(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 970 490**

(51) Int. CI.:

H04N 21/81 (2011.01)
H04N 21/218 (2011.01)
H04N 21/6587 (2011.01)
H04N 21/2343 (2011.01)
H04N 21/439 (2011.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **11.10.2018 PCT/EP2018/077770**

(87) Fecha y número de publicación internacional: **18.04.2019 WO19072984**

(96) Fecha de presentación y número de la solicitud europea: **11.10.2018 E 18783491 (6)**

(97) Fecha y número de publicación de la concesión europea: **03.01.2024 EP 3695613**

(54) Título: **Optimización de la transmisión de audio para aplicaciones de realidad virtual**

(30) Prioridad:

12.10.2017 EP 17196259

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
29.05.2024

(73) Titular/es:

FRAUNHOFER-GESELLSCHAFT ZUR FÖRDERUNG DER ANGEWANDTEN FORSCHUNG E.V. (100.0%)
Hansastr. 27c
80686 München, DE

(72) Inventor/es:

MURTAZA, ADRIAN;
FUCHS, HARALD;
CZELHAN, BERND;
PLOGSTIES, JAN;
AGNELLI, MATTEO y
HOFMANN, INGO

(74) Agente/Representante:

ARIZTI ACHA, Monica

DESCRIPCIÓN

Optimización de la transmisión de audio para aplicaciones de realidad virtual

Introducción

En un entorno de realidad virtual (VR) o de manera similar en un entorno de realidad aumentada (AR) o realidad mixta (MR) o de vídeo de 360 grados, el usuario generalmente puede visualizar un contenido completo de 360 grados utilizando, por ejemplo, una pantalla montada en la cabeza (HMD) y escuchándolo a través de los auriculares (o de manera similar a través de los altavoces, incluida la reproducción correcta en función de su posición).

En un caso de uso simple, el contenido se crea de tal manera que sólo una escena de audio/ vídeo (es decir, un vídeo de 360 grados, por ejemplo) se reproduce en un momento determinado. La escena de audio/ vídeo tiene una ubicación fija (por ejemplo, una esfera con el usuario posicionado en el centro), y el usuario no puede moverse en la escena, pero sólo puede girar su cabeza en varias direcciones (desvío, inclinación, balanceo). En este caso, se reproducen diferentes vídeos y audios (se visualizan diferentes ventanas gráficas) para el usuario según la orientación de su cabeza.

Mientras que para el vídeo, el contenido de vídeo se entrega para toda la escena de 360 grados, junto con los metadatos para describir el proceso de renderizado (por ejemplo, información de costura, mapeo de proyección, etc.) y se selecciona en función de la ventana gráfica del usuario actual, para el audio, el contenido es el mismo para toda la escena. En función de los metadatos, el contenido de audio se adapta a la ventana gráfica del usuario actual (por ejemplo, un objeto de audio se procesa de forma diferente según la información de la ventana gráfica/orientación del usuario). Debe tenerse en cuenta que el contenido de 360 grados se refiere a cualquier tipo de contenido que comprenda en más de un ángulo de visión en el mismo momento en el que el usuario puede elegir (por ejemplo, por su orientación de la cabeza o mediante un dispositivo de control remoto).

En un escenario más complejo, cuando el usuario puede moverse en la escena VR, o "saltar" de una escena a la siguiente, el contenido de audio también puede cambiar (por ejemplo, las fuentes de audio que no son audibles en una escena pueden volverse audibles en la siguiente escena – "se abre una puerta"). Con los sistemas existentes, las escenas de audio completas pueden codificarse en una secuencia y, si es necesario, en secuencias adicionales (dependiendo de la secuencia principal). Dichos sistemas se conocen como sistemas de audio de próxima generación (por ejemplo, MPEG–H 3D Audio). Los ejemplos de tales casos de uso pueden contener:

- Ejemplo 1: • El usuario selecciona ingresar a una nueva sala, y la escena de audio/ vídeo cambia

-Ejemplo 2: • El usuario se mueve en la escena VR, abre la puerta y camina, lo que implica una transición de audio de una escena a la siguiente escena requerida

Con el fin de describir este escenario, se introduce la noción de puntos de vista discretos en el espacio, como ubicación discreta en el espacio (o en el entorno de realidad virtual), para la cual se dispone de diferentes contenidos de audio/ vídeo.

