For example, at a certain time threshold the virtual position 602p of the second virtual sound object 602 may begin to move towards the centre front of the user's virtual sound scene. Accordingly, when such a movement occurs, the second user 802 is notified that the certain time threshold has been crossed.

[0095] FIG. 10A schematically illustrates a time line of an example of the rendering of spatial RTLT according to the present disclosure. In this example, the RTLT is provided by a local service on a device of the first user 801. The first user 801 talks to the second user 802 via a voice communication channel. The first user's speech/talking corresponds to first audio content 101, which is transmitted, via first audio data, to a device of the second user 802. The first audio data, representative of the first audio content 101/first user's speech, is spatially rendered to the second user 802 as a first virtual sound object 601.

[0096] The first user's speech is translated, via RTLT, to translated speech which corresponds to second audio content 102 that is transmitted, via second audio data, to the second user 802. The second audio data, representative of the second audio content 102/first user's translated speech, is spatially rendered to the second user 802 as a second virtual sound object 602.

[0097] In this example, there is a time offset 1001a between the start of the first audio content 101 and a start of the second audio content 102, wherein the time offset 1001a is less that the duration of the first audio content 101. Similarly, there is a time offset 1002a between the start of the rendering of the first virtual sound object 601 and a start of the rendering of the second virtual sound object 602, wherein the time offset 1002a is less that the duration of the first virtual sound object 601, such that the rendering of the second virtual sound object 602 begins before the end of the rendering of the first virtual sound object 601. Thus, in effect the translation of the speech can be played back simultaneously with/overlapping with the original speech. For example, the system may translate word-by-word or sentence-by-sentence instead of one active passage at a time. The time offsets between the original voice segment and the translated segments need not be fixed. For example, the length of a segment of original speech that triggers the activation of translation of the same may vary. The provision of the separate voice tracks (original voice/speech and translation of the same, i.e. the first and second audio data representative of the first and second audio content) via an encoding/decoding (such as via IVAS encoders) and their spatial rendering enables such simultaneous playback/rendering to be feasible and intelligible by the second user 802 (taking advantage of the "cocktail party effect").

[0098] The time offset 1001a (secondary track offset) and the time offset 1001b (secondary track end offset) or any equivalent signalling concerning the same provides information about how long the duration of the active RTLT track/second audio content 102 is and how its presentation time relates to the presentation time of the current active audio passage/first audio content 101. This information can be provided at least to the receiving de-

vice and the second user 802 (as well as to the encoder-side device and a first user 801).

[0099] An active signal tail reproduction 602t is an example of a locally generated/rendered downstream audio indication for the first user 801. The tail of the translation 602 can, e.g., based on the time offset and duration signalling, be spatially rendered to the talker/first user 801. This way the first user 801 receives an indication on how long the second user 802 is still listening to incoming audio/second virtual sound object 602.

[0100] In some examples, there can be a signalling to indicate that the receiving second user 802 wishes to end the current secondary track playback/ second virtual sound object 602. This can be used to control the tail reproduction on the transmitting side, by ending reproduction the tail reproduction 602t of the upon recipient request. Thus, not only is the first user 801 made aware of the time delay between the systems (e.g. not least a time delay in the generation of the first audio content 101 and the rendering of the translation of the same/second virtual sound object 602), but the first user 801 also receives an indication of the dismissing of the secondary audio/ second virtual sound object 602 by the second user 802.

[0101] FIG. 10B schematically illustrates a time line of a further example of the rendering of spatial RTLT according to the present disclosure, somewhat similar to that of FIG. 10A, except that there is no overlapping/simultaneous rendering of the original speech and translation.

[0102] The speech/talking/first audio content 101 of the first user 801 is transmitted, via first audio data, to a device of the second user 802. The first audio data, representative of the first audio content 101/first user's speech, is spatially rendered to the second user 802 as a first virtual sound object 601. The first user's speech is translated, via a RTLT service local to the first user's device, to translated speech which corresponds to second audio content 102 that is transmitted, via second audio data, to the second user 802.

[0103] The transmission of the second audio data occurs sequentially to the transmission of the first audio data, taking into account the duration of the first audio content 101 as well as accommodating for any lag/delay e.g. during transmission.

[0104] The second audio data, representative of the second audio content 102/first user's translated speech, is spatially rendered to the second user 802 as a second virtual sound object 602. Furthermore, the second audio content 102 may also be spatially rendered locally to the first user 801 as a third virtual sound object 603, wherein the timing of the spatial rendering of the third virtual sound object 603 on first user's device is in synchronisation with the timing of the rendering of the second virtual sound object 602 on second user's device.

[0105] The above process may be repeat from the perspective of the second user 802, i.e. wherein the speech/first audio content 101 (2) of the second user 802 is spatially rendered to the first user 801 as first virtual sound object 601 (2), and the translated speech/second audio content 102(2) of the second user 802 is rendered to the first user 801 as second virtual sound object 602(2). The translated speech/second audio content 102(2) may also be rendered to the second user 802 as virtual sound object 604(2).

[0106] FIGs.11A-C schematically illustrate an example of user control of the spatial rendering of one of the first and second virtual sound objects 601, 602, which may represent a first voice track and a secondary related voice track (e.g. a RTLT track of the first voice track). In this example, the second user 802 receives two voice tracks: an original voice track and an RTLT targeting the user's own language which are spatially rendered as first and second virtual sound objects 601,602. However, in this example, the second user 802 is not able to understand much/any of the original voice track/first virtual sound object 601, and wishes to make it less disturbing/distracting. The user provides a user input to reduce the volume of (or completely mute) the rendering of the original voice track/first virtual sound object 601 so as to enable the user to better concentrate on the translated voice track/second virtual sound object 602. Such a user-controlled reduction in the volume of the original voice track/first virtual sound object 601 automatically controls the rendering of the RTLT track/second virtual sound object 602 so as to make the RTLT track/second virtual sound object 602 more pronounced/emphasised and easier to listen to in some additional way. For example, it is here illustrated that the rendering of the RTLT track/second virtual sound object 602 is automatically repositioned to the front centre of the second user 802, as shown in FIG. 9C. An alternative modification to the spatial rendering would be to increase the spatial separation of the first and second virtual sound objects 601,602 (e.g. where the volume of one is reduced - but not muted), for example not least by increasing angles a and b as (discussed with respect to FIG 9B).

[0107] FIGs. 12A and 12B schematically illustrate an example of user control of the spatial rendering of one of the first and second virtual sound objects 601, 602, wherein upon a first user 801 controlling the movement of one of the first and second virtual sound objects 601, 602, the other of the first and second virtual sound objects 601, 602 is automatically moved. For example, upon a user command to change the virtual position of the second virtual sound object, this triggers the virtual position of the second virtual sound object to be automatically repositioned, thereby maintaining a spatial separation between the first and second virtual sound objects 601,602.

[0108] FIGs. 13A -13C schematically illustrate a user control use case (referred to as "secondary track dismissal") that particularly relates to RTLT systems (and any other system where the length of an active audio passage is transmitted or otherwise known upon its the rendering). As illustrated in FIG.10A, there can be a tail of RTLT voice track/second virtual sound object 602 that

is presented also to the first user 801 whose voice is being translated, namely the virtual sound object portion/tail reproduction 602t. Thus, the first user 801 knows there is active content being consumed by the receiving/second user 802. In FIG 13A, the talker (user 1) 801 is still talking, but in FIG. 13B the first user 801 has stopped talking with the tail 602t playback continuing. The second user 802 is not talking. In this case the two users 801, 802 may, e.g., at least partly understand each other. For example, the second user 802 to some degree understands the first language of first user's 801 speech, but he requires help via the RTLT/second virtual sound object 602. When the second user 802 has understood the message (in this case the original voice in language 1 may have been enough), it may be indicated to the second user 802 that the translation is still significantly behind (e.g. 10 secs for a long utterance). As is seen in FIG 13C, the second user 802 at this point dismisses the secondary audio track/second virtual sound object 602 via their UI. The corresponding tail reproduction 602t ends also for the first user 801. In FIG 13D, there is thus no more playback (although the RTLT track from user 1 may still have content, i.e. remaining tail which is not being rendered to either user). Both users know that it is now fine to continue talking.

[0109] A further user command/user selection that may be signalled from the receiving device to the transmitting device relates to a selection of the language track. For example, in some cases it may be that it is not known which translation is desired by the recipient/second user 802. A transmitting device (or service) may thus send more than one translation as separate virtual sound objects, i.e. there may be a plurality of second virtual sound objects each spatially rendered with differing virtual positions. The second user 802 may then indicate which one of the plurality of second virtual sound objects/ plurality of translations he wished to receive. The user selection is provided to the transmitting device, which can discontinue sending the remaining, non-selected, plurality of second virtual sound objects, as these are unnecessary, thereby conserving bandwidth.

[0110] In some embodiments, additional indications can be signalled and conveyed to the user(s) 801, 802 based on the usage of the secondary track/second virtual sound object at the other end. Possible indications include signalling how the second user 802 utilizes the original voice track/first virtual sound object 601 relative to the secondary track/second virtual sound object 602 and vice versa. By default, the first user 801 does not know how well the second user 802 understands the original language and how well the second user 802 understands the translation. These pieces of information can be valuable for the first user 801. Thus, it can be signalled to the first user 801, e.g., what type of changes the second user 802 makes in their virtual sound scene. For example, it can be tracked the relative volumes at playback, the spatial modification by the receiving second user 802, the usage of replay functionality (discussed below) and so on. This can be indicated to first user 801, e.g., by modifying the RTLT tail position 604t and/or volume. For example, a user adjusting the spatial placement of a virtual sound object towards back is indicative that it is not considered so important. Whereas moving a virtual sound object towards front is indicative that it is considered important. This can allow for the first user 801 to get feedback spatially on how important the second user 802 feels the translation is.

[0111] A further use case relates to a conference call system or any other suitable service based on at least one audio input, where a virtual sound scene is delivered to a receiving user that is created, e.g. by a conference call service. The virtual sound scene may include, e.g., at least two independent talkers (for example users calling from their respective homes) or at least two talkers from the same capture space (for example a shared meeting room), where each talker may be represented by at least one virtual sound object for their native language speech and a further virtual sound object for their RTLT speech. For the receiving user the virtual sound scene may be presented such that in a first direction the user hears a first talker's speech in a first language, in a second direction the user hears a second talker's speech in second language, in a third direction the user hears the first talker's speech in a third language, and in a fourth direction the user hears the second talker's speech in a fourth language. Alternatively, e.g., based on a user input, the virtual sound scene of the receiving user may be presented such that the user hears the first talker's speech in first language in the first direction or the first talker's speech in the third language in the third direction; and the second talker's speech in the second language in the second direction or the second talker's speech in the fourth language in the fourth direction. In the latter case, the first and the third direction may be the same direction, and the second and the fourth direction may be the same direction.

