(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 974 219**

(51) Int. Cl.:

H04R 3/00 (2006.01)
H04S 7/00 (2006.01)
H04N 7/15 (2006.01)
H04M 3/56 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **12.11.2019 PCT/US2019/060855**

(87) Fecha y número de publicación internacional: **22.05.2020 WO20102153**

(96) Fecha de presentación y número de la solicitud europea: **12.11.2019 E 19836164 (4)**

(97) Fecha y número de publicación de la concesión europea: **14.02.2024 EP 3881559**

(54) Título: **Procesamiento de audio en servicios de audio inversivos**

(30) Prioridad:

13.11.2018 US 201862760262 P
17.01.2019 US 201962793666 P
22.01.2019 US 201962795236 P
28.01.2019 US 201962797563 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
26.06.2024

(73) Titular/es:

DOLBY LABORATORIES LICENSING CORPORATION (50.0%)
1275 Market Street
San Francisco, CA 94103, US y
DOLBY INTERNATIONAL AB (50.0%)

(72) Inventor/es:

BRUHN, STEFAN;
TORRES, JUAN FELIX;
MCGRATH, DAVID S. y
LEE, BRIAN

(74) Agente/Representante:

LINAGE GONZÁLEZ, Rafael

DESCRIPCIÓN

Procesamiento de audio en servicios de audio inversivos

Referencia cruzada a solicitudes relacionadas

Esta solicitud reivindica el beneficio de prioridad de las solicitudes de patente provisional de los Estados Unidos núm. 62/760,262 presentada el 13 de noviembre de 2018; núm. 62/793.666 presentada el 17 de enero de 2019, núm. 62/795.236 presentada el 22 de enero de 2019; y núm. 62/797.563 presentada el 28 de enero de 2019.

Campo técnico

La divulgación en el presente documento se relaciona generalmente con la captura, el preprocesamiento acústico, la codificación, la decodificación y el renderizado de audio direccional de una escena de audio. En particular, se refiere a un dispositivo adaptado para modificar una propiedad direccional de un audio direccional capturado en respuesta a datos espaciales de un sistema de micrófono que captura el audio direccional. La divulgación se refiere además a un dispositivo de renderizado configurado para modificar una propiedad direccional de un audio direccional recibido en respuesta a datos espaciales recibidos.

Antecedentes

La introducción del acceso inalámbrico de alta velocidad 4G/5G a las redes de telecomunicaciones, combinada con la disponibilidad de plataformas de hardware cada vez más potentes, ha proporcionado una base para que los servicios multimedia y de comunicaciones avanzados sean desplegados más rápida y fácilmente que nunca.

El códec de servicios de voz mejorados (EVS) del proyecto de asociación de tercera generación (3GPP) ha brindado una mejora muy significativa en la experiencia del usuario con la introducción de codificación de voz y audio de banda súper ancha (SWB) y banda completa (FB), junto con resiliencia a la pérdida paquetes mejorada. Sin embargo, el ancho de banda de audio ampliado es solo una de las dimensiones necesarias para una experiencia verdaderamente inmersiva. Idealmente, se requiere soporte más allá del mono y multimono que ofrece EVS actualmente para sumergir al usuario en un mundo virtual convincente de una manera eficiente en el uso de recursos.

Además, los códecs de audio especificados actualmente en 3GPP proporcionan una calidad y compresión adecuadas para contenido estéreo, pero carecen de las funciones conversacionales (por ejemplo, una latencia suficientemente baja) necesarias para la voz conversacional y las teleconferencias. Estos codificadores también carecen de la funcionalidad multicanal que es necesaria para servicios inmersivos, como transmisión de contenido en vivo y generado por el usuario, realidad virtual (VR) y teleconferencias inmersivas.

Se ha propuesto el desarrollo de una extensión del códec EVS para los servicios inmersivos de voz y audio (IVAS) para llenar este vacío tecnológico y abordar la creciente demanda de servicios multimedia enriquecidos. Además, las aplicaciones de teleconferencia a través de 4G/5G se beneficiarán de un códec IVAS usado como codificador conversacional mejorado que soporta codificación de múltiples transmisiones (por ejemplo, audio basado en canales, objetos y escenas). Los casos de uso para este códec de próxima generación incluyen, entre otros, voz conversacional, teleconferencias de transmisión múltiple, conversación de realidad virtual y transmisión de contenido en vivo y no en vivo generado por el usuario.

Por lo tanto, se espera que IVAS ofrezca experiencias de usuario inmersivas, VR, AR y/o XR. En muchas de estas aplicaciones, un dispositivo (por ejemplo, un teléfono móvil) que captura audio direccional (inmersivo) puede en muchos casos moverse durante la sesión en relación con la escena acústica, provocando una rotación espacial y/o un movimiento de traslación de la escena de audio capturada. Dependiendo del tipo de experiencia proporcionada, por ejemplo, inmersivo, VR, AR o XR y, dependiendo del caso de uso específico, este comportamiento puede ser deseado o no. Por ejemplo, puede resultar perturbador para un oyente si la escena renderizada siempre gira cada vez que gira el dispositivo de captura. En el peor de los casos, puede provocar mareo.

Un ejemplo de un aparato para mezclar al menos dos señales de audio usando metadatos espaciales es el documento WO 2017/182714 A1.

Por tanto, es necesario introducir mejoras en este contexto.

Breve descripción de los dibujos

A continuación se describirán realizaciones de ejemplo con referencia a los dibujos adjuntos, en los que:

La figura 1 muestra un método para codificar audio direccional de acuerdo con realizaciones,

la figura 2 muestra un método para renderizar audio direccional de acuerdo con realizaciones,

la figura 3 muestra un dispositivo codificador configurado para realizar el método de la figura 1 de acuerdo con realizaciones,

la figura 4 muestra un dispositivo de renderizado configurado para realizar el método de la figura 2 de acuerdo con realizaciones,

la figura 5 muestra un sistema que comprende los dispositivos de la figura 3 y la figura 4 de acuerdo con realizaciones,

la figura 6 muestra un escenario de conferencia de realidad virtual física de acuerdo con realizaciones,

la figura 7 muestra un espacio de conferencia virtual de acuerdo con realizaciones.

Todas las figuras son esquemáticas y generalmente solo muestran partes que son necesarias para aclarar la divulgación, mientras que otras partes pueden omitirse o simplemente sugerirse. A menos que se indique lo contrario, números de referencia similares se refieren a partes similares en figuras diferentes.

Descripción detallada

En vista de lo anterior, es por tanto un objeto proporcionar dispositivos y métodos asociados según las reivindicaciones adjuntas para captura, preprocesamiento acústico y/o codificación para compensar movimientos no deseados de la escena sonora espacial que pueden resultar de movimientos inadvertidos de un sistema de micrófono que captura audio direccional. Un objeto adicional es proporcionar un decodificador y/o dispositivo de renderizado correspondiente y métodos asociados para decodificar y renderizar audio direccional según las reivindicaciones adjuntas. Sistemas que comprenden, por ejemplo, el dispositivo codificador y el dispositivo de renderizado también se proporcionan según las reivindicaciones adjuntas.

I. Descripción general: lado de emisión

De acuerdo con un primer aspecto, se proporciona un dispositivo que comprende o está conectado a un sistema de micrófono que comprende uno o más micrófonos para capturar audio. El dispositivo (también denominado en el presente documento lado de emisión o dispositivo de captura) comprende una unidad de recepción configurada para:

- recibir audio direccional capturado por el sistema de micrófono;

- recibir metadatos asociados con el sistema de micrófono, comprendiendo los metadatos datos espaciales del sistema de micrófono, siendo los datos espaciales indicativos de una orientación espacial y/o posición espacial del sistema de micrófono y comprendiendo al menos uno de la lista de: un acimut, un cabeceo, ángulo o ángulos de balanceo y coordenadas espaciales del sistema de micrófono.

En esta divulgación, el término "audio direccional" (sonido direccional) generalmente se refiere a audio inmersivo, es decir, audio capturado por sistemas de micrófonos direccionales que pueden captar sonidos, incluidas las direcciones desde las que llegan. La reproducción de audio direccional permite una experiencia de sonido tridimensional natural (renderizado binaural). El audio, que puede comprender objetos de audio y/o canales (por ejemplo, que representan audio basado en escenas en formato Ambisonics B o audio basado en canales), se asocia así con las direcciones desde las que se recibe. En otras palabras, el audio direccional proviene de las fuentes direccionales e incide desde una dirección de llegada (DOA) representada, por ejemplo, por ángulos de acimut y elevación. Por el contrario, se supone que el sonido ambiental difuso es omnidireccional, es decir, espacialmente invariante o espacialmente uniforme. Otras expresiones que pueden usarse para la característica de "audio direccional" incluyen "audio espacial", "sonido espacial", "audio inmersivo", "sonido inmersivo", "estéreo" y "audio envolvente".

En esta divulgación, el término "coordenadas espaciales" generalmente se refiere a la posición espacial del sistema de micrófono o el dispositivo de captura en el espacio. Las coordenadas cartesianas son una realización de las coordenadas espaciales. Otros ejemplos incluyen coordenadas cilíndricas o esféricas. Cabe señalar que la posición en el espacio puede ser relativa (por ejemplo, coordenadas en una habitación o relativa a otro dispositivo/unidad, etc.) o absoluta (por ejemplo, coordenadas GPS o similar).

En esta divulgación, los "datos espaciales" generalmente indican una orientación rotacional y/o posición espacial actual del sistema de micrófono o un cambio en la orientación rotacional y/o posición espacial en comparación con una orientación/posición anterior del sistema de micrófono.

De este modo, el dispositivo recibe metadatos que comprenden datos espaciales indicativos de una orientación espacial y/o posición espacial del sistema de micrófono que captura el audio direccional.

El dispositivo comprende además una unidad informática configurada para: modificar al menos parte del audio direccional para producir audio direccional modificado, mediante el cual se modifica una propiedad direccional del audio en respuesta a la orientación espacial y/o la posición espacial del sistema de micrófono.