La solución "directa" es tener un codificador en tiempo real que cambie la codificación (número de elementos de audio, información espacial, etc.) en función de los comentarios del dispositivo de reproducción sobre la posición/orientación del usuario. Esta solución implicaría, por ejemplo, en un entorno de transmisión, una comunicación muy compleja entre un cliente y un servidor:

• El cliente (que normalmente se supone que usa sólo lógica simple) requeriría mecanismos avanzados para transmitir no sólo las solicitudes de diferentes a la ventana gráfica actual del usuario y/o la orientación s, sino también información compleja sobre los detalles de codificación que permitirían procesar el contenido correcto en función de la posición del usuario.

• El servidor de medios generalmente se rellena previamente con diferentes transmisiones (formateadas de una manera específica que permite la entrega “segmentada”) y la función principal del servidor es proporcionar información sobre las transmisiones disponibles y provocar su entrega cuando se solicita. Para habilitar escenarios que permitan la codificación basada en la retroalimentación del dispositivo de reproducción, el servidor de medios requeriría enlaces de comunicación avanzados con múltiples codificadores de medios en vivo, y la capacidad de crear toda la información de señalización sobre la marcha (por ejemplo, descripción de presentación de medios) que podría cambiar en tiempo real.

Aunque tal sistema podría imaginarse, su complejidad y requisitos computacionales están más allá de la funcionalidad y las características de los equipos y sistemas disponibles en la actualidad o incluso en los que se desarrollarán en las próximas décadas.

Alternativamente, el contenido que representa el entorno completo de VR ("el mundo completo") se puede entregar todo el tiempo. Esto resolvería el problema, pero requeriría una tasa de bits enorme que está más allá de la capacidad de los enlaces de comunicaciones disponibles.

Esto es complejo para un entorno en tiempo real, y para habilitar tales casos de uso utilizando los sistemas disponibles, se proponen soluciones alternativas que permiten esta funcionalidad con una baja complejidad.

2. Terminología y definiciones

Se usa la siguiente terminología en el campo técnico:

• **Elementos de audio**: señales de audio que se pueden representar, por ejemplo, como objetos de audio, canales de audio, audio basado en escenas (Higher Order Ambisonics – HOA), o cualquier combinación de todos.

• **Región de interés** (ROI): Una región del contenido de vídeo (o del entorno visualizado o simulado) que interesa al usuario en un momento dado. Esto puede ser comúnmente una región en una esfera, por ejemplo, o una selección poligonal de un mapa 2D. La ROI identifica una región específica para un propósito particular, definiendo los bordes de un objeto en consideración.

• **Información de la posición del usuario**: información de la ubicación (por ejemplo, coordenadas x, y, z), información de la orientación (desvío, inclinación, balanceo), dirección y velocidad de movimiento, etc.

• **Ventana gráfica**: Parte del vídeo esférico que actualmente muestra y ve el usuario.

• **Punto de vista**: el punto central de la ventana gráfica.

• **vídeo de 360 grados** (también conocido como vídeo inmersivo o vídeo esférico): representa en el contexto de este documento un contenido de vídeo que contiene más de una vista (es decir, ventana gráfica) en una dirección en el mismo momento. Dicho contenido puede crearse, por ejemplo, utilizando una cámara omnidireccional o una colección de cámaras. Durante la reproducción, el espectador tiene el control de la dirección de visualización.

• La **descripción de la presentación de medios (MPD)** es una sintaxis, por ejemplo, XML que contiene información sobre segmentos de medios, sus relaciones e información necesaria para elegir entre ellos.

• Los **conjuntos de adaptación** contienen una secuencia de medios o un conjunto de secuencias de medios. En el caso más simple, un conjunto de adaptación que contiene todo el audio y el vídeo para el contenido, pero para reducir el ancho de banda, cada transmisión se puede dividir en un conjunto de adaptación diferente. Un caso común es tener un conjunto de adaptación de vídeo y múltiples conjuntos de adaptación de audio (uno para cada idioma admitido). Los conjuntos de adaptación también pueden contener subtítulos o metadatos arbitrarios.