[0112] An additional use case relates to spatial replay of the translation. A first user talks in a first language, which is translated into second language. Both language tracks are transmitted to second user and presented spatially as virtual sound objects. The first user may, e.g., have a problem to which they need help from second user. The first user describes the issue and second user answers. The first user has trouble understanding the answer, and wishes to replay a part of it. The first user thus rewinds the translation track. This change in playback duration can be transmitted to the second user, who thus knows there is still active playback at the other end (even though second user is not talking anymore). In some embodiments, the replay can be rendered in a different spatial position as the real-time translation playback. The replay may hence be a new spatially rendered virtual sound object. Thus, user may locally create new instances of the utterance and, e.g., save at least one of them for later consumption.

[0113] A second example related to the spatial replay

of the translation is a language training service. For example, a student calls to a tutor. The student practices pronunciation of an utterance. The tutor may rate the utterance and transmit back a correct pronunciation and the student's pronunciation. Student can spatially replay and compare the pronunciations, i.e. the virtual sound objects. The replay in particular can allow the tutor to instruct several students in parallel, where each student is able to use the "downtime" when the tutor is instructing (actively sending voice/speech to) another student in a constructive way by comparing pronunciations. The relative spatial positioning can indicate the closeness of the pronunciation to the correct one. For example, this may be based on the rating of the expert tutor or an automatic analysis system.

[0114] Thus, examples of the disclosure, the: volume, spatial position, and usage of certain related audio tracks/virtual sound objects may be controlled. At least for some audio tracks/virtual sound objects, such as the RTLT audio tracks that are not from any live/real talker, there may also be temporal modifications, such as the replay functionality discussed above.

[0115] A further use case relates to a voice memo of the utterances/key words that may be collected during the translated call to aid in understanding the translation. This can relate to additional usage of an augmented reality (AR) device. In particular this can relate to embodiments where there is text being transmitted. For example, a user has a problem with a rental car. The user places a call using RTLT with a rental company service person. The user discusses his problem with the service person, and the translated messages are saved. Subsequently, an error message is displayed on the car that the user does not understand. When user views the error message via their AR device (mobile device, AR glasses, etc.), the corresponding key words (vs. the error message) from the original language are mapped with the translated and saved comments. Such translated and saved comments may be replayed to the user.

[0116] The spatial rendering and functionality of the various examples described above can also be utilized for other use cases. One such vast area of use cases is different entertainment audio effects and modifications. People in general like to utilize various modifications and funny effects in their communications, which is currently seen in various social media applications and services. When it comes to conversational voice services, a problem with standardized legacy voice systems however is that any funny modifications (if they were applied for the codec ingest) would override the original voice. This is generally not desirable. For example, a child having a voice call with their parent might wish to use a funny filter by default, while their parent receiving the call could like to have the option to hear the original voice regardless of the modification. Examples of the disclosure would allow to solve this mismatch in a straightforward manner by providing at least two virtual sound objects with spatial rendering.

[0117] For example, a user may apply a filter to their voice and make it cartoony. Both the original voice track and the modified version of the same, i.e. a cartoony voice track may be delivered and spatially rendered to the recipient. The spatial audio controls discussed above may also apply.

[0118] Such entertainment use case examples enabled by examples of the present disclose include:

- Cartoony voice manipulation where at least two voices are delivered to recipient. For example, a user's voice may be filtered to "sound like a cat", where special effects can also be added (as schematically illustrated in FIG. 14
- wherein, for the purposes of illustration, the secondary/modified version of the user's original voice track visually represented as a cat-like avatar representation of the user)
- A second (cartoony) voice that reacts to what the user is saying and provides jokes or funny sounds for the recipient. This may be considered to be a "translation service evolution" for entertainment purposes.
- "Spatialized animoji" with modified audio for video calls.

[0119] Yet another use case is an advertisement object. For example, a local service on the device can analyse a user's speech and fetch a suitable audio advertisement. This is added to the stream as a separate spatially rendered virtual sound object. The receiving user may receive some perks (e.g., free call or other service) based on how they treat the advertisement virtual sound object rendering. For example, should the recipient mute or dismiss the advertisement, they receive nothing; should they let it play, they receive points/credit.

[0120] Further use cases relate to the provision of an alert. In a first example, there is a local alert service on the device (or a connected device). For example, a user may be walking while on a voice call, slip and fall. A locally generated alert is sent as separate virtual sound object to the recipient of the call to inform them of a potential medical emergency. In a second example, there is a remote/network operated alert service on the device. For example, a user is on a call and the authorities are aware of a bear moving about in the area of the cell. The network adds a virtual spatial object to the user's downstream audio package giving a warning to the user.

[0121] It is assumed above that a codec, and codec capability, negotiation takes place to establish the audio call between the at least two parties. Depending on the final codec standard/implementation, such negotiation can explicitly include, e.g., RTLT features (or any other suitable features), or it can implicitly allow for such features. For example, supporting several virtual sound object streams (at least two for the RTLT) and suitable dependency signalling for the streams, it is possible to pass an RTLT signal through the IVAS codec. Advanced user

manipulations, such as dismissing a secondary audio track playback such that this information is available for the encoder-side device will require additional signalling such as suitable codec mode request (CMR). Such request could be called, e.g., Alternative audio Track Request (ATR).

[0122] FIG. 15 schematically illustrates two further high-level illustrations of systems (with differing system architecture/signal routing) to that of FIG. 7. In the top system, a speech signal in a first language from a first user 801 is encoded using an IVAS encoder 803d of an apparatus 803 of the first user 801. A node/server 1501 in the network processes and decodes the incoming IVAS bitstream, translates the speech signal into a second language (i.e. understood by a second user 802), and encodes at least the translated signal and repacketizes the audio for transmission. Depending on the implementation one or two streams are thus sent from the network node to an IVAS decoder 500d of an apparatus 500 of a receiving second user 802. The IVAS decoder/renderer 500d,500 outputs the two language tracks 701, 702 as first and second virtual sound objects for spatial rendering/audio presentation such that the second user 802 can hear the original audio and the translation.

[0123] In the bottom system, a speech signal in a first language from a first user 801 is encoded using an IVAS encoder 803d of an apparatus 803 of the first user 801. It is transmitted to a second user 802 for decoding and spatial rendering.

[0124] In addition, it is also transmitted to an external RTLT service 1502 that decodes, translates, and encodes the speech signal then transmits the translated signal to a second user 802 for decoding and spatial rendering.

[0125] The system architecture of FIG.7 has some benefits over those of FIG.15, namely relating to 1) reduced delay between the original speech (1st language) and the translation (2nd language), and 2) user control. The delay reduction happens for two reasons. First, the local RTLT can in some implementations bypass at least some audio processing (that will introduce delay) that the regular IVAS input will be subject to. This can relate to, for example, equalization of the microphone signal(s) and so on. Such processing may be bypassed for the RTLT ingest because the output from the RTLT is a synthetic speech which can be automatically controlled. The RTLT ingest need not sound optimal to a human listener, only its output should appear natural. Second, there is no additional decoding/encoding delay in the path, which would introduce delay for the signal. Furthermore, maximal control over the features is allowed with encoder-side operation and in-band signalling. This advantage relates also to other use cases than RTLT.

[0126] There are use cases where a network service, e.g. as per FIG. 15, is needed or preferred to that of FIG. 8. For example, in a conference call use case it can be the only practical solution to carry out a translation in the cloud/ as a network service, when there are for example several different languages required for the downstream audio (e.g. to different participants). This is because there would be more computational capacity in the cloud and the bitrate available for each participant upstream may be limited (so, it is not possible to send, e.g., 10 translations to the server for each participant).

[0127] Various, but not necessarily all, examples of the present disclosure are described using flowchart illustrations and schematic block diagrams. It will be understood that each block (of the flowchart illustrations and block diagrams), and combinations of blocks, can be implemented by computer program instructions of a computer program. These program instructions can be provided to one or more processor(s), processing circuitry or controller(s) such that the instructions which execute on the same create means for causing implementing the functions specified in the block or blocks, i.e. such that the method can be computer implemented. The computer program instructions can be executed by the processor(s) to cause a series of operational steps/actions to be performed by the processor(s) to produce a computer implemented process such that the instructions which execute on the processor(s) provide steps for implementing the functions specified in the block or blocks.

[0128] Accordingly, the blocks support: combinations of means for performing the specified functions; combinations of actions for performing the specified functions; and computer program instructions/algorithm for performing the specified functions. It will also be understood that each block, and combinations of blocks, can be implemented by special purpose hardware-based systems which perform the specified functions or actions, or combinations of special purpose hardware and computer program instructions.

[0129] Various, but not necessarily all, examples of the present disclosure provide both a method and corresponding apparatus comprising various modules, means or circuitry that provide the functionality for performing/applying the actions of the method. The modules, means or circuitry can be implemented as hardware, or can be implemented as software or firmware to be performed by a computer processor. In the case of firmware or software, examples of the present disclosure can be provided as a computer program product including a computer readable storage structure embodying computer program instructions (i.e. the software or firmware) thereon for performing by the computer processor.

[0130] Where a structural feature has been described, it can be replaced by means for performing one or more of the functions of the structural feature whether that function or those functions are explicitly or implicitly described.

[0131] Although specific terms are employed herein, they are used in a generic and descriptive sense only and not for purposes of limitation.

[0132] Features described in the preceding description can be used in combinations other than the combinations

explicitly described.

[0133] Although functions have been described with reference to certain features, those functions can be performable by other features whether described or not.

[0134] Although features have been described with reference to certain examples, those features can also be present in other examples whether described or not. Accordingly, features described in relation to one example/aspect of the disclosure can include any or all of the features described in relation to another example/aspect of the disclosure, and vice versa, to the extent that they are not mutually inconsistent.

[0135] Although various examples of the present disclosure have been described in the preceding paragraphs, it should be appreciated that modifications to the examples given can be made without departing from the scope of the invention as set out in the claims. For example, whilst examples have been disclosed with: one/a single first audio content, one/a single first virtual sound object one/a single second audio content, one/a single second virtual sound object; in some examples, there may be plural first and second audio contents and plural first and second virtual sound objects. Whilst examples of the disclosure are foreseen to be implemented using the IVAS codec, implementation it can also be used with other codecs and communication protocols, for example, such implementations could carry text instead of voice during transmission, which text is then converted to voice and is received by the apparatus for spatial rendering.