La modificación se puede realizar usando cualquier medio adecuado, por ejemplo, definiendo una matriz de rotación/traslación basada en los datos espaciales, y multiplicando el audio direccional con esta matriz para lograr el audio direccional modificado. La multiplicación de matrices es adecuada para audio espacial no paramétrico. El audio espacial paramétrico se puede modificar ajustando los metadatos espaciales como, por ejemplo, los parámetros direccionales del objeto u objetos de sonido.

El audio direccional modificado luego se codifica en datos de audio digitales, datos que se transmiten mediante una unidad de transmisión del dispositivo.

Los inventores se han dado cuenta de que los movimientos de rotación/traslación del dispositivo de captura de sonido (sistema de micrófono) se compensan mejor en el extremo de emisión, es decir, en el extremo de captura del audio. Es probable que esto permita la mejor estabilización posible de la escena de audio capturada con respecto a, por ejemplo, movimientos no deseados. Tal compensación puede ser parte del proceso de captura, es decir, durante el preprocesamiento acústico, o como parte de la etapa de codificación IVAS. Además, al realizar la compensación en el extremo de emisión, se reduce la necesidad de transmitir los datos espaciales desde el extremo de emisión al extremo de recepción. En caso de que la compensación de los movimientos de rotación/traslación del dispositivo de captura de sonido tuviera que realizarse en el receptor del audio, los datos espaciales completos tendría que transmitirse al extremo de recepción. Suponiendo que las coordenadas de rotación en los tres ejes se representan con 8 bits cada una y se estiman y transmiten a una velocidad de 50 Hz, la velocidad de bits resultante sería de 1,2 kbps. Se puede hacer la suposición analógica para las coordenadas espaciales del sistema de micrófono.

De acuerdo con algunas realizaciones, la orientación espacial del sistema de micrófono se representa con parámetros que describen el movimiento/orientación rotacional con un grado de libertad, DoF, en los datos espaciales. Por ejemplo, puede ser suficiente considerar únicamente el ángulo de acimut para las conferencias telefónicas.

De acuerdo con algunas realizaciones, la orientación espacial del sistema de micrófono se representa con parámetros que describen la orientación/movimiento rotacional con tres grados de libertad, DoF, en los datos espaciales.

De acuerdo con algunas realizaciones, los datos espaciales del sistema de micrófono se representan en seis DoF. En esta realización, los datos espaciales del sistema de micrófono capturan una posición cambiada (denominada en el presente documento coordenadas espaciales) del sistema de micrófono como traslación hacia adelante/atrás (sobretensión), arriba/abajo (alzado), izquierda/derecha (oscilación) en tres ejes perpendiculares, combinados con cambios en la orientación (u orientación rotacional actual) del sistema de micrófono a través de la rotación alrededor de tres ejes perpendiculares, a menudo denominados guiñada o azimut (eje normal/vertical), cabeceo (eje transversal) y balanceo (eje longitudinal).

De acuerdo con algunas realizaciones, el audio direccional recibido comprende audio que comprende metadatos direccionales. Por ejemplo, tal audio puede comprender objetos de audio, es decir, audio basado en objetos (OBA). OBA es una forma paramétrica de audio espacial/direccional con metadatos espaciales. Una forma particular de audio espacial paramétrico es el audio espacial asistido por metadatos (MASA).

De acuerdo con algunas realizaciones, la unidad informática está configurada además para codificar al menos partes de los metadatos que comprenden datos espaciales del sistema de micrófono en dichos datos de audio digitales. Ventajosamente, esto permite la compensación del ajuste direccional realizado en el audio capturado en el extremo de recepción. Sujeto a la definición de un marco de referencia de rotación adecuado, por ejemplo, con el eje z correspondiente a la dirección vertical, en muchos casos puede ser necesario transmitir simplemente el ángulo de acimut (por ejemplo, 400 bps). Es posible que los ángulos de cabeceo y balanceo del dispositivo de captura en el marco de referencia de rotación solo sean necesarios en determinadas aplicaciones de realidad virtual. Compensando los datos espaciales del sistema de micrófono en el lado de emisión, e incluyendo condicionalmente al menos partes de los datos espaciales en los datos de audio digitales codificados, el caso en el que la escena acústica renderizada debe ser invariante de la posición del dispositivo de captura y se admiten ventajosamente el resto de casos en los que la escena acústica renderizada deba girar con los movimientos correspondientes del dispositivo de captura.

De acuerdo con algunas realizaciones, la unidad de recepción está configurada además para recibir primeras instrucciones que indican a la unidad informática si se deben incluir dichas al menos partes de los metadatos que

comprenden datos espaciales del sistema de micrófono en dichos datos de audio digitales, por lo que la unidad informática actúa en consecuencia. En consecuencia, el lado de emisión incluye condicionalmente partes de los datos espaciales en los datos de audio digitales, para ahorrar velocidad de bits cuando sea posible. La instrucción puede recibirse más de una vez durante una sesión, de modo que si los datos espaciales (o partes de estos) deben incluirse o no en los datos de audio digitales cambia con el tiempo. Es decir, pueden existir adaptaciones en sesión donde las primeras instrucciones puedan ser recibidas por el dispositivo tanto de forma continua como discontinua. Continua sería, por ejemplo, ser una vez en cada marco. Discontinua podría ser solo una vez que se deba dar una nueva instrucción. También existe la posibilidad de recibir la primera instrucción solo una vez en la configuración de una sesión.

De acuerdo con algunas realizaciones, la unidad de recepción está configurada además para recibir segundas instrucciones que indican a la unidad informática qué parámetro o parámetros de los datos espaciales del sistema de micrófono incluir en los datos de audio digitales, por lo que la unidad informática actúa en consecuencia. Como se ejemplificó anteriormente, se puede indicar al lado de emisión que incluya solo el acimut o que incluya todos los datos que definen la orientación espacial del sistema de micrófono. La instrucción puede recibirse más de una vez durante una sesión de modo que el número de parámetros incluidos en los datos de audio digitales cambie con el tiempo. En otras palabras, puede haber adaptaciones en sesión donde las segundas instrucciones pueden ser recibidas por el dispositivo tanto de forma continua como discontinua. Continua sería, por ejemplo, ser una vez en cada marco. Discontinua podría ser solo una vez que se deba dar una nueva instrucción. También existe la posibilidad de recibir la segunda instrucción solo una vez en la configuración de una sesión.

De acuerdo con algunas realizaciones, la unidad de transmisión está configurada para transmitir los datos de audio digitales a un dispositivo adicional, en el que se reciben indicaciones sobre la primera y/o la segunda instrucción desde dicho dispositivo adicional. En otras palabras, el lado de recepción (que comprende un renderizador para renderizar el audio decodificado recibido) puede, dependiendo del contexto, indicar al lado de emisión si debe incluir parte de los datos espaciales o no en los datos de audio digitales, y/o qué parámetros incluir. En otras realizaciones, se pueden recibir indicaciones sobre la primera y/o la segunda instrucción desde, por ejemplo, una unidad coordinadora (servidor de llamadas) para una conferencia de audio/vídeo inmersiva multiusuario, o cualquier otra unidad que no esté directamente involucrada en el renderizado del audio direccional.

De acuerdo con algunas realizaciones, la unidad de recepción está configurada además para recibir metadatos que comprenden una marca de tiempo que indica un tiempo de captura del audio direccional, en donde la unidad informática está configurada para codificar dicha marca de tiempo en dichos datos de audio digitales. Ventajosamente, esta marca de tiempo se puede usar para sincronizar en un lado de recepción, por ejemplo, sincronizar en el renderizador de audio con el renderizador de vídeo, o sincronizar una pluralidad de datos de audio digitales recibidos desde diferentes dispositivos de captura.

De acuerdo con algunas realizaciones, la codificación de las señales de audio modificadas comprende mezclar de manera descendente el audio direccional modificado, en donde la mezcla descendente se realiza teniendo en cuenta la orientación espacial del sistema de micrófono, y codificar la mezcla descendente y una matriz de mezcla descendente usada en la mezcla descendente en dichos datos de audio digitales. Por ejemplo, la formación de haces acústicos hacia una fuente direccional específica del audio direccional se adapta ventajosamente basándose en la modificación direccional realizada en el audio direccional.

De acuerdo con algunas realizaciones, el dispositivo se implementa en un equipo de realidad virtual, VR o equipo de realidad aumentada, AR, que comprende el sistema de micrófono y un dispositivo de seguimiento de la cabeza configurado para determinar datos espaciales del dispositivo en 3 a 6 DoF. En otras realizaciones, el dispositivo se implementa en un teléfono móvil que comprende un sistema de micrófono.

II. Descripción general: lado de recepción

De acuerdo con un segundo aspecto, se proporciona un dispositivo para renderizar señales de audio. El dispositivo (también denominado en el presente documento lado de recepción o dispositivo de renderizado) comprende una unidad de recepción configurada para recibir datos de audio digitales. El dispositivo comprende además una unidad de decodificación configurada para decodificar los datos de audio digitales recibidos en audio direccional y en metadatos, comprendiendo los metadatos datos espaciales al menos uno de la lista de: acimut, cabeceo, ángulo o ángulos de balanceo y coordenadas espaciales. Los datos espaciales pueden recibirse, por ejemplo, en forma de parámetros, por ejemplo, los 3 ángulos DoF. En otras realizaciones, los datos espaciales pueden recibirse como una matriz de rotación/traslación.

El dispositivo comprende además una unidad de renderizado configurada para:

modificar una propiedad direccional del audio direccional usando los datos espaciales de rotación; y

renderizar el audio direccional modificado.

Ventajosamente, el dispositivo de acuerdo con este aspecto puede modificar el audio direccional como se indica en los metadatos. Por ejemplo, los movimientos de un dispositivo que captura el audio pueden considerarse durante el renderizado.

De acuerdo con algunas realizaciones, los datos espaciales indican la orientación espacial y/o la posición espacial de un sistema de micrófono que comprende uno o más micrófonos que capturan el audio direccional, en donde la unidad de renderizado modifica la propiedad direccional del audio direccional para reproducir al menos parcialmente un entorno de audio del sistema de micrófono. En esta realización, el dispositivo aplica la rotación de la escena acústica reaplicando al menos partes de la rotación de la escena acústica (relativa, es decir, la rotación de la escena es relativa al sistema de micrófono en movimiento) que fue compensada en el dispositivo de captura.