• Las **Representaciones** permiten que un conjunto de adaptación contenga el mismo contenido codificado de diferentes maneras. En la mayoría de los casos, las representaciones se proporcionarán en múltiples tasas de bits. Esto permite a los clientes solicitar el contenido de mayor calidad que pueden reproducir sin esperar a un búfer. Las representaciones también pueden codificarse con diferentes códecs, lo que permite el soporte para clientes con diferentes códecs compatibles.

En el contexto de esta aplicación, las nociones de los conjuntos de adaptación se usan de manera más genérica, a veces refiriéndose realmente a las representaciones. Además, las secuencias de medios (transmisiones de audio/vídeo) generalmente se encapsulan primero en segmentos de medios que son los archivos de medios reales reproducidos por el cliente (por ejemplo, el cliente DASH). Se pueden usar varios formatos para los segmentos de medios, como el formato de archivo de medios de base ISO (ISOBMFF), que es similar al formato de contenedor MPEG–4, o ella ventana gráfica actual del usuario y/o la orientación de transporte MPEG–2 (TS). La encapsulación en segmentos de medios y en diferentes representaciones/conjuntos de adaptación es independiente de los métodos descritos aquí, los métodos se aplican a todas las diversas opciones.

Además, la descripción de los métodos en este documento se centra en una comunicación DASH Servidor-Cliente, pero los métodos son lo suficientemente genéricos para trabajar con otros entornos de entrega, como MMT, MPEG-2 TS, DASH-ROUTE, formato de archivo para archivo reproducción etc.

En términos generales, un conjunto de adaptación está en una capa superior con respecto a una transmisión y puede comprender metadatos (por ejemplo, asociados a posiciones). Una transmisión puede comprender una pluralidad de elementos de audio. Una escena de audio puede asociarse a una pluralidad de transmisiones entregados como una parte de una pluralidad de conjuntos de adaptación.

3. Soluciones actuales

Las soluciones actuales son:

[1]. ISO/IEC 23008-3: 2015, Information technology -- High efficiency coding and media delivery in heterogeneous environments -- Parte 3: audio 3D

[2]. N16950, Estudio del formato de medio omnidireccional ISO/IEC DIS 23000–20

Las soluciones actuales se limitan a proporcionar una experiencia VR independiente en una ubicación fija que permite al usuario cambiar su orientación pero no moverse en el entorno VR.

A discussion on the scalable system design for the creation of a realistic voice communication service for crowded virtual spaces es proporcionada por Boustead P. et al., "DICE: Internet delivery of immersive voice communication for crowded virtual spaces", Virtual Reality, 2005, Proceedings. VR 2005. IEEE Bonn, Alemania 12-16 de marzo de 2005, Piscataway, NJ, EE.UU., IEEE Piscataway, NJ, EE.UU. 12 de marzo de 2005, páginas 35-41.

El documento EP 3 065 406 A1 describe un método de transmisión de vídeo en el que la resolución de un vídeo panorámico se ajusta en función del campo de visión actual.

El documento WO 2017/120681 A1 describe un método para determinar automáticamente una salida tridimensional posicional de información de audio basada en la orientación del usuario dentro de un entorno inmersivo de activación.

El documento US 2017/127118 A1 describe un aparato de procesamiento de información capaz de mejorar la eficiencia de adquisición de un tipo predeterminado de datos de audio entre una pluralidad de tipos de datos de audio.

El documento US 2010/040238 A1 describe un aparato para procesar el sonido en un sistema de realidad virtual.

El documento US 2006/0212147 A1 describe un sistema de chat según el cual varios usuarios se encuentran en diferentes salas.

Sumario

La invención se define en las reivindicaciones independientes.

De acuerdo con una forma de realización, un sistema para una realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR o entorno de vídeo de 360 grados puede configurarse para recibir transmisiones de vídeo y audio para ser reproducidas en un dispositivo de consumo de medios, en donde el sistema puede comprender: al menos un decodificador de vídeo multimedia configurado para decodificar señales de vídeo de transmisiones de vídeo para la representación de escenas de entornos de vídeo VR, AR, MR o 360 grados para un usuario, y al menos un decodificador de audio configurado para decodificar señales de audio de al menos una transmisión de audio, en el que el sistema puede configurarse para solicitar al menos una transmisión de audio y/o un elemento de audio de una transmisión de audio y/o un conjunto de adaptación a un servidor sobre la base de al menos la ventana gráfica actual del usuario y/u orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales.