[0136] The term 'comprise' is used in this document with an inclusive not an exclusive meaning. That is any reference to X comprising Y indicates that X can comprise only one Y or can comprise more than one Y. If it is intended to use 'comprise' with an exclusive meaning then it will be made clear in the context by referring to "comprising only one ..." or by using "consisting".

[0137] In this description, the wording 'communication' and its derivatives mean operationally in communication. It should be appreciated that any number or combination of intervening components can exist (including no intervening components), i.e. so as to provide direct or indirect communication. Any such intervening components can include hardware and/or software components.

[0138] As used herein, the "determining" (and grammatical variants thereof) can include, not least: calculating, computing, processing, deriving, investigating, looking up (e.g., looking up in a table, a database or another data structure), ascertaining and the like. Also, "determining" can include receiving (e.g., receiving information), accessing (e.g., accessing data in a memory) and the like. Also, "determining" can include resolving, selecting, choosing, establishing, and the like.

[0139] In this description, reference has been made to various examples. The description of features or functions in relation to an example indicates that those features or functions are present in that example. The use of the term 'example' or 'for example', 'can' or 'may' in the text denotes, whether explicitly stated or not, that such features or functions are present in at least the described example, whether described as an example or not, and that they can be, but are not necessarily, present in some or all other examples. Thus 'example', 'for example', 'can' or 'may' refers to a particular instance in a class of examples. A property of the instance can be a property of only that instance or a property of the class or a property of a sub-class of the class that includes some but not all of the instances in the class.

[0140] In this description, references to "a/an/the" [feature, element, component, means ...] are to be interpreted as "at least one" [feature, element, component, means ...] unless explicitly stated otherwise. That is any reference to X comprising a/the Y indicates that X can comprise only one Y or can comprise more than one Y unless the context clearly indicates the contrary. If it is intended to use 'a' or 'the' with an exclusive meaning then it will be made clear in the context. In some circumstances the use of 'at least one' or 'one or more' can be used to emphasis an inclusive meaning but the absence of these terms should not be taken to infer and exclusive meaning.

[0141] The presence of a feature (or combination of features) in a claim is a reference to that feature) or combination of features) itself and also to features that achieve substantially the same technical effect (equivalent features). The equivalent features include, for example, features that are variants and achieve substantially the same result in substantially the same way. The equivalent features include, for example, features that perform substantially the same function, in substantially the same way to achieve substantially the same result.

[0142] In this description, reference has been made to various examples using adjectives or adjectival phrases to describe characteristics of the examples. Such a description of a characteristic in relation to an example indicates that the characteristic is present in some examples exactly as described and is present in other examples substantially as described.

[0143] The above description describes some examples of the present disclosure however those of ordinary skill in the art will be aware of possible alternative structures and method features which offer equivalent functionality to the specific examples of such structures and features described herein above and which for the sake of brevity and clarity have been omitted from the above description. Nonetheless, the above description should be read as implicitly including reference to such alternative structures and method features which provide equivalent functionality unless such alternative structures or method features are explicitly excluded in the above description of the examples of the present disclosure.

[0144] Whilst endeavouring in the foregoing specification to draw attention to those features of examples of the present disclosure believed to be of particular importance it should be understood that the applicant claims protection in respect of any patentable feature or combination of features hereinbefore referred to and/or shown

in the drawings whether or not particular emphasis has been placed thereon.

[0145] The examples of the present disclosure and the accompanying claims can be suitably combined in any manner apparent to one of ordinary skill in the art.

[0146] Each and every claim is incorporated as further disclosure into the specification and the claims are embodiment(s) of the present invention. Further, while the claims herein are provided as comprising specific dependencies, it is contemplated that any claims can depend from any other claims and that to the extent that any alternative embodiments can result from combining, integrating, and/or omitting features of the various claims and/or changing dependencies of claims, any such alternative embodiments and their equivalents are also within the scope of the disclosure.

Claims

1. An apparatus comprising means configured to cause:

receiving first audio data representative of first audio content;
receiving second audio data representative of second audio content, wherein the second audio content is derived from the first audio content;
rendering the first audio data as a first virtual sound object in a virtual sound scene such that it is spatially rendered with a first virtual position within the virtual sound scene;
rendering the second audio data as a second virtual sound object in the virtual sound scene such that it is spatially rendered with a second virtual position within the virtual sound scene; and
controlling the spatial rendering of the first and second virtual sound objects such that the first and second virtual positions differ.

2. The apparatus according to claim 1, further comprising means configured to cause:

simultaneously rendering the first and second virtual sound objects in the virtual sound scene; and
controlling the spatial rendering of the first and second virtual sound objects such that, at least whilst the first and second virtual sound objects are simultaneously spatially rendered, the first and second virtual positions differ.

3. The apparatus according to any of the previous claims, wherein the first and second virtual positions comprise first and second virtual orientations respectively, and wherein the apparatus further comprises means configured to cause:

controlling the spatial rendering of the first and second virtual sound objects such that the first virtual orientation of the spatially rendered first virtual sound object differs from the second virtual orientation of the spatially rendered second virtual sound object.

4. The apparatus according to claim 3, further comprising means configured to cause:
controlling the spatial rendering of the first and second virtual sound objects such that the first virtual orientation differs from the second virtual orientation by an azimuthal angle of greater than: 15°, 30°, 45°, 60°, 75°, 90°, 105°, 120°, 135°, 150°, or 165°.

5. The apparatus according to any of the previous claims, further comprising means configured to cause:
moving, during a first time period between a start of the spatial rendering of the first virtual sound object and a start of the rendering of the second virtual sound object, the first virtual position.

6. The apparatus according to any of the previous claims, further comprising means configured to cause:
moving, during a second time period following an end of the spatial rendering of the first virtual sound object, the second virtual position.

7. The apparatus according to any of the previous claims, further comprising means configured to cause:

receiving a user input to control the spatial rendering of one of the first or second virtual sound objects;
changing, responsive to receipt of the user input, the spatial rendering of the one of the first or second virtual sound objects; and
changing, responsive to the user-controlled change of the spatial rendering of the one of the first or second virtual sound objects, the spatial rendering of the other of the second or first virtual sound objects.

8. The apparatus according to claim 7, wherein changing the spatial rendering of one of the first or second virtual sound objects comprises one or more of:

changing a virtual position of one of the rendered first or second virtual sound objects;
changing a virtual orientation of one of the rendered first or second virtual sound objects;
changing a volume level output of one of the rendered first or second virtual sound objects;
terminating the rendering of one of the first or second virtual sound objects; and
replaying a rendering of at least a part of one of

the first or second virtual sound objects.

9. The apparatus of any of the previous claims, wherein the apparatus is at least part of: a portable device, a handheld device, a wearable device, a head mountable device, a wireless communications device, a user equipment device, a client device or a server.

10. The apparatus according to any of the previous claims, wherein the second audio content is a translation of the first audio content.

11. A system comprising:

the apparatus according to any of the previous claims; and
a further apparatus;
wherein the apparatus comprises means configured to cause:

receiving a user input to control the spatial rendering of one of the first or second virtual sound objects,
changing, responsive to receipt of the user input, the spatial rendering of the one of the first or second virtual sound objects, and
changing, responsive to the user-controlled change of the spatial rendering of the one of the first or second virtual sound objects, the spatial rendering of the other of the second or first virtual sound objects; and

wherein the further apparatus comprises means configured to cause:

spatially rendering:

the first audio data in a second virtual sound scene such that it is rendered with a third virtual position within the second virtual sound scene, and/or
the second audio data in the second virtual sound scene such that it is rendered with a fourth virtual position within the second virtual sound scene;

receiving the signal from the first apparatus; and
controlling the spatial rendering of the third and/or fourth virtual sound objects based on the received signal.

12. A system comprising:

the apparatus according to any of the previous claims 1 to 10; and
a further apparatus;
wherein the apparatus comprises means configured to cause:

receiving a user input to control the spatial rendering of one of the first or second virtual sound objects,
changing, responsive to receipt of the user input, the spatial rendering of the one of the first or second virtual sound objects, and
changing, responsive to the user-controlled change of the spatial rendering of the one of the first or second virtual sound objects, the spatial rendering of the other of the second or first virtual sound objects;
and

wherein the further apparatus comprises means configured to cause:

transmitting, to the apparatus, one or more of the first and second audio data representative of the first and second audio content;
receiving the signal from the first apparatus; and
controlling the transmission of one or more of the first and second audio data based on the received signal.

13. A system comprising:

the apparatus according to any one of more of previous claims 1 to 10; and
a further apparatus, wherein the further apparatus comprises means configured to cause:

determining a second time period between an end of the spatial rendering of the first virtual sound object and an end of the rendering of the second virtual sound object; and
rendering an end portion of the second audio content, during the determined second time period after the end of the rendering of the first virtual sound object.

14. A method comprising causing, at least in part, actions that result in:

receiving first audio data representative of first audio content;
receiving second audio data representative of second audio content, wherein the second audio content is derived from the first audio content;
rendering the first audio data as a first virtual sound object in a virtual sound scene such that it is spatially rendered with a first virtual position within the virtual sound scene;
rendering the second audio data as a second

virtual sound object in the virtual sound scene such that it is spatially rendered with a second virtual position within the virtual sound scene; and

controlling the spatial rendering of the first and second virtual sound objects such that the first and second virtual positions differ.

15. Computer program instructions for causing an apparatus to perform:

receiving first audio data representative of first audio content;

receiving second audio data representative of second audio content, wherein the second audio content is derived from the first audio content;

rendering the first audio data as a first virtual sound object in a virtual sound scene such that it is spatially rendered with a first virtual position within the virtual sound scene;

rendering the second audio data as a second virtual sound object in the virtual sound scene such that it is spatially rendered with a second virtual position within the virtual sound scene; and

controlling the spatial rendering of the first and second virtual sound objects such that the first and second virtual positions differ.