De acuerdo con algunas realizaciones, los datos espaciales comprenden parámetros que describen el movimiento/orientación rotacional con un grado de libertad, DoF.

De acuerdo con algunas realizaciones, los datos espaciales comprenden parámetros que describen el movimiento/orientación rotacional con tres grados de libertad, DoF.

De acuerdo con algunas realizaciones, el audio direccional decodificado comprende audio que comprende metadatos direccionales. Por ejemplo, el audio direccional decodificado puede comprender objetos de audio, es decir, audio basado en objetos (OBA). En otras realizaciones, el audio direccional decodificado puede estar basado en canales, por ejemplo, que representa audio basado en escenas en formato Ambisonics B o audio basado en canales.

De acuerdo con algunas realizaciones, el dispositivo comprende una unidad de transmisión configurada para transmitir instrucciones a un dispositivo adicional desde el cual se recibe el audio digital, indicando las instrucciones al dispositivo adicional qué parámetro o parámetros (si los hay) deben comprender los datos de rotación. En consecuencia, el dispositivo de renderizado puede ordenar al dispositivo de captura que transmita, por ejemplo, solo parámetros de rotación, solo el parámetro de azimut o 6 parámetros DoF completos, dependiendo del caso de uso y/o el ancho de banda disponible. Además, el dispositivo de renderizado puede tomar esta decisión basándose en los recursos informáticos disponibles en el renderizador para aplicar la rotación acústica de la escena, o el nivel de complejidad de la unidad de renderizado. Las instrucciones pueden transmitirse más de una vez durante una sesión y, por tanto, cambiar con el tiempo, es decir, basándose en lo anterior. En otras palabras, pueden existir adaptaciones en sesión donde el dispositivo puede transmitir las instrucciones tanto de forma continua como discontinua. Continua sería, por ejemplo, una vez en cada marco. Discontinua podría ser solo una vez que se deba dar una nueva instrucción. También existe la posibilidad de transmitir la instrucción solo una vez en la configuración de una sesión.

De acuerdo con algunas realizaciones, la unidad de decodificación está configurada además para extraer una marca de tiempo que indica un tiempo de captura del audio direccional a partir de los datos de audio digitales. Esta marca de tiempo se puede usar por motivos de sincronización como se discutió anteriormente.

De acuerdo con algunas realizaciones, la decodificación de los datos de audio digitales recibidos en audio direccional mediante la unidad de decodificación comprende:

decodificar los datos de audio digitales recibidos en audio mezclado de manera descendente,

mezclar de manera ascendente, mediante la unidad de decodificación, el audio mezclado de manera descendente en el audio direccional usando una matriz de mezcla descendente incluida en los datos de audio digitales recibidos.

De acuerdo con algunas realizaciones, los datos espaciales incluyen coordenadas espaciales y en donde la unidad de renderizado está configurada además para ajustar un volumen del audio renderizado basándose en las coordenadas espaciales. En esta realización, el volumen del audio recibido desde "lejos" puede atenuarse en comparación con el audio recibido desde una ubicación más cercana. Cabe señalar que la cercanía relativa del audio recibido se puede determinar basándose un espacio virtual, donde la posición del dispositivo de captura en este espacio con relación al dispositivo de recepción se determina basándose en las coordenadas espaciales de los dispositivos, aplicando una métrica de distancia adecuada, por ejemplo, métrica euclidiana. Un paso adicional puede implicar el uso de algún esquema de mapeo arbitrario para determinar a partir de la distancia métrica parámetros de renderización de audio, tales como un nivel de sonido. Ventajosamente, en esta realización, se puede mejorar la experiencia de inmersión del audio renderizado.

De acuerdo con algunas realizaciones, el dispositivo se implementa en un equipo de realidad virtual, VR, o equipo de realidad aumentada, AR, que comprende un dispositivo de seguimiento de la cabeza configurado para

medir la orientación espacial y la posición espacial del dispositivo en seis DoF. En esta realización, también se pueden usar los datos espaciales del dispositivo de renderizado al modificar una propiedad direccional del audio direccional. Por ejemplo, la matriz de rotación/traslación recibida se puede multiplicar con una matriz similar que defina, por ejemplo. el estado de rotación del dispositivo de renderizado, y la matriz resultante puede usarse entonces para modificar la propiedad direccional del audio direccional. Ventajosamente, en esta realización, se puede mejorar la experiencia de inmersión del audio renderizado. En otras realizaciones, el dispositivo se implementa en un dispositivo de conferencia telefónica o similar, que se supone que es estacionario y en el que se ignora cualquier estado de rotación del dispositivo.

De acuerdo con algunas realizaciones, la unidad de renderizado está configurada para el renderizado de audio binaural.

III. Descripción general: sistema

De acuerdo con un tercer aspecto, se proporciona un sistema que comprende:

un primer dispositivo de acuerdo con el primer aspecto configurado para transmitir datos de audio digitales a un segundo dispositivo de acuerdo con el segundo aspecto, en el que el sistema está configurado para audio y/o videoconferencia.

De acuerdo con algunas realizaciones, el primer dispositivo comprende además una unidad de grabación de video y está configurado para codificar video grabado en datos de video digital y transmitir los datos de video digital al segundo dispositivo, en donde el segundo dispositivo comprende además un visualizador para visualizar datos de video digital decodificados.

De acuerdo con un cuarto aspecto, se proporciona un sistema que comprende:

un primer dispositivo de acuerdo con el primer aspecto configurado para transmitir datos de audio digitales a un segundo dispositivo, comprendiendo el segundo dispositivo:

una unidad de recepción configurada para recibir datos de audio digitales,

una unidad de decodificación configurada para:

decodificar los datos de audio digitales recibidos en audio direccional y en metadatos, comprendiendo los metadatos datos espaciales que comprenden al menos uno de la lista de: acimut, cabeceo, ángulo o ángulos de balanceo y coordenadas espaciales;

una unidad de renderizado para renderizar audio;

en el que la unidad de renderizado está configurada para, cuando el segundo dispositivo recibe además datos de vídeo codificados desde el primer dispositivo:

modificar una propiedad direccional del audio direccional usando los datos espaciales, y

renderizar el audio direccional modificado;

en el que la unidad de renderizado está configurada para, cuando el segundo dispositivo no recibe datos de vídeo codificados desde el primer dispositivo:

renderizar el audio direccional.

Ventajosamente, la decisión de reproducir un entorno de audio del sistema de micrófono compensando la orientación espacial y/o la posición espacial del sistema de micrófono se toma basándose en si el vídeo se transmite o no. En esta realización, es posible que el dispositivo de emisión no siempre sea consciente de cuándo es necesaria o deseable la compensación de su movimiento. Consideremos, por ejemplo, la situación en la que el audio es renderizado junto con el vídeo. En ese caso, al menos cuando la captura de vídeo se realiza con el mismo dispositivo que captura el audio, puede ser posible ventajosamente rotar la escena de audio junto con la escena visual en movimiento o mantener estable la escena de audio. Mantener estable la escena de audio compensando los movimientos del dispositivo de captura puede ser la opción preferida si no se consume vídeo.

De acuerdo con un quinto aspecto, se proporciona un medio no transitorio legible por ordenador que almacena instrucciones que, cuando son ejecutadas por uno o más procesadores, hacen que uno o más procesadores realicen operaciones de cualquiera de los aspectos uno a cuatro.

IV. Descripción general - Generalmente

Los aspectos segundo a quinto generalmente pueden tener características y ventajas iguales o correspondientes al primer aspecto.

Otros objetivos, características y ventajas de la presente invención aparecerán a partir de la siguiente divulgación detallada, de las reivindicaciones dependientes adjuntas así como de los dibujos.

Los pasos de cualquier método, o un dispositivo que implemente una serie de pasos, divulgados en el presente documento no tienen que realizarse en el orden exacto divulgado, a menos que se indique explícitamente.

V - Ejemplos de realización

Se espera que los servicios de voz y audio inmersivos ofrezcan experiencias de usuario inmersivas y de realidad virtual (VR). También se pueden ofrecer experiencias de realidad aumentada (AR) y realidad extendida (XR). Esta divulgación trata del hecho de que los dispositivos móviles como los UE portátiles que capturan una escena inmersiva o AR/VR/XR pueden en muchos casos estar moviéndose durante la sesión en relación con la escena acústica. Destaca los casos en los que se debe evitar que los movimientos de rotación del dispositivo de captura se reproduzcan como la correspondiente rotación de la escena renderizada por el dispositivo de recepción. Esta divulgación se relaciona con cómo lo anterior puede manejarse eficientemente para cumplir con los requisitos que tiene el usuario sobre audio inmersivo dependiendo del contexto.

Cabe señalar que, si bien algunos ejemplos en el presente documento se describirán en el contexto de un codificador, decodificador y/o renderizador IVAS, cabe señalar que este es simplemente un tipo de codificador/decodificador/renderizador en el que se pueden aplicar los principios generales de la invención, y que puede haber muchos otros tipos de codificadores, decodificadores y renderizadores que se pueden usar junto con las diversas realizaciones descritas en el presente documento.

También se debe tener en cuenta que si bien los términos "mezcla ascendente" y "mezcla descendente" se usan en todo este documento, es posible que no impliquen necesariamente aumentar y reducir, respectivamente, el número de canales. Si bien este puede ser el caso a menudo, debe tenerse en cuenta que cualquiera de los términos puede referirse a reducir o aumentar el número de canales. Por tanto, ambos términos caen bajo el concepto más general de "mezcla".

Volviendo ahora a la figura 1, se describe un método 1 para codificar y transmitir una representación de audio direccional, de acuerdo con una realización. Un dispositivo 300 configurado para realizar el método 1 se muestra en la figura 3.