De acuerdo con un aspecto, el sistema puede configurarse para proporcionar al servidor la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o los datos de posición virtuales para obtener al menos una secuencia de audio y/o un elemento de audio de una transmisión de audio y/o un conjunto de adaptación del servidor.

Una forma de realización puede configurarse de modo que al menos una escena esté asociada a al menos un elemento de audio, donde cada elemento de audio está asociado a una posición y/o área en el entorno visual donde el elemento de audio es audible, de modo que se proporcionan diferentes transmisiones de audio para diferentes posiciones del usuario y/o ventanas gráficas y/u orientaciones de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos virtuales de posición en la escena.

De acuerdo con otro aspecto, el sistema puede configurarse para decidir si al menos un elemento de audio de una transmisión de audio y/o un conjunto de adaptación se reproducirán para la ventana del usuario actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o posición virtual en la escena, y en el que el sistema puede configurarse para solicitar y/o recibir el al menos un elemento de audio en la posición virtual del usuario actual.

De acuerdo con un aspecto, el sistema puede configurarse para decidir de manera predecible si al menos un elemento de audio de una transmisión de audio y/o un conjunto de adaptación se volverán relevantes y/o audibles en función de al menos la vista actual y la orientación de la cabeza del usuario y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales, y en el que el sistema puede configurarse para solicitar y/o recibir al menos un elemento de audio y/o transmisión de audio y/o conjunto de adaptación en una determinada posición virtual del usuario antes del movimiento y/o interacción predichos del usuario en la escena, en donde el sistema puede configurarse para reproducir al menos en el elemento de audio y/o transmisión de audio, cuando se recibe, en la posición virtual del usuario particular después del movimiento del usuario y/o la interacción en la escena.

Una forma de realización del sistema puede configurarse para solicitar y/o recibir el al menos un elemento de audio a una tasa de bits y/o nivel de calidad más bajos, en la posición virtual del usuario ante el movimiento y/o interacción de un usuario en la escena, en donde el sistema puede configurarse para solicitar y/o recibir el al menos un elemento de audio a una tasa de bits y/o nivel de calidad más altos, en la posición virtual del usuario después del movimiento y/o interacción del usuario en la escena.

De acuerdo con un aspecto, el sistema puede configurarse de modo que al menos un elemento de audio esté asociado a al menos una escena, en donde cada elemento de audio está asociado a una posición y/o área en el entorno visual asociado a la escena, en donde el sistema puede ser configurado para solicitar y/o recibir transmisiones a mayor tasa de bits y/o calidad para elementos de audio más cercanos al usuario que para elementos de audio más alejados del usuario.

De acuerdo con un aspecto en el sistema, al menos un elemento de audio puede estar asociado a al menos una escena, al menos un elemento de audio está asociado a una posición y/o área en el entorno visual asociado a la escena, en donde el sistema puede ser configurado para solicitar diferentes transmisiones a diferentes velocidades de bits y/o niveles de calidad para elementos de audio según su relevancia y/o nivel de audibilidad en la posición virtual de cada usuario en la escena, en el que el sistema puede configurarse para solicitar una transmisión de audio a una tasa de bits mayor y/o nivel de calidad para los elementos de audio que son más relevantes y/o más audibles en la posición virtual del usuario actual, y/o una transmisión de audio a una tasa de bits más baja y/o nivel de calidad para los elementos de audio que son menos relevantes y/o menos audibles en la posición virtual del usuario actual.

En una realización en el sistema, al menos un elemento de audio puede estar asociado a una escena, en donde cada elemento de audio está asociado a una posición y/o área en el entorno visual asociado a la escena, en donde el sistema puede configurarse para enviarse periódicamente al servidor la ventana gráfica actual y/o los datos de movimiento y/u orientación de la cabeza y/o metadatos de interacción y/o datos posicionales virtuales, de modo que: para una primera posición, se proporciona una transmisión a mayor tasa de bits y/o calidad, desde el servidor, y para una segunda posición, se proporciona una transmisión a menor tasa de bits y/o calidad, desde el servidor, en donde la primera posición está más cerca del al menos un elemento de audio que la segunda posición.

Una pluralidad de escenas están definidas para múltiples entornos de vídeo, como entornos adyacentes y/o vecinos, de modo que se proporcionen primeras secuencias asociadas a una primera escena actual y, en caso de que el usuario realice la transición a una segunda escena posterior, se proporcionen tanto las secuencias asociadas a la primera escena como las segundas secuencias asociadas a la segunda escena.