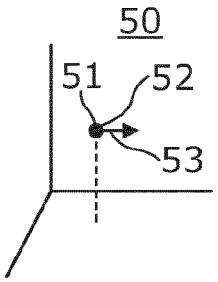

Fig. 1A

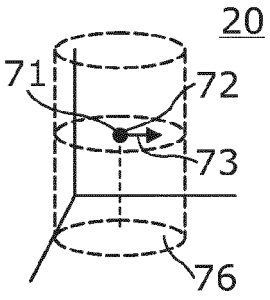

Fig. 2A

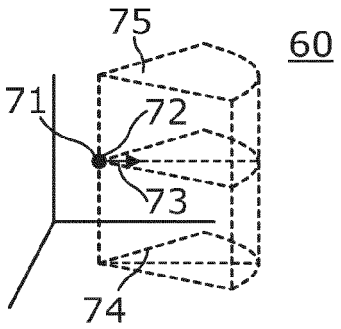

Fig. 3A

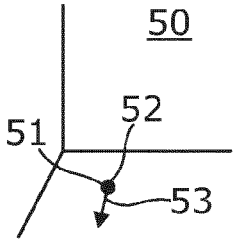

Fig. 1B

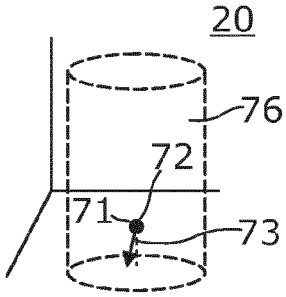

Fig. 2B

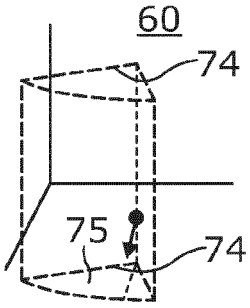

Fig. 3B

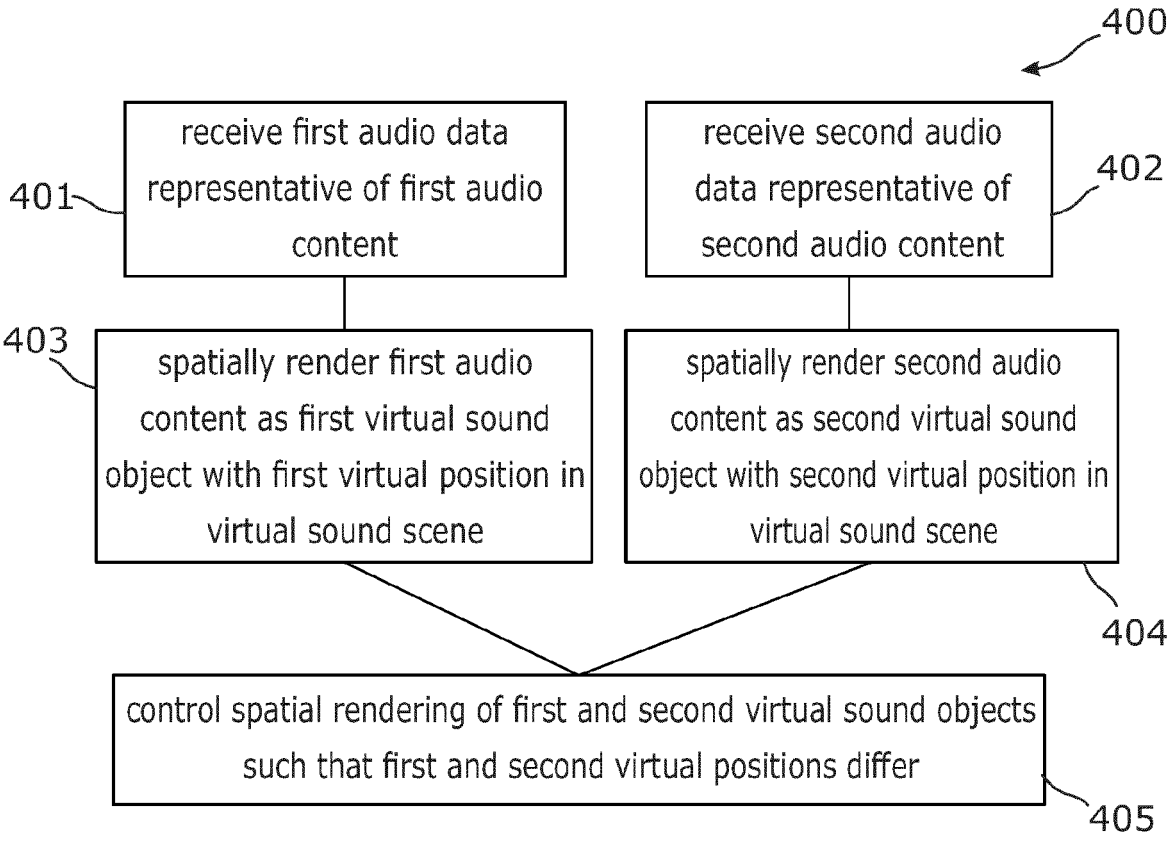

Fig. 4

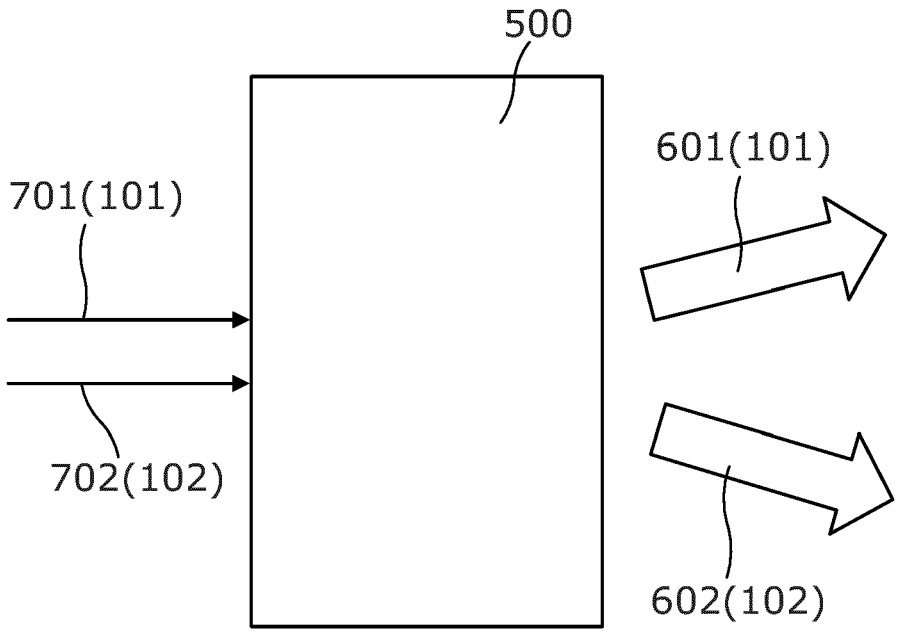
500
701(101)
702(102)
601(101)
602(102)

Fig. 5

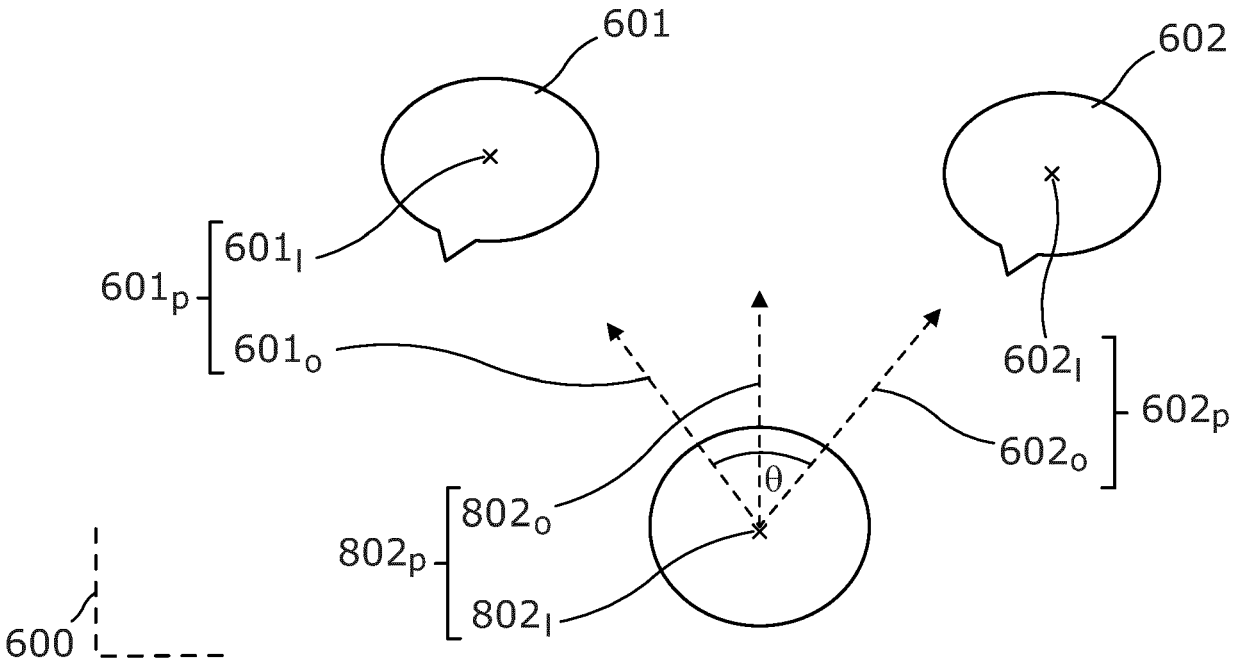
601
602
601_p
601_I
601_O
602_I
602_O
602_p
θ
802_O
802_p
802_I
600