El dispositivo 300 puede ser generalmente un teléfono móvil (teléfono inteligente), sin embargo, el dispositivo también puede ser parte de un equipo VR/AR/XR o cualquier otro tipo de dispositivo que comprenda o esté conectado a un sistema 302 de micrófono que comprenda uno o más micrófonos para capturar audio de dirección. El dispositivo 300 puede así comprender el sistema 302 de micrófono o estar conectado (por cable o inalámbrico) a un sistema 302 de micrófono ubicado remotamente. En algunas realizaciones, el dispositivo 300 se implementa en un equipo VR o equipo AR que comprende el sistema 302 de micrófono y un dispositivo de seguimiento de la cabeza configurado para determinar datos espaciales del dispositivo en 1 a 6 DoF.

En algunos escenarios de captura de audio, una posición y/o la orientación espacial del sistema 302 de micrófono pueden estar cambiando durante la captura del audio direccional.

A continuación se describirán dos escenarios de ejemplo.

Un cambio de posición y/u orientación espacial del sistema 302 de micrófono durante la captura de audio puede causar rotación/traslación espacial de la escena renderizada en un dispositivo de recepción. Dependiendo del tipo de experiencia proporcionada, por ejemplo, inmersiva, VR, AR o XR y, dependiendo del caso de uso específico, este comportamiento puede ser deseado o no. Un ejemplo en el que esto puede desearse es cuando el servicio proporciona adicionalmente un componente visual y donde la cámara de captura (por ejemplo, captura de vídeo de 360 grados, no mostrada en la figura 1) y los micrófonos 302 están integrados en el mismo dispositivo. En ese caso, debería esperarse que una rotación del dispositivo de captura dé como resultado una rotación correspondiente de la escena audiovisual renderizada.

Por otro lado, si la captura audiovisual no se realiza mediante el mismo dispositivo físico o en caso de que no haya un componente de vídeo, puede resultar molesto para el oyente si la escena renderizada gira cada vez que gira el dispositivo de captura. En el peor de los casos, puede provocar mareo. Por tanto, es deseable compensar los cambios posicionales (traslación y/o rotaciones) del dispositivo de captura. Los ejemplos incluyen telefonía inmersiva y aplicaciones de conferencias inmersivas que usan un teléfono inteligente como dispositivo de captura (es decir, que comprende el conjunto de micrófonos 302). En estos casos de uso, puede suceder frecuentemente que el conjunto de micrófonos se mueva sin querer ya sea porque es de mano o porque el usuario lo toca

durante el funcionamiento. Es posible que el usuario del dispositivo de captura no sea consciente de que moverlo puede provocar inestabilidades en el audio espacial renderizado en los dispositivos de recepción. Por lo general, no se puede esperar que el usuario mantenga quieto el teléfono mientras se encuentra en una situación de conversación.

Los métodos y dispositivos descritos a continuación se definen para algunos o todos los escenarios descritos anteriormente.

Por lo tanto, el dispositivo 300 comprende o está conectado a un sistema 302 de micrófono que comprende uno o más micrófonos para capturar audio. El sistema de micrófono puede así comprender 1, 2, 3, 5, 10, etc., micrófonos. En algunas realizaciones, el sistema de micrófono comprende una pluralidad de micrófonos. El dispositivo 300 comprende una pluralidad de unidades funcionales. Las unidades pueden implementarse en hardware y/o software y pueden comprender uno o más procesadores para manejar la funcionalidad de las unidades.

El dispositivo 300 comprende una unidad 304 de recepción que está configurada para recibir audio direccional S13 320 capturado por el sistema 302 de micrófono. El audio direccional 320 es preferiblemente una representación de audio que permite fácilmente la rotación y/o traslación de la escena de audio. El audio direccional 320 puede comprender, por ejemplo, objetos y/o canales de audio que permitan la rotación y/o traslación de la escena de audio. El audio direccional puede comprender

- audio basado en canales (CBA) como estéreo, multicanal/envolvente, 5.1, 7.1, etc.

- audio basado en escenas (SBA), como Ambisonics de primer orden y de orden superior.

- audio basado en objetos (OBA).

CBA y SBA son formas no paramétricas de audio espacial/direccional, mientras que OBA es paramétrica con metadatos espaciales. Una forma particular de audio espacial paramétrico es el audio espacial asistido por metadatos (MASA).

La unidad 304 de recepción está configurada además para recibir metadatos 322 de S14 asociados con el sistema 302 de micrófono. Los metadatos 322 comprenden datos espaciales del sistema 302 de micrófono. Los datos espaciales son indicativos de una orientación espacial y/o posición espacial del sistema 302 de micrófono. Los datos espaciales del sistema de micrófono comprenden al menos uno de la lista de: acimut, cabeceo, ángulo o ángulos de balanceo y coordenadas espaciales del sistema de micrófono. Los datos espaciales se pueden representar en 1 grado de libertad, DoF (por ejemplo, solo el ángulo de acimut del sistema de micrófono), tres DoF (por ejemplo, la orientación espacial del sistema de micrófono en 3 DoF) o en seis DoF (ambas con orientación espacial en 3 DoF y posición espacial en 3 DoF). Por supuesto, los datos espaciales pueden representarse en cualquier DoF del uno al seis.

El dispositivo 300 comprende además una unidad informática 306 que recibe el audio direccional 320 y los metadatos 322 desde la unidad 304 de recepción y modifica S15 al menos parte del audio direccional 320 (por ejemplo, al menos algunos de los objetos de audio del audio direccional) para producir audio direccional modificado. Esta modificación da como resultado que se modifica una propiedad direccional del audio en respuesta a la orientación espacial y/o posición espacial del sistema de micrófono.

La unidad informática 306 codifica S16 entonces datos digitales codificando S17 el audio direccional modificado en datos 328 de audio digital. El dispositivo 300 comprende además una unidad 310 de transmisión configurada para transmitir (por cable o inalámbrica) los datos 328 de audio digital, por ejemplo, como un flujo de bits.

Al compensar los movimientos de rotación y/o traslación del sistema 302 de micrófono ya en el dispositivo 300 de codificación (también puede denominarse dispositivo de emisión, dispositivo de captura, dispositivo de transmisión, lado de emisión), disminuyen los requisitos para transmitir los datos espaciales del sistema 302 de micrófono. está relajado. Si dicha compensación la realizara un dispositivo que recibe el audio direccional codificado (por ejemplo, un renderizador de audio inmersivo), todos los metadatos requeridos siempre tendrían que incluirse en los datos 328 de audio digital. Suponiendo que las coordenadas de rotación del sistema 302 de micrófono en los tres ejes se representan con 8 bits cada uno y se estiman y transmiten a una velocidad de 50 Hz, el aumento resultante en la velocidad de bits de la señal 332 sería de 1,2 kbps. Además, es probable que las variaciones de la escena auditiva en caso de que no haya compensación de movimiento en el lado de captura puedan hacer que la codificación de audio espacial sea más exigente y potencialmente menos eficiente.

Además, como la información subyacente a la decisión de modificación está fácilmente disponible en el dispositivo 300, es apropiado compensar los movimientos de rotación/traslación del sistema 302 de micrófono ya aquí, lo que de este modo se puede realizar de manera eficiente. De este modo se puede reducir el retardo algorítmico máximo para esta operación.

Otra ventaja más es que al compensar siempre (en lugar de condicionalmente, a petición) los movimientos de rotación/traslación en el dispositivo 300 de captura y proporcionar condicionalmente a los extremos de recepción datos de orientación espacial del sistema de captura, se evitan conflictos potenciales si se sirven múltiples puntos finales con diferentes necesidades de renderizado, como en casos de uso de conferencias con múltiples interlocutores.

Lo anterior cubre todos los casos en los que la escena acústica renderizada debe ser invariante con respecto a la posición y rotación del sistema 302 de micrófono que captura el audio direccional. Para abordar los casos restantes en los que la escena acústica renderizado debe girar con los movimientos correspondientes del sistema 302 de micrófono, la unidad informática 306 puede configurarse opcionalmente para codificar S18 al menos partes de los metadatos 322 que comprenden datos espaciales del sistema de micrófono en dichos datos 328 de audio digital. Por ejemplo, sujeto a la definición de un marco de referencia de rotación adecuado, por ejemplo, con el eje z correspondiente a la dirección vertical, en muchos casos puede ser necesario transmitir simplemente el ángulo de acimut (por ejemplo, 400 bps). Es posible que los ángulos de cabeceo y balanceo del sistema 302 de micrófono en el marco de referencia de rotación solo sean necesarios en ciertas aplicaciones VR.

Los parámetros de rotación/traslación proporcionados condicionalmente típicamente pueden transmitirse como un elemento condicional del formato de carga útil IVAS RTP. Por tanto, estos parámetros requerirán una pequeña porción del ancho de banda asignado.

Para cumplir con los diferentes escenarios, la unidad 304 de recepción puede configurarse opcionalmente para recibir S10 instrucciones sobre cómo manejar los metadatos 322 cuando la unidad informática 306 está codificando los datos 328 de audio digital. Las instrucciones pueden recibirse S10 desde un dispositivo de renderizado (por ejemplo, otra parte de la audioconferencia) o desde un dispositivo de coordinación tal como un servidor de llamadas o similar.

En algunas realizaciones, la unidad 304 de recepción está configurada además para recibir S11 primeras instrucciones que indican a la unidad informática 306 si se deben incluir dichas al menos partes de los metadatos 322 que comprenden datos espaciales del sistema de micrófono en dichos datos de audio digitales. En otras palabras, las primeras instrucciones informan al dispositivo 300 si alguno o ninguno de los metadatos debe incluirse en los datos 328 de audio digital. Por ejemplo, si el dispositivo 300 está transmitiendo los datos 328 de audio digital como parte de una audioconferencia, las primeras instrucciones pueden definir que no se debe incluir ninguna parte de los metadatos 322.

Alternativamente, o adicionalmente, en algunas realizaciones, la unidad 304 de recepción está configurada además para recibir segundas instrucciones que indican a la unidad informática qué parámetro o parámetros de los datos espaciales del sistema de micrófono incluir en los datos de audio digitales, por lo que la unidad informática actúa respectivamente. Por ejemplo, por razones de ancho de banda u otras razones, las segundas instrucciones pueden definir a la unidad informática 306 para incluir solo el ángulo de acimut en los datos 328 de audio digital.