Se puede definir una pluralidad de escenas para un primer y un segundo entorno de vídeo, siendo el primero y el segundo entornos adyacentes y/o entornos adyacentes, en donde se proporcionan las primeras secuencias asociadas a la primera escena, desde el servidor, para la reproducción de la primera escena en caso de que la posición del usuario o la posición virtual se encuentre en un primer entorno asociado a la primera escena, se proporcionan las segundas transmisiones asociadas a la segunda escena, desde el servidor, para la reproducción de la segunda escena en caso de la posición del usuario o la posición virtual en un segundo entorno asociado a la segunda escena, y tanto las primeras transmisiones asociadas a la primera escena como las segundas transmisiones asociadas a la segunda escena se proporcionan en caso de que la posición del usuario o la posición virtual estén en una posición de transición entre la primera escena y la segunda escena.

Se puede definir una pluralidad de escenas para un primer y un segundo entornos visuales, que son entornos adyacentes y/o vecinos. El sistema está configurado para solicitar y/o recibir primeras transmisiones asociadas a una primera escena asociada al primer entorno, para la reproducción de la primera escena en caso de que la posición virtual del usuario se encuentre en el primer entorno. El sistema puede configurarse para solicitar y/o recibir segundas transmisiones asociadas a la segunda escena asociada al segundo entorno, para la reproducción de la segunda escena en caso de que la posición virtual del usuario se encuentre en el segundo entorno. El sistema está configurado para solicitar y/o recibir las dos primeras secuencias asociadas a la primera escena y las segundas secuencias asociadas a la segunda escena en el caso la posición virtual del usuario está en una posición de transición entre el primer entorno y el segundo entorno.

El sistema está configurado de modo que las primeras transmisiones asociadas a la primera escena se obtienen a una

velocidad de bits y una calidad superiores cuando el usuario se encuentra en el primer entorno asociado a la primera escena, mientras que las segundas transmisiones asociadas a la segunda escena asociada al segundo entorno se obtienen a una velocidad de bits y una calidad inferiores cuando el usuario se encuentra al principio de una posición de transición de la primera escena a la segunda escena, y las primeras transmisiones asociadas a la primera escena se obtienen a una tasa de bits y una calidad inferiores y las segundas transmisiones asociadas a la segunda escena se obtienen a una tasa de bits y una calidad superiores cuando el usuario se encuentra al final de una posición de transición desde la primera escena a la segunda escena, en la que la tasa de bits y/o la calidad inferiores son inferiores a la tasa de bits y/o la calidad superiores.

De acuerdo con un aspecto, el sistema puede configurarse de modo que se pueda definir una pluralidad de escenas para múltiples entornos, tales como entornos adyacentes y/o vecinos, para que el sistema pueda obtener las transmisiones asociadas a una primera escena actual asociada a un primer entorno actual, y, en caso de que la distancia de la posición del usuario o la posición virtual desde un límite de la escena esté por debajo de un umbral predeterminado, el sistema puede obtener transmisiones de audio asociadas a un segundo entorno adyacente y/o adyacente asociado a la segunda escena.

De acuerdo con un aspecto, el sistema puede configurarse de modo que se pueda definir una pluralidad de escenas para múltiples entornos visuales, de modo que el sistema solicite y/u obtenga las transmisiones asociadas a la escena actual a una tasa de bits y/o calidad superior y las transmisiones asociadas a la segunda escena a una tasa de bits y/o calidad inferior, en donde la tasa de bits y/o calidad inferior es menor que la tasa de bits y/o calidad superior.

De acuerdo con un aspecto, el sistema puede configurarse de modo que se pueda definir una pluralidad de N elementos de audio, y, en caso de que la distancia del usuario a la posición o área de estos elementos de audio sea mayor que un umbral predeterminado, los N elementos de audio se procesan para obtener un número M más pequeño de elementos de audio (M<N) asociados a una posición o área cercana a la posición o área de los N elementos de audio, a fin de proporcionar al sistema al menos una transmisión de audio asociada a los N elementos de audio, en caso de que la distancia del usuario a la posición o área de los N elementos de audio sea menor que un umbral predeterminado, o para proporcionar al sistema al menos una transmisión de audio asociada a los M elementos de audio, en caso de que la distancia del usuario a la posición o el área de los N elementos de audio es mayor que un umbral predeterminado.