Fig. 6

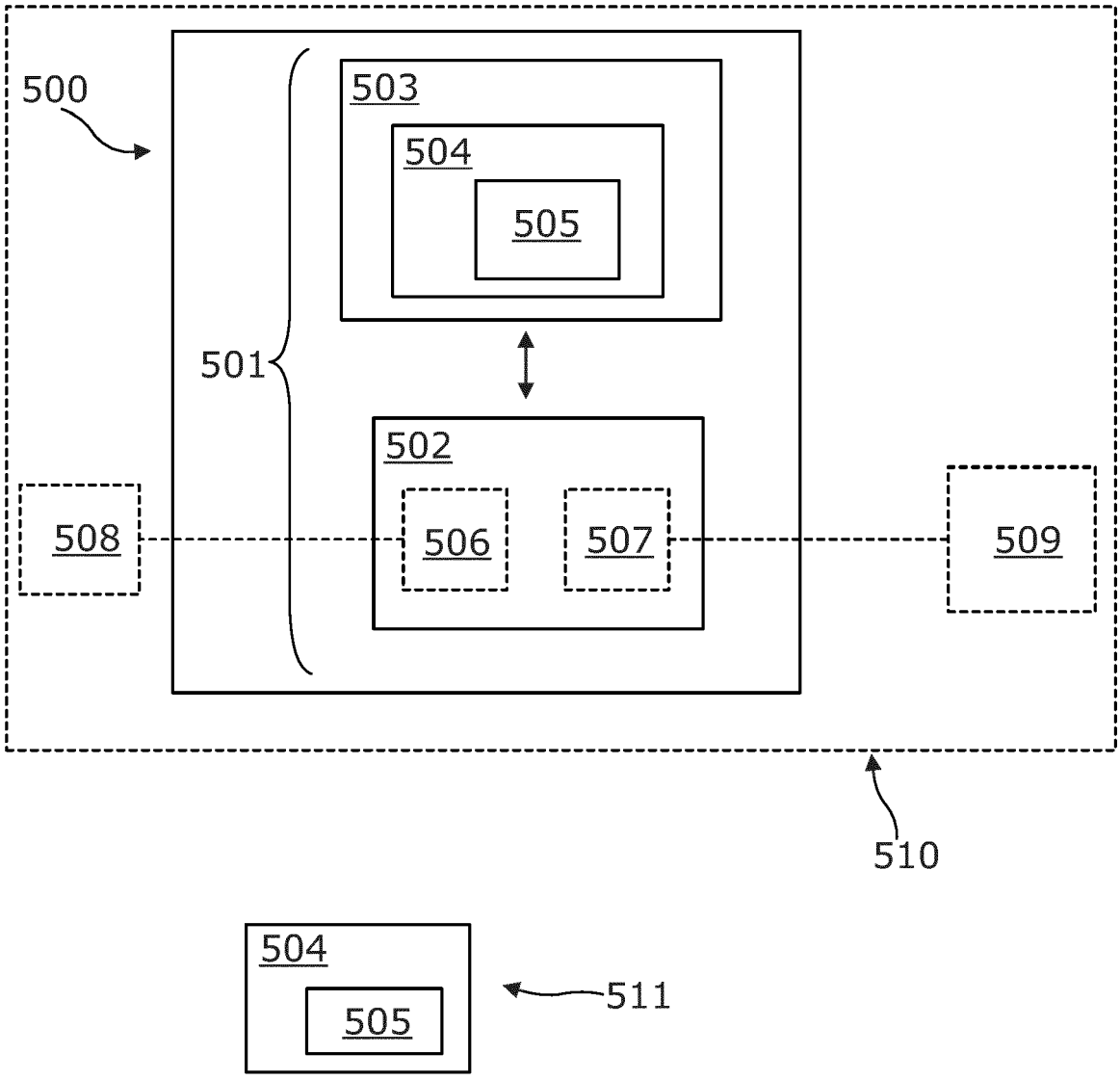

Fig. 7

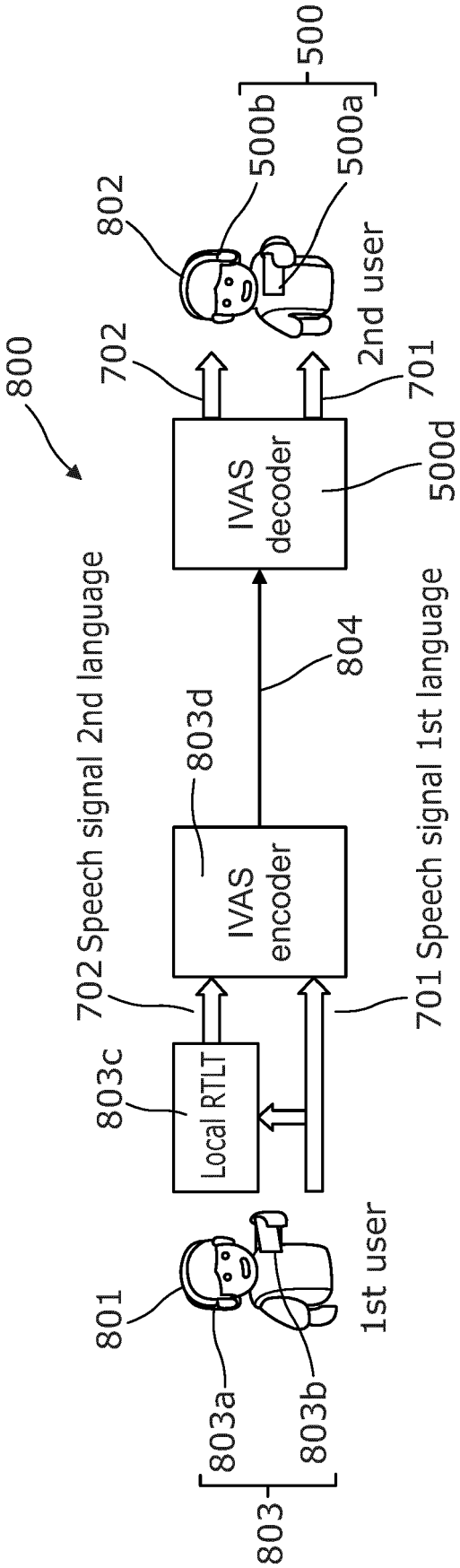
800
802
500b
500a
500
2nd user
702
701
IVAS decoder
500d
804
803d
IVAS encoder
702 Speech signal 2nd language
701 Speech signal 1st language
803c
Local RTLT
801
1st user
803a
803b
803

Fig. 8

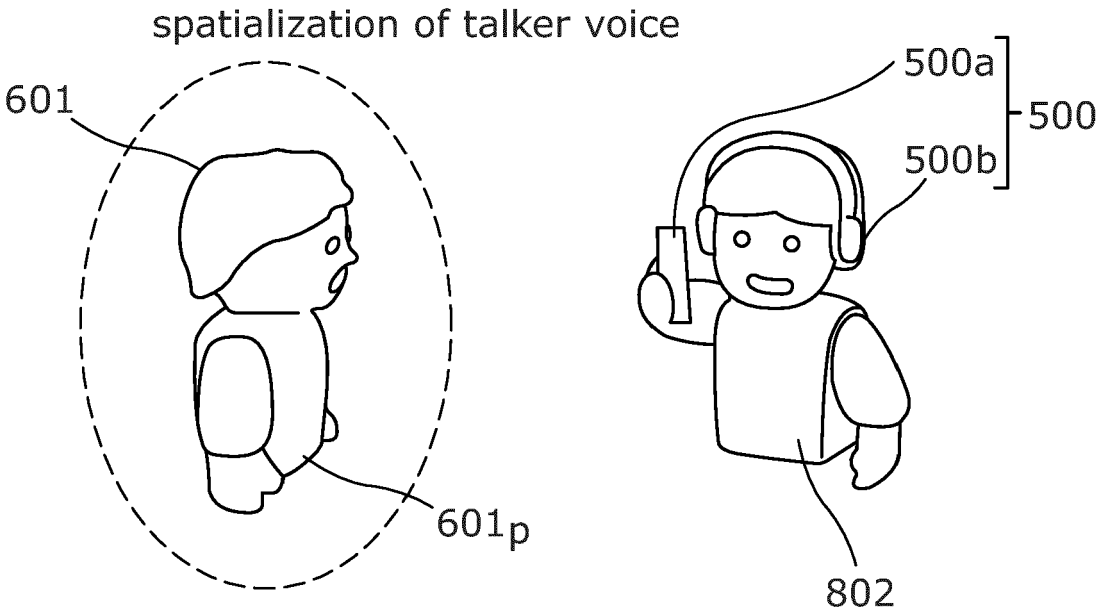
spatialization of talker voice
601
601p
500a
500
500b
802

Fig. 9A

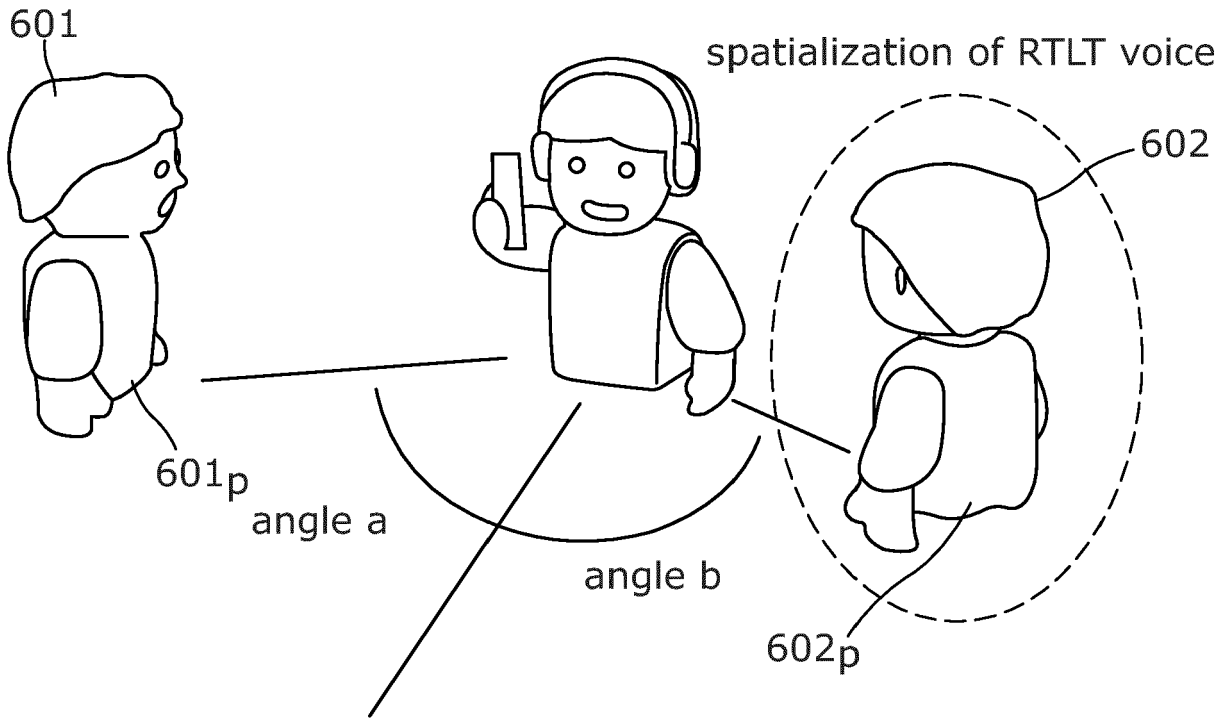
601
spatialization of RTLT voice
602
601p
angle a
angle b
602p