La primera y/o la segunda instrucción típicamente pueden estar sujetas a negociación de configuración de sesión. Por lo tanto, ninguna de estas instrucciones requiere transmisiones durante la sesión y no requerirá nada del ancho de banda asignado, por ejemplo, la conferencia de audio/video inmersiva.

Como se mencionó anteriormente, el dispositivo 300 puede ser parte de una videoconferencia. Por esta razón, la unidad 304 de recepción puede configurarse además para recibir metadatos (no mostrados en la figura 1) que comprenden una marca de tiempo que indica un tiempo de captura del audio direccional, en donde la unidad informática 306 está configurada para codificar dicha marca de tiempo en dichos datos de audio digitales. Ventajosamente, el audio direccional modificado se puede sincronizar con el vídeo capturado en el lado de renderizado.

En algunas realizaciones, la codificación S17 del audio direccional modificado comprende mezclar de manera descendente el audio direccional modificado, en donde la mezcla descendente se realiza teniendo en cuenta la orientación espacial del sistema 302 de micrófono y codificar la mezcla descendente y una matriz de mezcla descendente usada en la mezcla descendente en dichos datos 328 de audio digital. La mezcla descendente puede comprender, por ejemplo, ajustar una operación de formación de haces del audio direccional 320 basándose en los datos espaciales del sistema 302 de micrófono.

Los datos de audio digitales se transmiten S19 así desde el dispositivo 300 como parte de transmisión de, por ejemplo, un escenario de audio/videoconferencia inmersivo. Los datos de audio digitales luego son recibidos por un dispositivo para renderizar señales de audio, por ejemplo, una parte de recepción del escenario de audio/videoconferencia inmersivo. El dispositivo 400 de renderizado se describirá ahora junto con las figuras 2 y 4.

El dispositivo 400 que renderiza señales de audio comprende una unidad de recepción 402 configurada para recibir S21 datos 328 de audio digital (por cable o inalámbricos).

El dispositivo 400 comprende además una unidad 404 de decodificación configurada para decodificar S22 los datos 328 de audio digital recibidos en audio direccional 420 y en metadatos 422, comprendiendo los metadatos 422 datos espaciales que comprenden al menos uno de la lista de: azimut, cabeceo, ángulo o ángulos de balanceo y coordenadas espaciales.

En algunas realizaciones, la mezcla ascendente se realiza mediante la unidad 404 de decodificación. En estas realizaciones, la decodificación de los datos 328 de audio digital recibidos en audio direccional 420 mediante la unidad 404 de decodificación comprende: decodificar los datos 328 de audio digital recibidos en audio mezclado de manera descendente, y mezclar de manera ascendente, mediante la unidad 404 de decodificación, el audio mezclado de manera descendente en el audio direccional 420 usando una matriz de mezcla descendente incluida en los datos 328 de audio digital recibidos.

El dispositivo comprende además una unidad 406 de renderizado configurada para modificar S23 una propiedad direccional del audio direccional usando los datos espaciales; y renderizar S24 el audio direccional modificado 424 usando altavoces o auriculares.

El dispositivo 400 (la unidad 406 de renderizado del mismo) está así configurado para aplicar rotación/traslación de escenas acústicas basándose en datos espaciales recibidos.

En algunas realizaciones, los datos espaciales indican la orientación espacial y/o la posición espacial de un sistema de micrófono que comprende uno o más micrófonos que capturan el audio direccional, en donde la unidad de renderizado modifica S23 la propiedad direccional del audio direccional para reproducir al menos parcialmente un entorno de audio del sistema de micrófono. En esta realización, el dispositivo 400 vuelve a aplicar al menos partes de la rotación de la escena acústica que fue compensada en el extremo de captura por el dispositivo 300 de la figura 3.

Los datos espaciales pueden comprender datos espaciales que comprenden datos rotacionales que representan movimiento en tres grados de libertad, DoF. De forma alternativa, o adicional, los datos espaciales pueden incluir coordenadas espaciales.

El audio direccional decodificado puede, en algunas realizaciones, comprender objetos de audio o, más generalmente, audio asociado con metadatos espaciales como se describió anteriormente.

La decodificación S22 de los datos de audio digitales recibidos en audio direccional mediante la unidad 404 de decodificación puede comprender en algunas realizaciones la decodificación de los datos de audio digitales recibidos en audio mezclado de manera descendente, y mezclar de manera ascendente, mediante la unidad 404 de decodificación, el audio mezclado de manera descendente en audio direccional usando una matriz de mezcla descendente incluida en los datos 328 de audio digital recibidos.

Para proporcionar una mayor flexibilidad y/o cumplir con los requisitos de ancho de banda, el dispositivo 400 puede comprender una unidad 306 de transmisión configurada para transmitir S20 instrucciones a un dispositivo adicional desde el cual se reciben los datos 328 de audio digital, indicando las instrucciones al dispositivo adicional qué parámetro o parámetros (si los hay) deben comprender los datos de rotación o traslación. De este modo, esta característica puede facilitar el cumplimiento de las preferencias de los posibles usuarios o de las preferencias relacionadas con el renderizado y/o el tipo de servicio usado.

En algunas realizaciones, el dispositivo 400 también puede configurarse para transmitir instrucciones que indiquen al dispositivo adicional si se deben incluir los metadatos que comprenden datos espaciales en los datos 328 de audio digital o no. En estas realizaciones, si los datos 328 de audio digital recibidos S21 no comprenden dichos metadatos, la unidad de renderizado renderizará audio direccional decodificado tal como se recibió (posiblemente mezclado de manera ascendente como se describió anteriormente), sin ninguna modificación de una propiedad direccional del audio direccional debido a compensaciones realizadas en el dispositivo 300 de captura. Sin embargo, en algunas realizaciones, el audio direccional recibido se modifica en respuesta a la información de seguimiento de la cabeza del renderizador (como se describe con más detalle a continuación).

En algunas realizaciones, el dispositivo 400 puede implementarse en un equipo VR o un equipo AR que comprende un dispositivo de seguimiento de la cabeza configurado para medir la orientación espacial del dispositivo en seis DoF. La unidad 406 de renderizado puede configurarse para renderizado de audio binaural.

En algunas realizaciones, la unidad 406 de renderizado está configurada para ajustar S25 un volumen del audio renderizado basándose en las coordenadas espaciales recibidas en los metadatos. Esta característica se describirá con más detalle a continuación junto con las figuras 6-7.

La figura 5 muestra un sistema que comprende un dispositivo 300 de captura (como se describe junto con la figura 3) y un dispositivo 400 de renderizado (como se describe junto con la figura 4). El dispositivo 300 de captura puede en algunas realizaciones recibir S10 instrucciones 334 transmitidas S20 desde el dispositivo 400 de renderizado que indican si y en qué medida el dispositivo 300 de captura debe incluir datos espaciales del sistema de micrófono del dispositivo de captura en los datos 328 de audio digital.

En algunas realizaciones, el dispositivo 300 de captura comprende además una unidad de grabación de video y está configurado para codificar video grabado en datos 502 de video digital y transmitir los datos de video digital al dispositivo 400 de renderizado, en donde el dispositivo 400 de renderizado comprende además un visualizador para visualizar datos de vídeo digital decodificados.

Como se describió anteriormente, un cambio de posición y/o la orientación espacial del sistema de micrófono del dispositivo 300 de captura durante la captura de audio puede causar rotación/traslación espacial de la escena renderizada en el dispositivo 400 de renderizado. Dependiendo del tipo de experiencia proporcionada, por ejemplo, inmersiva, VR, AR o XR y, dependiendo del caso de uso específico, este comportamiento puede ser deseado o no. Un ejemplo en el que esto puede desearse es cuando el servicio proporciona adicionalmente un componente visual 502 y donde la cámara de captura y uno o más micrófonos 302 están integrados en el mismo dispositivo. En ese caso, se debería esperar que una rotación del dispositivo 300 de captura dé como resultado una rotación correspondiente de la escena audiovisual renderizada en el dispositivo 400 de renderizado.

Por otro lado, si la captura audiovisual no se realiza mediante el mismo dispositivo físico o en caso de que no haya un componente de vídeo, puede resultar molesto para el oyente si la escena renderizada gira cada vez que gira el dispositivo 300 de captura. En el peor de los casos, puede provocar mareo.

Por esta razón, de acuerdo con algunas realizaciones, la unidad de renderizado del dispositivo 400 de renderizado puede configurarse para, cuando el dispositivo 400 de renderizado recibe además datos 502 de vídeo codificados desde el dispositivo 300 de captura, modificar una propiedad direccional del audio direccional (recibido en los datos 328 de audio digital) usando los datos espaciales y renderizar el audio direccional modificado.

Sin embargo, cuando el dispositivo 400 de renderizado no recibe datos de vídeo codificados desde el dispositivo 300 de captura, la unidad de renderizado del dispositivo 400 de renderizado puede configurarse para renderizar el audio direccional sin ninguna modificación direccional.

En otras realizaciones, se informa al dispositivo 400 de renderizado antes de la conferencia que no se incluirá ningún componente de vídeo en los datos recibidos desde el dispositivo 300 de captura. En este caso, el dispositivo 400 de renderizado puede indicar en las instrucciones 334 que no es necesario incluir ningún dato espacial del sistema de micrófono del dispositivo 300 de captura en los datos 328 de audio digital, por lo que la unidad de renderizado del dispositivo 400 de renderizado está configurada para renderizar el audio direccional recibido en los datos 328 de audio digital sin ninguna modificación direccional.

En lo anterior, se ha descrito brevemente la mezcla descendente y/o codificación del audio direccional en el dispositivo de captura. Esto se desarrollará ahora más detalladamente.