De acuerdo con un aspecto, el sistema puede configurarse de modo que al menos una escena de entorno visual esté asociada a al menos una pluralidad de N elementos de audio (N>=2), en donde cada elemento de audio está asociado a una posición y/o área en el entorno visual, en el que al menos al menos una pluralidad de N elementos de audio se proporciona en al menos una representación a un alto nivel de velocidad de bits y/o calidad, y en el que al menos una pluralidad de N elementos de audio se proporciona en al menos una representación a baja tasa de bits y/o nivel de calidad, donde la al menos una representación se obtiene procesando los N elementos de audio para obtener un número menor de elementos de audio (M <N) asociados a una posición o área cercana a la posición o área de los N elementos de audio, en donde el sistema puede configurarse para solicitar la representación a una tasa de bits y/o nivel de calidad más altos para los elementos de audio, en caso de que los elementos de audio sean más relevantes y/o más audibles en la posición virtual del usuario actual en la escena, en donde el sistema puede configurarse para solicitar la representación a menor tasa de bits y/o nivel de calidad para los elementos de audio, en caso de que los elementos de audio sean menos relevantes y/o menos audibles en la posición virtual del usuario actual en la escena.

De acuerdo con un aspecto, el sistema puede configurarse de modo que, en caso de que la distancia del usuario y/o la relevancia y/o el nivel de audibilidad y/o la orientación angular sean inferiores a un umbral predeterminado, se obtengan diferentes a la ventana gráfica actual del usuario y/o la orientación s para los diferentes elementos de audio.

En una forma de realización, el sistema puede configurarse para solicitar y/u obtener las transmisiones en función de la orientación del usuario y/o la dirección del movimiento del usuario y/o las interacciones del usuario en la escena.

En una forma de realización del sistema, la ventana gráfica puede estar asociada a la posición y/o posición virtual y/o datos de movimiento y/u orientación de la cabeza.

De acuerdo con un aspecto, el sistema puede configurarse de manera que se proporcionen diferentes elementos de audio en distintas ventanas gráficas, en donde el sistema puede configurarse para solicitar y/o recibir, en caso de que un primer elemento de audio caiga dentro de una vista, el primer elemento de audio en una mayor tasa de bits que un segundo elemento de audio que no cae dentro de la ventana gráfica.

De acuerdo con un aspecto, el sistema puede configurarse para solicitar y/o recibir primeras transmisiones de audio y segundas transmisiones de audio, en donde los primeros elementos de audio en las primeras transmisiones de audio son más relevantes y/o más audibles que los segundos elementos de audio en las segundas transmisiones de audio, en donde las primeras transmisiones de audio se solicitan y/o reciben a una tasa de bits y/o calidad superior a la velocidad de bits y/o calidad de las segundas transmisiones de audio.

De acuerdo con un aspecto, el sistema puede configurarse de modo que se definan al menos dos escenas del entorno visual, en donde al menos uno de los elementos de audio primero y segundo está asociado a una primera escena asociada a un primer entorno visual, y al menos un tercer elemento de audio es asociado a una segunda escena asociada a un segundo entorno visual, en el que el sistema puede configurarse para obtener metadatos que describen que el al menos un segundo elemento de audio está asociado adicionalmente con la segunda escena del entorno visual, y en el que el sistema puede configurarse para solicitar y/o reciba al menos el primer y segundo elementos de audio, en caso de que la posición virtual del usuario se encuentre en el primer entorno visual, y en el que el sistema pueda configurarse para solicitar y/o recibir al menos el segundo y tercer elementos de audio, en caso de que la posición virtual del usuario está en la segunda escena del entorno visual, y en el que el sistema puede configurarse para solicitar y/o recibir al menos el primero y el segundo y el tercer elemento de audio, en caso de que la posición virtual del usuario esté en transición entre la primera escena del entorno visual y la segunda escena del entorno visual.