Fig. 9B

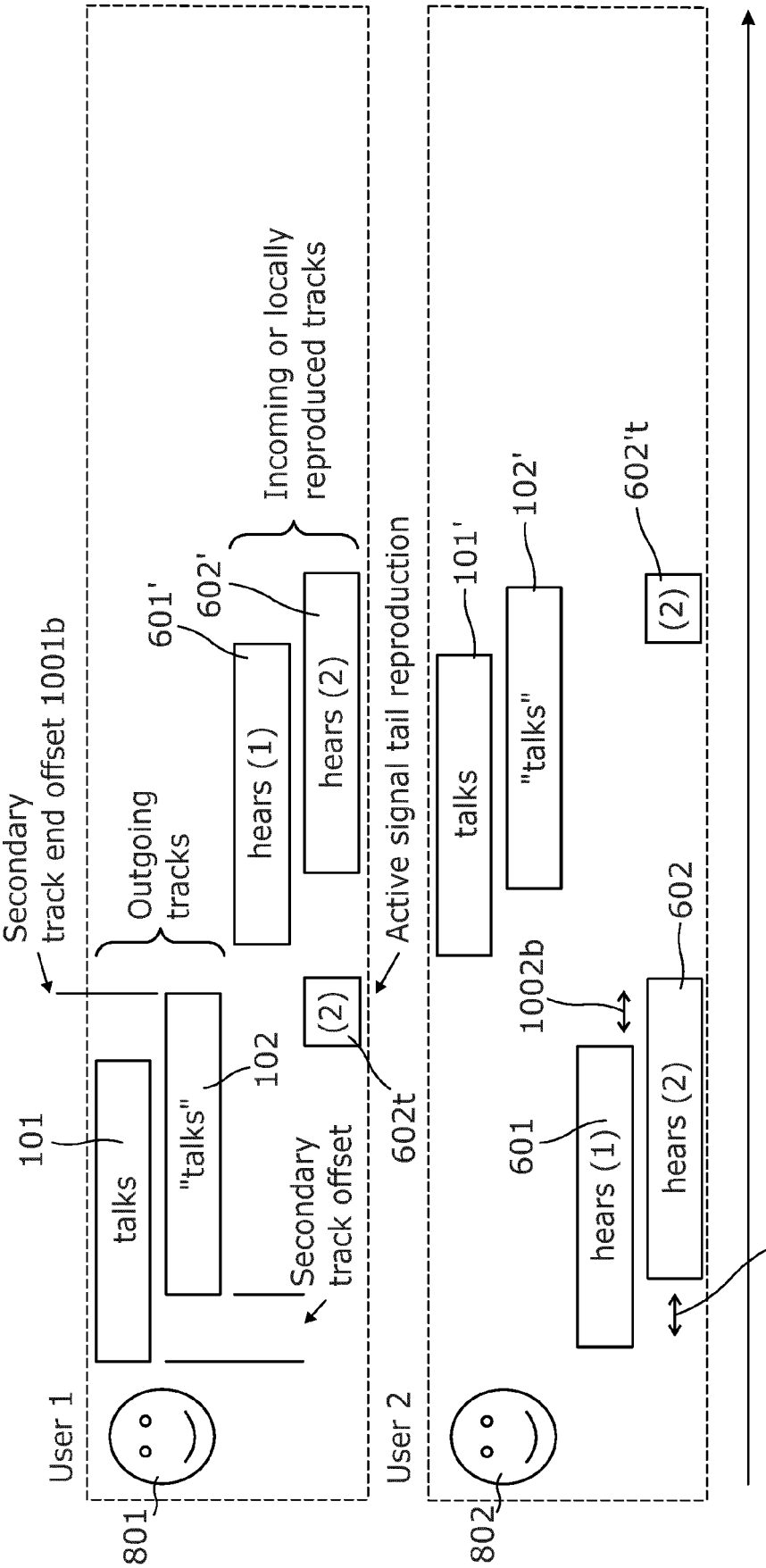
User 1
User 2
801
802
101
talks
102
"talks"
Secondary track offset
Secondary track end offset 1001b
Outgoing tracks
601'
602'
hears (1)
hears (2)
Incoming or locally reproduced tracks
(2)
602t
Active signal tail reproduction
101'
talks
102'
"talks"
602't
(2)
1002b
601
hears (1)
602
hears (2)
1002a
time

Fig. 10A

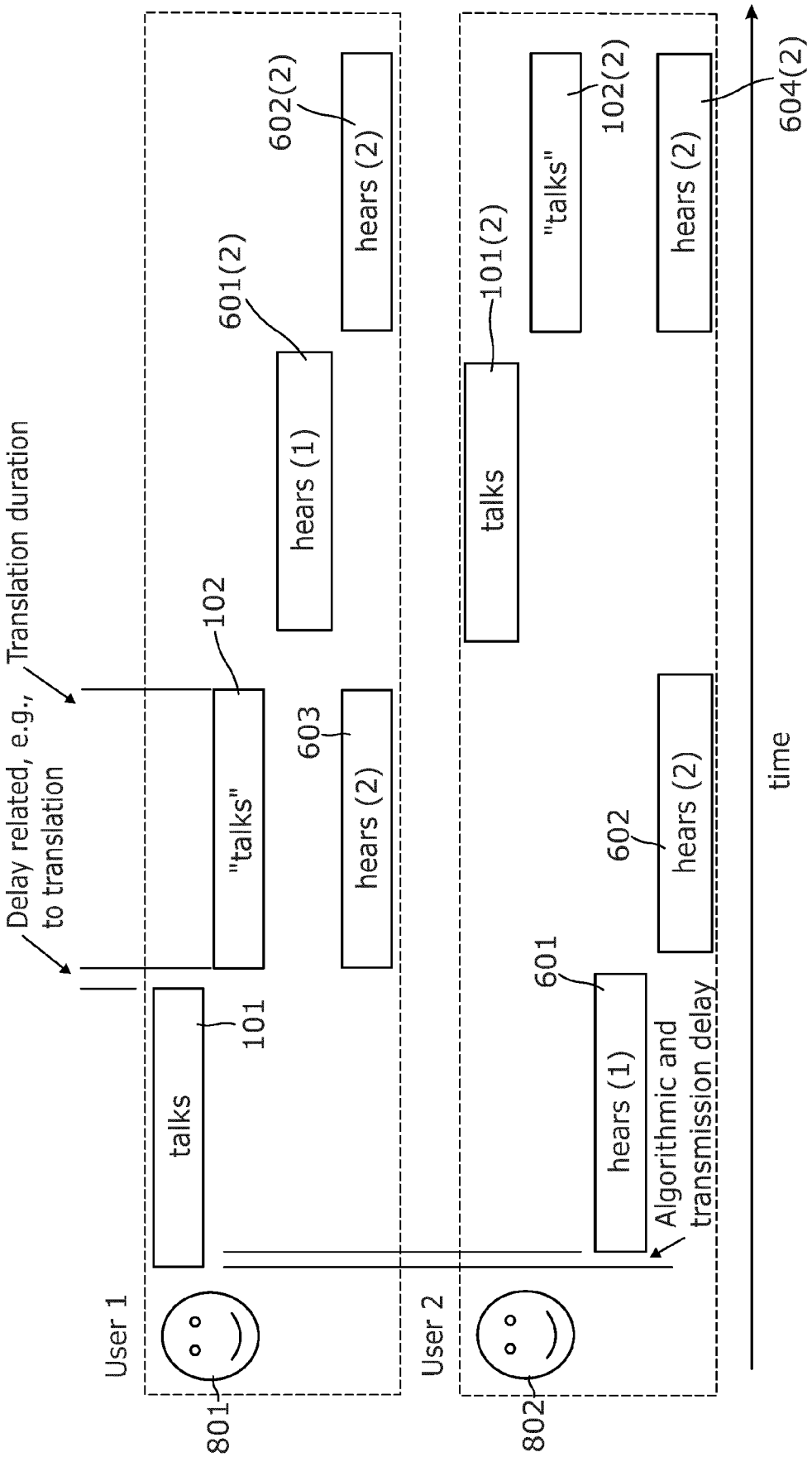
Translation duration
Delay related, e.g., to translation
User 1
User 2
talks
"talks"
hears (1)
hears (2)
101
102
601
602
603
101(2)
102(2)
601(2)
602(2)
604(2)
Algorithmic and transmission delay
time
801
802