En muchos casos, el dispositivo 300 de captura no tiene información sobre si la presentación decodificada (en el dispositivo de renderizado) será para un único altavoz mono, altavoces estéreo o auriculares. El escenario de renderizado real también puede variar durante una sesión de servicio, por ejemplo, con equipos de reproducción conectados que pueden cambiar, como la conexión o desconexión de auriculares a un teléfono móvil. Otro escenario más en el que se desconocen las capacidades del dispositivo de renderizado es cuando un único dispositivo 300 de captura necesita soportar múltiples puntos finales (dispositivos 400 de renderizado). Por ejemplo, en un caso de uso de conferencia IVAS o distribución de contenido de realidad virtual, un punto final podría estar usando unos auriculares y otro podría renderizar en altavoces estéreo, pero sería ventajoso poder suministrar una única codificación a ambos puntos finales, ya que reduciría complejidad en el lado de la codificación y también puede reducir el ancho de banda de red agregado requerido.

Una forma sencilla, aunque menos deseable, de soportar estos casos sería asumir siempre la capacidad más baja del dispositivo de recepción, es decir, mono, y seleccionar el modo de operación de audio correspondiente. Sin embargo, es más sensato exigir que el códec usado (por ejemplo, el códec IVAS), incluso si se opera en un modo de presentación que soporta audio espacial, binaural o estéreo, siempre puede producir una señal de audio decodificada que se puede presentar en dispositivos 400 con capacidad de audio respectivamente inferior. En algunas realizaciones, una señal codificada como una señal de audio espacial también puede ser decodificable para renderización binaural, estéreo y/o mono. Asimismo, una señal codificada como binaural puede ser decodificable como estéreo o mono, y una señal codificada como estéreo puede ser decodificable para presentación mono. A modo de ilustración, un dispositivo 300 de captura solo debería necesitar implementar una única codificación (datos 328 de audio digital) y enviar la misma codificación a múltiples puntos finales 400, algunos de los cuales pueden soportar presentación binaural y otros pueden ser solo estéreo.

Cabe señalar que el códec discutido anteriormente puede implementarse en el dispositivo de captura o en el servidor de llamadas. En el caso del servidor de llamadas, el servidor de llamadas recibirá los datos 328 de audio digital desde el dispositivo de captura y realizará una transcodificación de los datos de audio digitales para cumplir con los requisitos anteriores, antes de enviar los datos de audio digitales transcodificados a uno o más dispositivos 400 de renderizado. Tal escenario se ejemplificará ahora junto con la figura 6.

El escenario físico 600 de conferencia VR se ilustra en la figura 6. Cinco usuarios 602a-e de conferencia VR/AR de diferentes sitios se están reuniendo virtualmente. Los usuarios 602a-e de conferencia VR/AR pueden estar habilitados para IVAS. Cada uno de ellos usa equipos VR/AR, incluida, por ejemplo, reproducción binaural y reproducción de vídeo usando un HMD. El equipo de todos los usuarios soporta movimientos en 6 DoF con el seguimiento de la cabeza correspondiente head-tracking. El equipo 602 de usuario, UE, de los usuarios intercambia audio codificado en sentido ascendente y descendente con un servidor 604 de llamada de conferencia. Visualmente, los usuarios pueden representarse a través de respectivos avatares que pueden renderizarse basándose en información relacionada con los parámetros de posición relativa y su orientación rotacional.

Para mejorar aún más la experiencia inmersiva del usuario, también se considera el movimiento de rotación y/o el movimiento de traslación de la cabeza de un oyente al renderizar el audio recibido de otros participantes en el escenario de la conferencia. En consecuencia, el seguimiento de la cabeza informa a la unidad de renderizado del dispositivo de renderizado de un usuario (referencia 400 en las figuras 4-5) sobre los datos espaciales actuales (6 DoF) del equipo VR/AR del usuario. Estos datos espaciales se combinan (por ejemplo, mediante multiplicación de matrices o modificación de metadatos asociados con el audio direccional) con datos espaciales recibidos en los datos de audio digitales recibidos de otro usuario 602, mediante lo cual la unidad de renderizado modifica una propiedad direccional del audio direccional recibido de dicho otro usuario 602 basándose en la combinación de datos espaciales. Luego, el audio direccional modificado es renderizado al usuario.

Además, el volumen del audio renderizado recibido de un usuario específico se puede ajustar basándose en las coordenadas espaciales recibidas en los datos de audio digitales. Basándose en una distancia virtual (o real) entre los dos usuarios (calculada por el dispositivo de renderizado o por el servidor 604 de llamadas), el volumen se puede aumentar o disminuir para mejorar aún más la experiencia inmersiva del usuario.

La figura 7 ilustra a modo de ejemplo un espacio 700 de conferencia virtual generado por el servidor de llamadas de conferencia. Inicialmente, el servidor coloca a los usuarios de conferencia Ui, i=1...5 (también denominados 702a-e), en las coordenadas de posición virtual $K_i=(x_i, y_i, z_i)$ El espacio de conferencia virtual se comparte entre los usuarios. En consecuencia, en ese espacio se realiza la renderización audiovisual de cada usuario. Por ejemplo, desde la perspectiva del usuario U5 (correspondiente al usuario 602d en la figura 6), el renderizado colocará virtualmente a los otros participantes de la conferencia en las posiciones relativas $K_i - K_5$, $i \neq 5$. Por ejemplo, el usuario U5 percibirá al usuario U2 a distancia $|K_i - K_5|$ y bajo la dirección del vector $(K_i - K_5)/|K_i - K_5|$, mediante el cual el renderizado direccional se realiza en relación con la posición de rotación de U5. También se ilustra en la figura 2 el movimiento de U5 hacia U4. Este movimiento afectará la posición de U5 en relación con los demás usuarios, lo que se tendrá en cuenta durante el renderizado. Al mismo tiempo, el UE de U5 envía su posición cambiante al servidor 604 de conferencia, que actualiza el espacio de conferencia virtual con las nuevas coordenadas de U5. A medida que se comparte el espacio de conferencia virtual, los usuarios U1-U4 se dan cuenta del movimiento del usuario U5 y pueden adaptar en consecuencia sus respectivas renderizaciones. El movimiento simultáneo del usuario U2 funciona de acuerdo con los principios correspondientes. El servidor 604 de llamadas está configurado para mantener los datos de posición de los participantes 702a-e en el espacio de reunión compartido.

En el escenario de la figura 6-7, uno o más de los siguientes requisitos 6 DoF pueden aplicarse al marco de codificación cuando se trata de audio:

- Ofrecer un marco de metadatos para la representación y transmisión ascendente de información posicional de un punto final de recepción, incluidas coordenadas espaciales y/o coordenadas rotacionales (como se describe anteriormente junto con las figuras 1-4).

- La capacidad de asociar elementos de audio de entrada (por ejemplo, objetos) con atributos 6 DoF, incluidas coordenadas espaciales, coordenadas de rotación y directividad.

- La capacidad de renderización espacial simultánea de múltiples elementos de audio recibidos respectivos de sus atributos 6 DoF asociados.

- Ajustes adecuados de la escena renderizada ante movimientos de rotación y traslación de la cabeza del oyente.

Cabe señalar que lo anterior también aplica para las reuniones XR, siendo una mezcla de reunión física y virtual. Los participantes físicos ven y escuchan avatares que representan a los participantes remotos a través de sus

gafas AR y auriculares. Interactúan con los avatares en las discusiones como si fueran participantes físicamente presentes. Para ellos, las interacciones con otros participantes físicos y virtuales ocurren en una realidad mixta. Las posiciones de los participantes reales y virtuales se fusionan en un espacio de reunión virtual compartido combinado (por ejemplo, mediante un servidor 604 de llamadas) que es consistente con las posiciones de las posiciones reales de los participantes en el espacio de reunión físico y se asignan al espacio de reunión virtual usando los datos de posición física/real relativa y absoluta.

En un escenario VR/AR/XR, se podrán formar subgrupos de la conferencia virtual. Estos subgrupos pueden usarse para informar al servidor 604 de llamadas entre qué usuarios, por ejemplo, la calidad del servicio, QoS, debe ser alta, y entre qué usuarios la QoS puede ser menor. En algunas realizaciones, solo los participantes de un mismo subgrupo se incluyen en un entorno virtual proporcionado a estos subgrupos a través del equipo VR/AR/XR. Por ejemplo, un escenario donde se pueden formar subgrupos es una sesión de pósteres que ofrece participación virtual desde una ubicación remota. Los participantes remotos están equipados con HMD y auriculares. Están prácticamente presentes y pueden caminar de un póster a otro. Pueden escuchar presentaciones de pósteres en curso y acercarse a una presentación si creen que el tema o la discusión en curso es interesante. Para mejorar la posibilidad de interacciones inmersivas entre los participantes virtuales y físicos, se pueden formar subgrupos, por ejemplo, basándose en qué póster de la pluralidad de pósteres en el que están interesados actualmente los participantes.

Las realizaciones de este escenario comprenden:

- recibir, mediante un sistema de teleconferencia, temas de los participantes de una conferencia virtual;

- agrupar, mediante el sistema de teleconferencia basándose en los temas, a los participantes en subgrupos de la conferencia virtual;

- recibir, por el sistema de teleconferencia, una petición desde un dispositivo de un nuevo participante para unirse a la conferencia virtual, estando asociada la petición con un indicador que indica un tema preferido;

- seleccionar, mediante el sistema de teleconferencia, un subgrupo de los subgrupos basándose en el tema preferido y los temas de los subgrupos;

- proporcionar, mediante el sistema de teleconferencia al dispositivo del nuevo participante, un entorno virtual de la conferencia virtual, indicando el entorno virtual al menos uno de una proximidad virtual visual o una proximidad virtual de audio entre el nuevo participante y uno o más participantes del subgrupo seleccionado.

En algunas realizaciones, el entorno virtual indica la proximidad virtual visual o la proximidad virtual de audio al menos proporcionando un visualizador de realidad virtual o un campo de sonido de realidad virtual donde un avatar del nuevo participante y uno o más avatares de los participantes del subgrupo seleccionado están cerca unos de otros.

En algunas realizaciones, cada participante está conectado mediante unos auriculares abiertos y gafas AR.