Una realización del sistema puede configurarse de modo que el al menos un primer elemento de audio se proporcione en al menos una transmisión de audio y/o conjunto de adaptación, y el al menos un segundo elemento de audio se proporcione en al menos una segunda transmisión de audio y/o conjunto de adaptación, y el al menos un tercer elemento de audio se proporcione en al menos una tercera transmisión de audio y/o conjunto de adaptación, y en donde la al menos primera escena de entorno visual se describa mediante metadatos como una escena completa que requiere al menos la primera y segunda transmisiones de audio y/o conjuntos de adaptación, y en donde la segunda escena de entorno visual es descrita por metadatos como una escena incompleta que requiere el al menos la tercera transmisión de audio y/o conjunto de adaptación y la al menos segunda transmisión de audio y/o conjuntos de adaptación asociados con la al menos primera escena de entorno visual, en donde el sistema comprende un procesador de metadatos configurado para manipular los metadatos, para permitir la fusión de la segunda transmisión de audio perteneciente al primer entorno visual y la tercera transmisión de audio asociada con el segundo entorno visual en una nueva transmisión única, en caso de que la posición virtual del usuario se encuentre en el segundo entorno visual.

De acuerdo con un aspecto, el sistema comprende un procesador de metadatos configurado para manipular los metadatos en al menos una transmisión de audio antes del al menos un decodificador de audio, en función de la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción. y/o datos posicionales virtuales.

De acuerdo con un aspecto, el procesador de metadatos puede configurarse para habilitar y/o deshabilitar al menos un elemento de audio en al menos una transmisión de audio antes del al menos un decodificador de audio, basado en la ventana gráfica actual y/o la orientación y/o los datos del movimiento de la cabeza y/o metadatos de interacción y/o datos posicionales virtuales, en donde el procesador de metadatos puede configurarse para deshabilitar al menos un elemento de audio en al menos una transmisión de audio antes de al menos un decodificador de audio, en caso de que el sistema decida que el elemento de audio ya no debe reproducirse como consecuencia de una vista actual y/o una orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales, y en el que el procesador de metadatos puede configurarse para permitir al menos un elemento de audio en al menos una transmisión de audio antes de al menos un decodificador de audio, en caso de que el sistema decida que el elemento de audio se va a reproducir como consecuencia de la vista actual de un usuario y/o la orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales.

De acuerdo con un aspecto, el sistema puede configurarse para deshabilitar la decodificación de elementos de audio seleccionados en función de la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o la posición virtual.

De acuerdo con un aspecto, el sistema puede configurarse para fusionar al menos una primera transmisión de audio asociada a la escena de audio actual a al menos una transmisión asociada a una escena de audio vecina, adyacente y/o futura.

De acuerdo con un aspecto, el sistema puede configurarse para obtener y/o recopilar datos estadísticos o agregados en la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o datos virtuales de posición, para transmitir una solicitud al servidor asociado a los datos estadísticos o agregados.

De acuerdo con un aspecto, el sistema puede configurarse para desactivar la decodificación y/o la reproducción de al menos una transmisión sobre la base de metadatos asociados a al menos una transmisión y sobre la base de la ventana gráfica actual y/o la orientación de la cabeza del usuario y/o datos de movimiento y/o metadatos y/o datos posicionales virtuales.

De acuerdo con un aspecto, el sistema puede configurarse para: manipular los metadatos asociados con un grupo de transmisiones de audio seleccionadas, en función de al menos la vista actual o estimada del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o los datos de posición virtuales para: seleccionar y/o

habilitar y/o activar elementos de audio que componen la escena de audio a reproducir; y/o habilite la combinación de todas las transmisiones de audio seleccionadas en una sola transmisión de audio.

De acuerdo con un aspecto, el sistema puede configurarse para controlar la solicitud de al menos una transmisión al servidor sobre la base de la distancia de la posición del usuario desde los límites de entornos vecinos y/o adyacentes asociados a diferentes escenas u otras métricas asociadas a la posición del usuario en el entorno actual o las predicciones sobre el entorno futuro.

De acuerdo con un aspecto en el sistema, la información puede proporcionarse desde el sistema de servidor, para cada elemento de audio u objeto de audio, en donde la información incluye información descriptiva sobre las ubicaciones en las que la escena de sonido o los elementos de audio están activos.

De acuerdo con un aspecto, el sistema puede configurarse para elegir entre reproducir una escena y componer o mezclar o multiplexar o superponer o combinar al menos dos escenas sobre la base de la orientación actual o futura o de la ventana gráfica y/o la orientación y/o movimiento, y/o los metadatos y/o la posición virtual y/o la selección de un usuario, donde las dos escenas están asociadas a diferentes entornos vecinos y/o adyacentes.