Fig. 10B

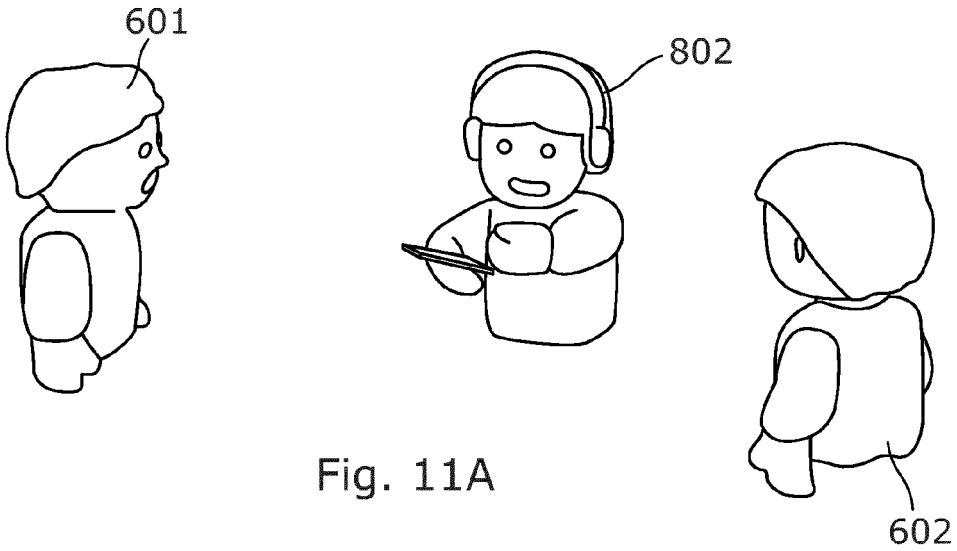

Fig. 11A

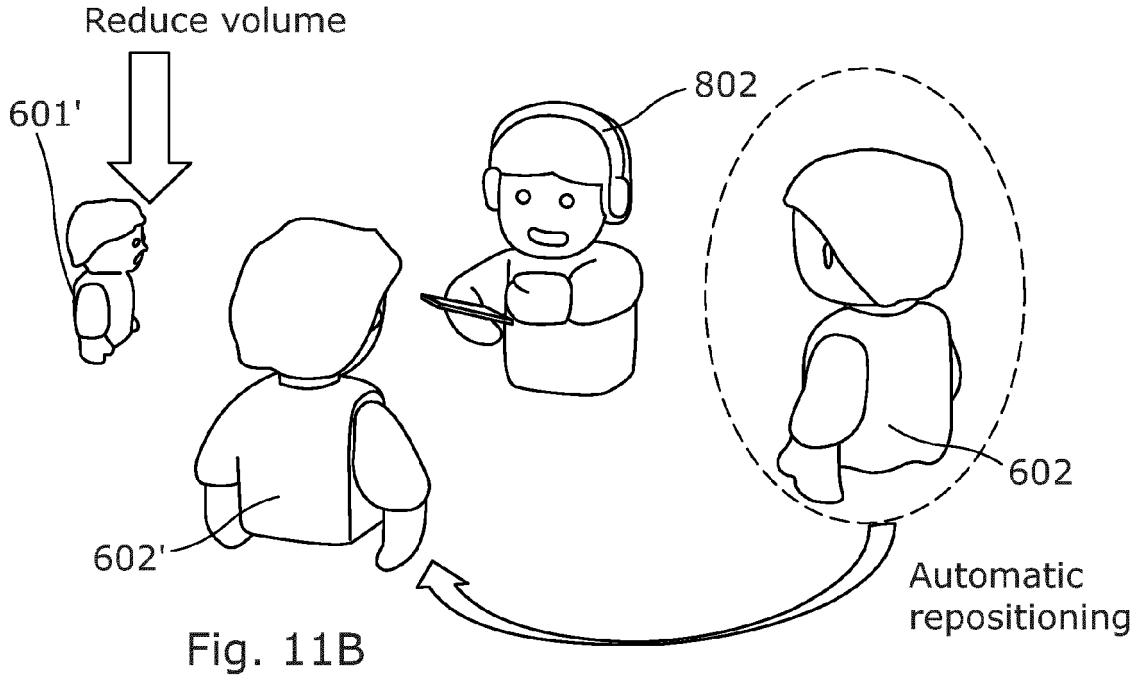

Fig. 11B

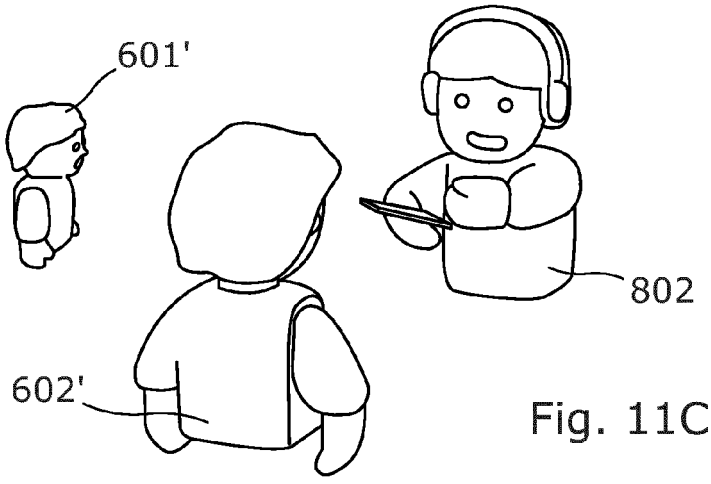

Fig. 11C

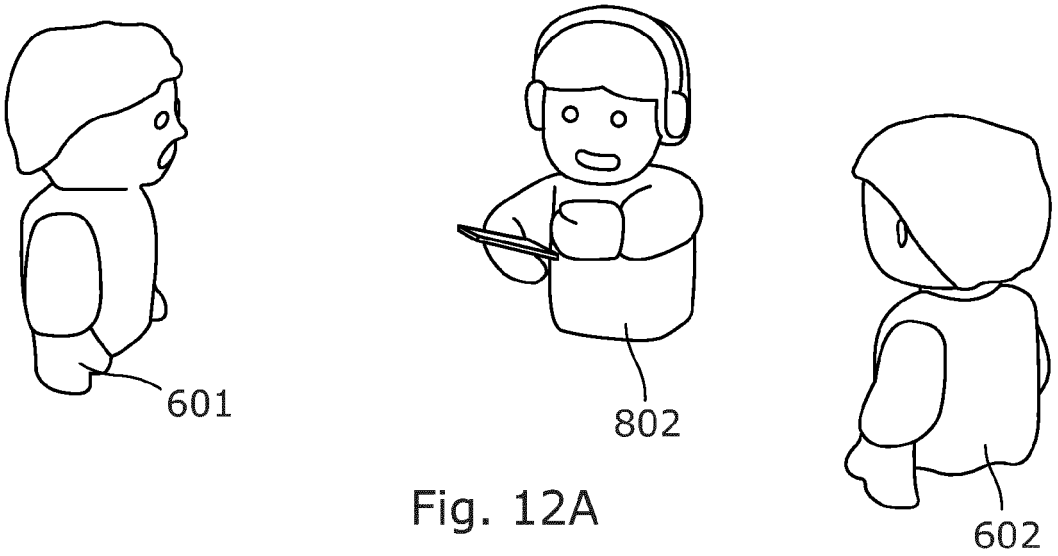

Fig. 12A

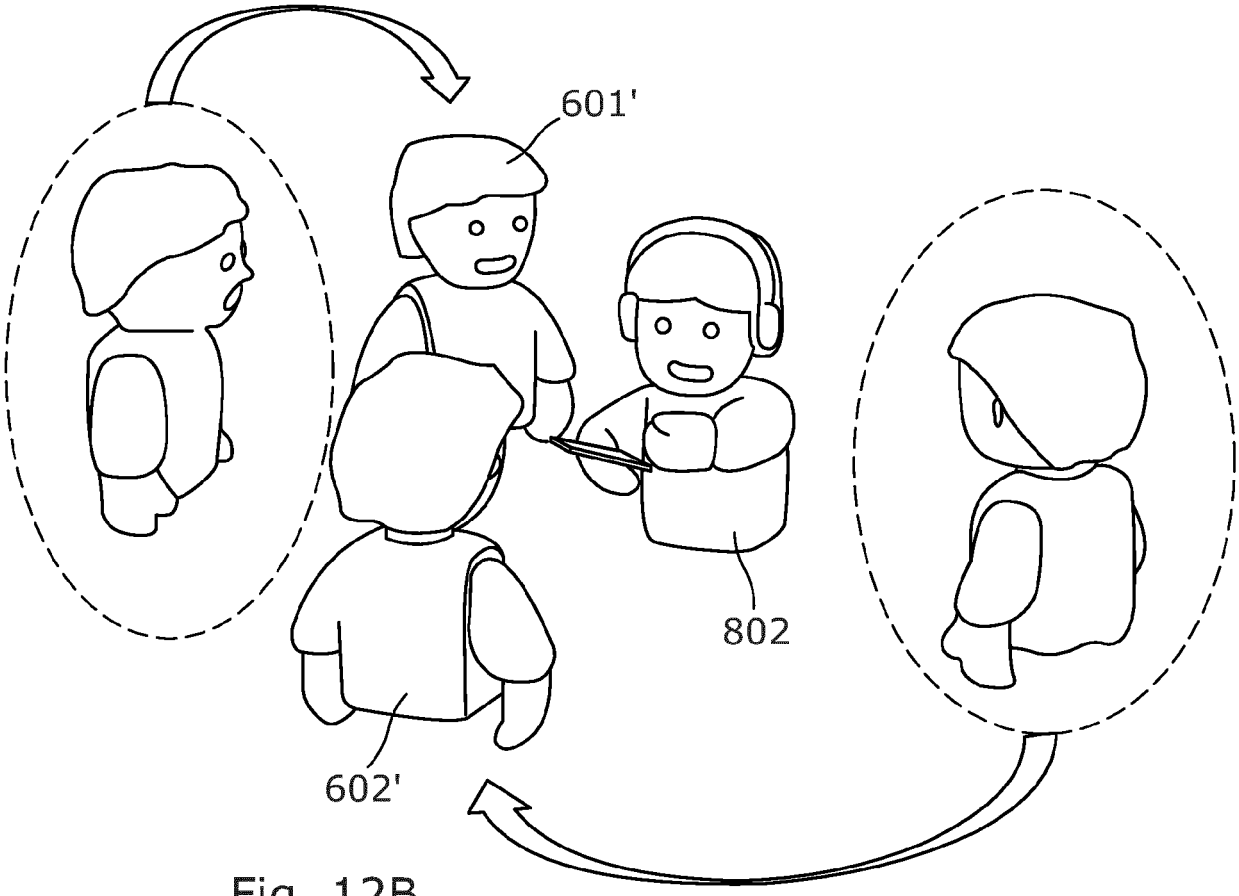

Fig. 12B

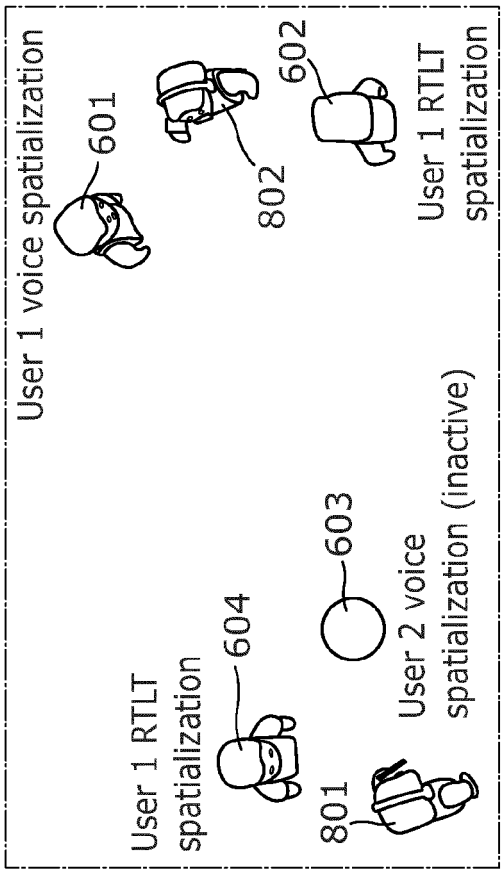

Fig. 13A

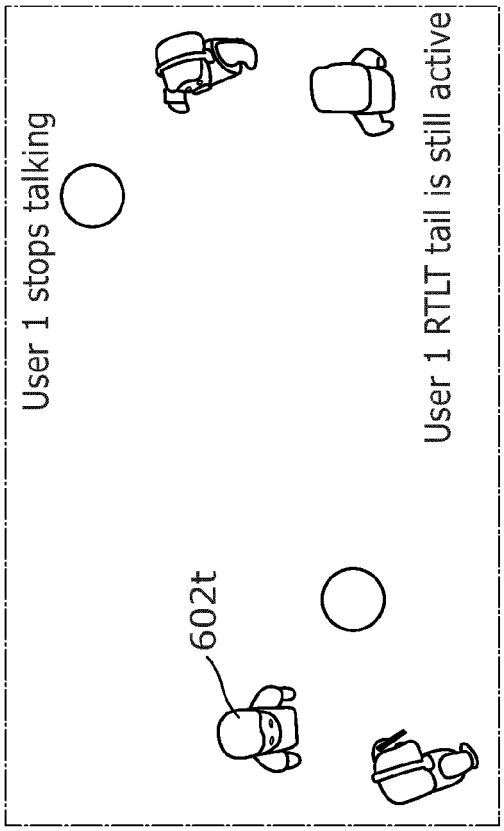

Fig. 13B

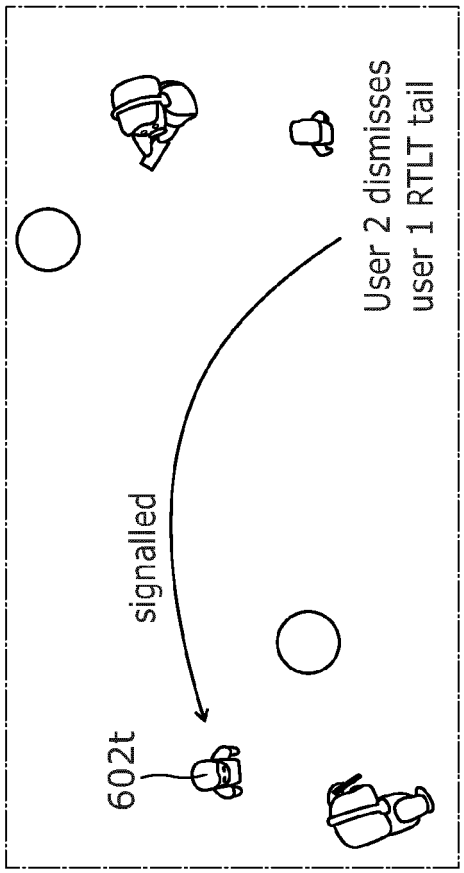

Fig. 13C

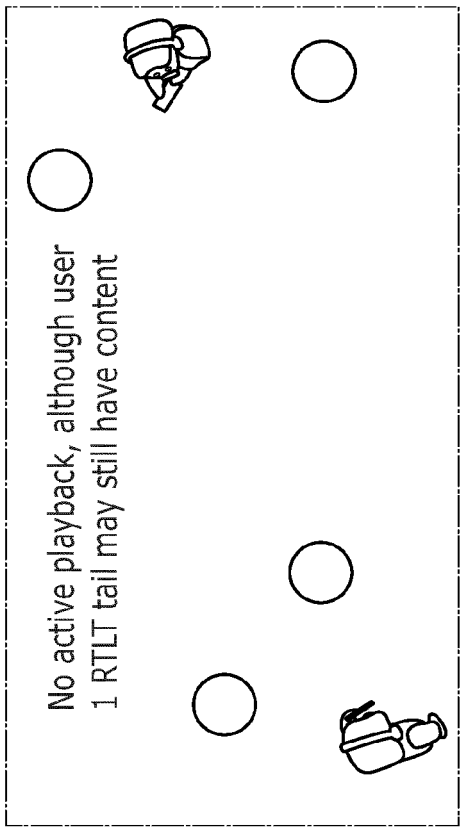

Fig. 13D

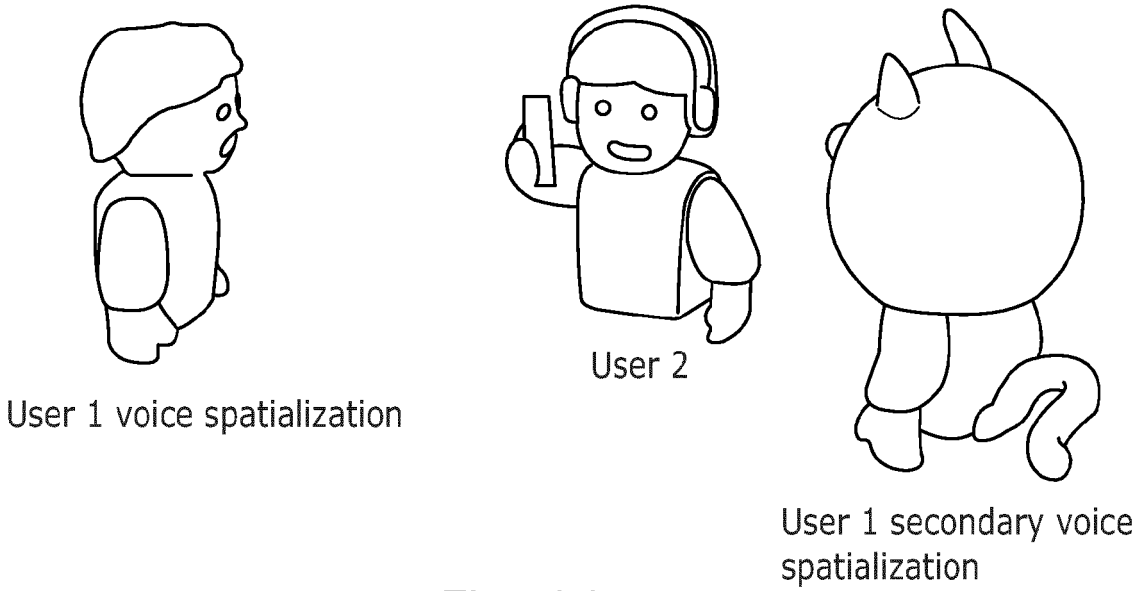
User 1 voice spatialization
User 2
User 1 secondary voice spatialization

Fig. 14

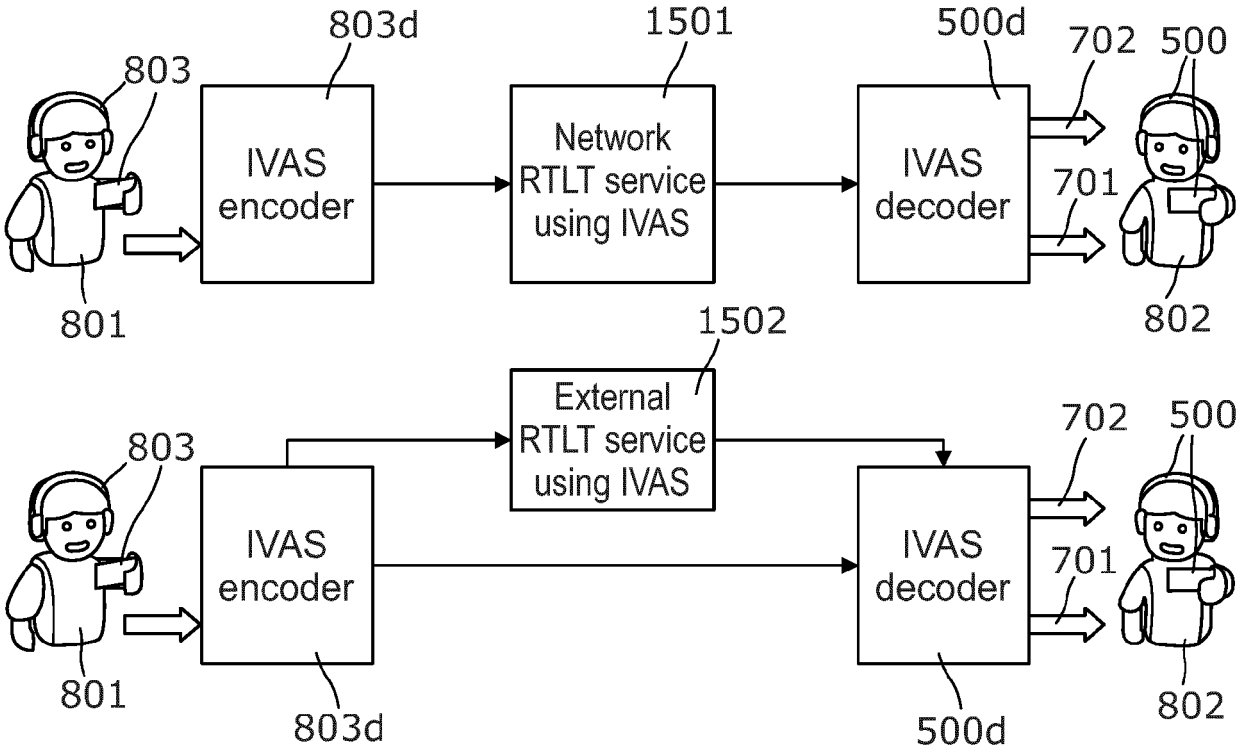
803
803d
IVAS encoder
1501
Network RTLT service using IVAS
500d
IVAS decoder
702
500
701
801
802
1502
External RTLT service using IVAS
803
IVAS encoder
IVAS decoder
702
500
701
801
803d
500d
802

Fig. 15

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 19 16 6572

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	WO 2014/130221 A1 (DOLBY LAB LICENSING CORP [US]) 28 August 2014 (2014-08-28) * the whole document *	1-15	INV. H04S5/00
X	EP 3 293 987 A1 (NOKIA TECHNOLOGIES OY [FI]) 14 March 2018 (2018-03-14) * the whole document *	1-15	
A	US 2007/016401 A1 (EHSANI FARZAD [US] ET AL) 18 January 2007 (2007-01-18) * the whole document *	1-15	

TECHNICAL FIELDS SEARCHED (IPC)
H04S
G06F

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
The Hague	28 June 2019	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 19 16 6572

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.

28-06-2019

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2014130221 A1	28-08-2014	CN 104010265 A	27-08-2014
		EP 2959697 A1	30-12-2015
		US 2015382127 A1	31-12-2015
		WO 2014130221 A1	28-08-2014
EP 3293987 A1	14-03-2018	CN 109691140 A	26-04-2019
		EP 3293987 A1	14-03-2018
		US 2019191264 A1	20-06-2019
		WO 2018050959 A1	22-03-2018
US 2007016401 A1	18-01-2007	NONE	

EPO FORM P0459

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