VI - Equivalentes, extensiones, alternativas y varios

Otras realizaciones de la presente divulgación resultarán evidentes para un experto en la técnica después de estudiar la descripción anterior. Aunque la presente descripción y los dibujos divulgan realizaciones y ejemplos, la divulgación no se limita a estos ejemplos específicos. Se pueden realizar numerosas modificaciones y variaciones sin apartarse del alcance de la presente divulgación, que está definido por las reivindicaciones adjuntas. Los signos de referencia que aparecen en las reivindicaciones no deben entenderse como limitativos de su alcance.

Además, el experto en la práctica de la divulgación puede comprender y efectuar variaciones de las realizaciones divulgadas, a partir de un estudio de los dibujos, la divulgación y las reivindicaciones adjuntas. En las reivindicaciones, la palabra "que comprende" no excluye otros elementos o pasos, y el artículo indefinido "un" o "una" no excluye una pluralidad. El mero hecho de que determinadas medidas se mencionen en reivindicaciones dependientes mutuamente diferentes no indica que una combinación de estas medidas no pueda usarse con beneficio.

Los sistemas y métodos divulgados anteriormente pueden implementarse como software, firmware, hardware o una combinación de los mismos. En una implementación de hardware, la división de tareas entre unidades funcionales a la que se hace referencia en la descripción anterior no corresponde necesariamente a la división en unidades físicas; por el contrario, un componente físico puede tener múltiples funcionalidades y una tarea puede ser realizada por varios componentes físicos en cooperación. Ciertos componentes o todos los componentes pueden implementarse como software ejecutado por un procesador o microprocesador de señales digitales, o implementarse como hardware o como un circuito integrado de aplicación específica. Tal software puede

distribuirse en medios legibles por ordenador, que pueden comprender medios de almacenamiento informático (o medios no transitorios) y medios de comunicación (o medios transitorios). Como es bien conocido por un experto en la técnica, el término medios de almacenamiento informático incluye medios tanto volátiles como no volátiles, extraíbles y no extraíbles implementados en cualquier método o tecnología para el almacenamiento de información, tales como instrucciones legibles por ordenador, estructuras de datos, módulos de programa u otros datos. Los medios de almacenamiento informático incluyen, entre otros, RAM, ROM, EEPROM, memoria flash u otra tecnología de memoria, CD-ROM, discos versátiles digitales (DVD) u otro almacenamiento en disco óptico, casetes magnéticos, cintas magnéticas, almacenamiento en disco magnético u otros dispositivos de almacenamiento magnético, o cualquier otro medio que pueda usarse para almacenar la información deseada y al que se pueda acceder mediante un ordenador. Además, el experto en la técnica sabe bien que los medios de comunicación típicamente incorporan instrucciones legibles por ordenador, estructuras de datos, módulos de programa u otros datos en una señal de datos modulada tal como una onda portadora u otro mecanismo de transporte e incluye cualquier medio de entrega de información.

Todas las figuras son esquemáticas y generalmente solo muestran partes que son necesarias para aclarar la divulgación, mientras que otras partes pueden omitirse o simplemente sugerirse. A menos que se indique lo contrario, números de referencia similares se refieren a partes similares en figuras diferentes.

REIVINDICACIONES

1.- Un dispositivo que comprende o se puede conectar a un sistema (302) de micrófono que comprende uno o más micrófonos para capturar audio, comprendiendo el dispositivo (300):

una unidad (304) de recepción configurada para:

recibir (S13) audio direccional (320) capturado por el sistema de micrófono;

recibir (S14) metadatos (322) asociados con el sistema de micrófono, comprendiendo los metadatos datos espaciales del sistema de micrófono, siendo los datos espaciales indicativos de un cambio en la orientación espacial y/o la posición espacial del sistema de micrófono en comparación con una orientación/posición anterior del sistema de micrófono y que comprende al menos uno de la lista de: acimut, cabeceo, ángulo o ángulos de balanceo y coordenadas espaciales del sistema de micrófono;

una unidad informática (306) configurada para:

modificar al menos parte del audio direccional para producir audio direccional modificado, mediante el cual se modifica una propiedad direccional del audio en respuesta a la orientación espacial y/o posición espacial del sistema de micrófono;

codificar el audio direccional modificado en datos (328) de audio digital;

una unidad (308) de transmisión configurada para transmitir los datos de audio digitales.

2.- Un dispositivo de acuerdo con la reivindicación 1, en el que la unidad informática está configurada además para codificar al menos partes de los metadatos que comprenden datos espaciales del sistema de micrófono en dichos datos de audio digitales.

3.- Un dispositivo de acuerdo con la reivindicación 2, en el que la unidad de recepción está configurada además para recibir (S11) primeras instrucciones (334) que indican a la unidad informática si incluir dichas al menos partes de los metadatos que comprenden datos espaciales del sistema de micrófono en dicho datos de audio digitales, por lo que la unidad informática actúa en consecuencia.

4.- Un dispositivo de acuerdo con cualquiera de las reivindicaciones 2 a 3, en el que la unidad de recepción está configurada además para recibir (S12) segundas instrucciones (334) que indican a la unidad informática qué parámetro o parámetros de los datos espaciales del sistema de micrófono incluir en los datos de audio digitales, actuando en consecuencia la unidad informática.

5.- Un dispositivo de acuerdo con cualquiera de las reivindicaciones 1 a 4, en el que la unidad de recepción está configurada además para recibir metadatos que comprenden una marca de tiempo que indica un tiempo de captura del audio direccional, en el que la unidad informática está configurada para codificar dicha marca de tiempo en dicha datos de audio digitales.

6.- Un dispositivo de acuerdo con cualquiera de las reivindicaciones 1 a 5, en el que la codificación del audio direccional modificado comprende mezclar de manera descendente el audio direccional modificado, en el que la mezcla descendente se realiza teniendo en cuenta la orientación espacial del sistema de micrófono, y codificar la mezcla descendente y una matriz de mezcla descendente usada en la mezcla descendente de dichos datos de audio digitales.

7.- Un dispositivo de acuerdo con cualquiera de las reivindicaciones 1 a 6, que se implementa en un equipo de realidad virtual, VR (602a-e) o un equipo de realidad aumentada, AR, (602 a-e) que comprende el sistema de micrófono y un dispositivo de seguimiento de la cabeza configurado para determinar datos espaciales del dispositivo en 3-6 DoF.

8.- Un dispositivo (400) para renderizar señales de audio, comprendiendo el dispositivo:

una unidad (402) de recepción configurada para recibir (S21) datos de audio digitales (328),

una unidad (404) de decodificación configurada para:

decodificar (S22) los datos de audio digitales recibidos en audio direccional (420) y en metadatos (422), comprendiendo los metadatos datos espaciales al menos uno de la lista de: acimut, cabeceo, ángulo o ángulos de balanceo y coordenadas espaciales;

una unidad (406) de renderizado configurada para:

modificar (S23) una propiedad direccional del audio direccional usando los datos espaciales, en donde los datos espaciales indican un cambio en la orientación rotacional y/o la posición espacial de un sistema (302) de micrófono que comprende uno o más micrófonos que han capturado el audio direccional, en comparación con una orientación/posición anterior del sistema de micrófono, en donde la unidad de renderizado modifica la propiedad direccional del audio direccional para reproducir al menos parcialmente un entorno de audio del sistema de micrófono; y

renderizar (S24) el audio direccional modificado (424).

9.- Un dispositivo de acuerdo con la reivindicación 8, que comprende además una unidad (306) de transmisión configurada para transmitir instrucciones (334) a un dispositivo adicional (300) desde el cual se recibe el audio digital, indicando las instrucciones al dispositivo adicional qué parámetro o parámetros deben comprender los datos de rotación.

10.- Un dispositivo de acuerdo con cualquiera de las reivindicaciones 8 a 9, en el que la unidad de decodificación está configurada además para extraer una marca de tiempo que indica un tiempo de captura del audio direccional a partir de los datos de audio digitales.

11.- Un dispositivo de acuerdo con cualquiera de las reivindicaciones 8 a 10, en el que la decodificación de los datos de audio digitales recibidos en audio direccional mediante la unidad de decodificación comprende:

decodificación de los datos de audio digitales recibidos en audio mezclado de manera descendente,

mezclar de manera ascendente, mediante la unidad de decodificación, el audio mezclado de manera descendente en el audio direccional usando una matriz de mezcla descendente incluida en los datos de audio digitales recibidos.

12.- Un dispositivo de acuerdo con cualquiera de las reivindicaciones 8 a 11, que se implementa en un equipo (602a-e) de realidad virtual, VR, o equipo (602a-e) de realidad aumentada, AR, que comprende un dispositivo de seguimiento de la cabeza configurado para medir la orientación espacial y la posición espacial del dispositivo en seis DoF.

13.- Un dispositivo de acuerdo con cualquiera de las reivindicaciones 8 a 12, en el que la unidad de renderizado está configurada para renderizado de audio binaural.

14.- Un sistema que comprende:

un primer dispositivo (300) de acuerdo con cualquiera de las reivindicaciones 1 a 7, configurado para transmitir datos de audio digitales a un segundo dispositivo (400) de acuerdo con cualquiera de las reivindicaciones 8 a 13, en el que el sistema está configurado para audio y/o videoconferencia.

15.- Un medio no transitorio legible por ordenador que almacena instrucciones que, cuando son ejecutadas por uno o más procesadores, hacen que uno o más procesadores realicen operaciones de:

recibir (S13) audio direccional (320) capturado por un sistema de micrófono;

recibir (S14) metadatos (322) asociados con el sistema de micrófono, comprendiendo los metadatos datos espaciales del sistema de micrófono, siendo los datos espaciales indicativos de un cambio en la orientación espacial y/o la posición espacial del sistema de micrófono en comparación con una orientación anterior/posición del sistema de micrófono y que comprende al menos uno de la lista de: acimut, cabeceo, ángulo o ángulos de balanceo y coordenadas espaciales del sistema de micrófono;

modificar al menos parte del audio direccional para producir audio direccional modificado, mediante el cual se modifica una propiedad direccional del audio en respuesta a la orientación espacial y/o la posición espacial del sistema de micrófono; y

codificar el audio direccional modificado en datos de audio digitales (328).

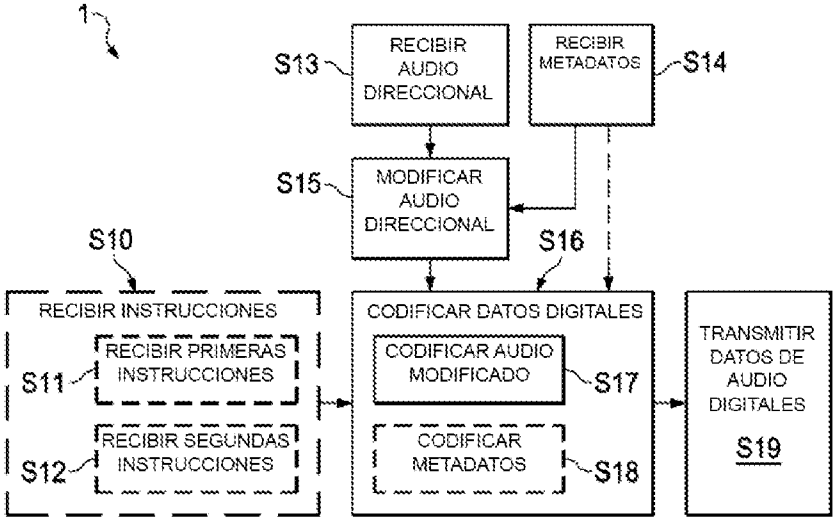
1
S13
RECIBIR AUDIO DIRECCIONAL
RECIBIR METADATOS
S14
S15
MODIFICAR AUDIO DIRECCIONAL
S10
S16
RECIBIR INSTRUCCIONES
RECIBIR PRIMERAS INSTRUCCIONES
S11
RECIBIR SEGUNDAS INSTRUCCIONES
S12
CODIFICAR DATOS DIGITALES
CODIFICAR AUDIO MODIFICADO
S17
CODIFICAR METADATOS
S18
TRANSMITIR DATOS DE AUDIO DIGITALES
S19

FIG. 1

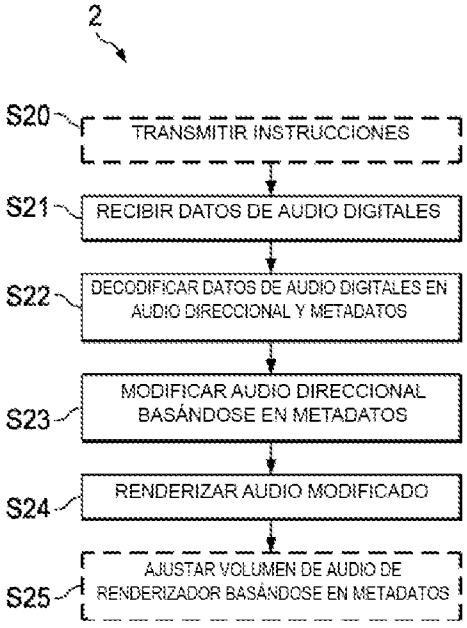
2
S20
TRANSMITIR INSTRUCCIONES
S21
RECIBIR DATOS DE AUDIO DIGITALES
S22
DECODIFICAR DATOS DE AUDIO DIGITALES EN AUDIO DIRECCIONAL Y METADATOS
S23
MODIFICAR AUDIO DIRECCIONAL BASÁNDOSE EN METADATOS
S24
RENDERIZAR AUDIO MODIFICADO
S25
AJUSTAR VOLUMEN DE AUDIO DE RENDERIZADOR BASÁNDOSE EN METADATOS

FIG. 2

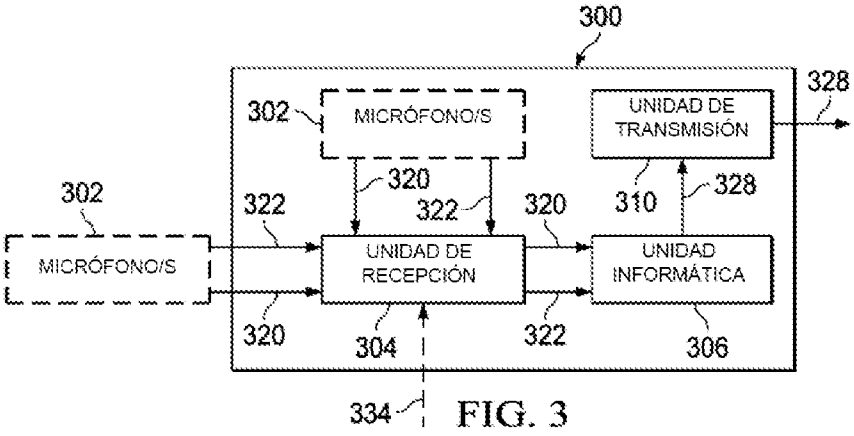

FIG. 3

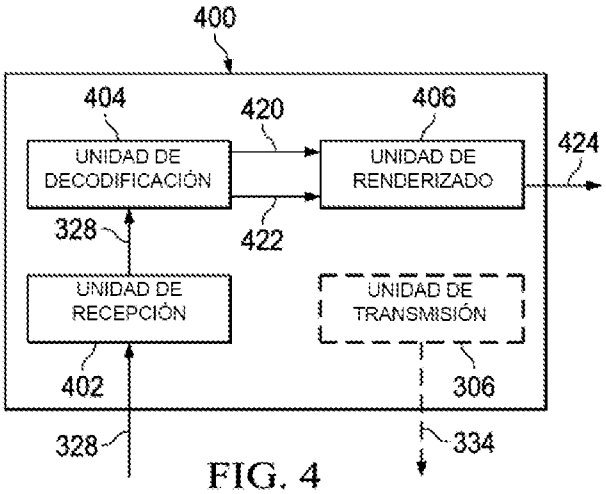

FIG. 4

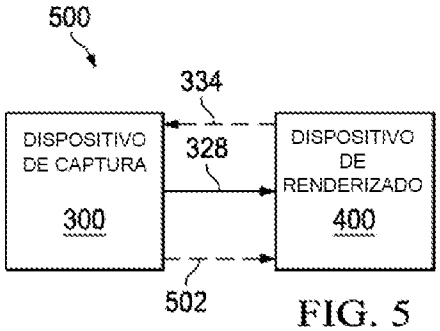

FIG. 5

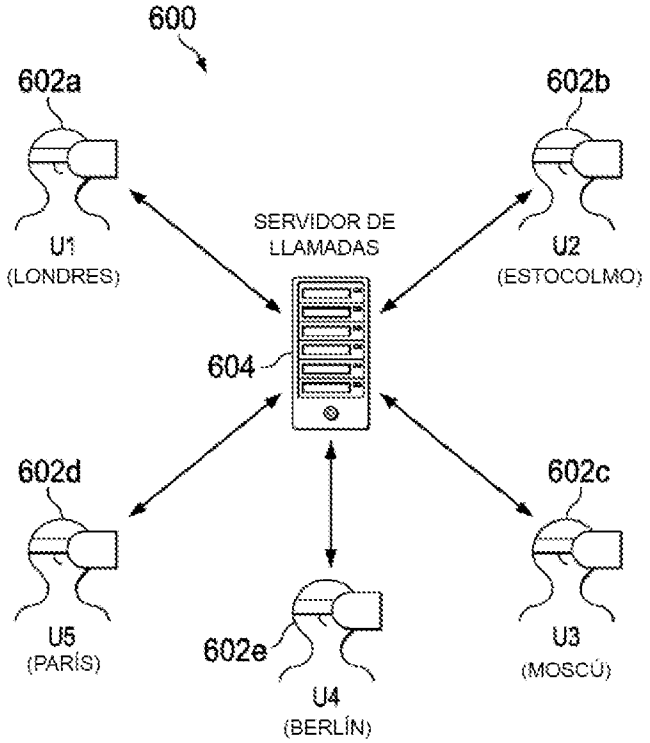
600
602a
602b
SERVIDOR DE LLAMADAS
U1
(LONDRES)
U2
(ESTOCOLMO)
604
602d
602c
U5
(PARÍS)
602e
U4
(BERLÍN)
U3
(MOSCÚ)

FIG. 6

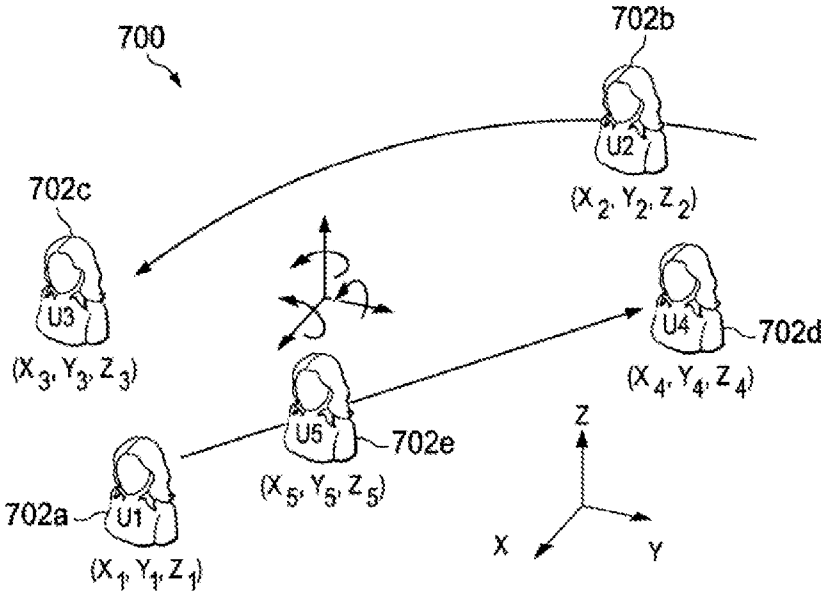
700
702b
U2
(X2, Y2, Z2)
702c
U3
(X3, Y3, Z3)
U4
702d
(X4, Y4, Z4)
U5
702e
(X5, Y5, Z5)
702a
U1
(X1, Y1, Z1)
Z
X
Y

FIG. 7