De acuerdo con un aspecto, el sistema puede configurarse para crear o usar al menos los conjuntos de adaptación de modo que: una serie de conjuntos de adaptación estén asociados con una escena de audio; y/o se proporciona información adicional que relaciona cada conjunto de adaptación con un punto de vista o una escena de audio; y/o se proporciona información adicional que puede incluir: información sobre los límites de una escena de audio y/o información sobre la relación entre un conjunto de adaptación y una escena de audio (por ejemplo, la escena de audio está codificada en tres secuencias que se encapsulan en tres conjuntos de adaptaciones) y/o información sobre la conexión entre los límites de la escena de audio y los múltiples Conjuntos de adaptación.

De acuerdo con un aspecto, el sistema puede configurarse para: recibir una transmisión para una escena asociada a un entorno vecino o adyacente; comience a decodificar y/o reproducir ella ventana gráfica actual del usuario y/o la orientación para el entorno vecino o adyacente en la detección de la transición de un límite entre dos entornos.

De acuerdo con un aspecto, el sistema puede configurarse para funcionar como un cliente y un servidor configurado para entregar transmisiones de vídeo y/o audio para reproducirse en un dispositivo de consumo de medios.

De acuerdo con un aspecto, el sistema puede configurarse para: solicitar y/o recibir al menos un primer conjunto de adaptación que comprende al menos una transmisión de audio asociada con al menos una primera escena de audio; solicitar y/o recibir al menos un segundo conjunto de adaptación que comprenda al menos una segunda transmisión de audio asociada con al menos dos escenas de audio, incluida la al menos una primera escena de audio; y permitir una fusión de la al menos una primera transmisión de audio y la de al menos una segunda transmisión de audio en una nueva transmisión de audio que se decodificará, en función de los metadatos disponibles con respecto a la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o metadatos y/o datos de posición virtual y/o información que describe una asociación de al menos un primer conjunto de adaptación a al menos una primera escena de audio y/o una asociación de al menos un segundo conjunto de adaptación a al menos una primera escena de audio.

De acuerdo con un aspecto, el sistema puede configurarse para recibir información sobre la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o los datos de posición virtuales y/o cualquier información que caracterice los cambios activados por las acciones del usuario; y recibir información sobre la disponibilidad de conjuntos de adaptación e información que describe una asociación de al menos un conjunto de adaptación a al menos una escena y/o punto de vista y/o ventana gráfica y/o posición y/o posición virtual y/o datos de movimiento y/u orientación.

De acuerdo con un aspecto, el sistema puede configurarse para decidir si al menos un elemento de audio de al menos una escena de audio incrustada en al menos una transmisión y al menos un elemento de audio adicional de al menos una escena de audio adicional incrustada en al menos una transmisión adicional deben ser reproducidos; y causar, en caso de una decisión positiva, una operación de fusionar, componer, multiplexar, superponer o combinar al menos una transmisión adicional de la escena de audio adicional a la al menos una transmisión de la al menos una escena de audio.

De acuerdo con un aspecto, el sistema puede configurarse para manipular los metadatos de audio asociados con las transmisiones de audio seleccionadas, en función de al menos la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o los datos de posición virtuales, para: seleccionar y/o habilitar y/o activar los elementos de audio que componen la escena de audio que se decidió reproducir; y habilitar la combinación de todas las transmisiones de audio seleccionadas en una sola secuencia de audio.

De acuerdo con un aspecto, se puede proporcionar un servidor para entregar transmisiones de audio y vídeo a un

cliente para una realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR o entorno de vídeo de 360 grados, las transmisiones de vídeo y audio serán reproducidos en un dispositivo de consumo de medios, en el que el servidor puede comprender un codificador para codificar y/o un almacenamiento para almacenar transmisiones de vídeo para describir un entorno visual, estando asociado el entorno visual a una escena de audio; en el que el servidor puede comprender además un codificador para codificar y/o un almacenamiento para almacenar una pluralidad de secuencias y/o elementos de audio y/o conjuntos de adaptación para ser entregados al cliente, las secuencias y/o elementos de audio y/o conjuntos de adaptación estar asociado a al menos una escena de audio, en donde el servidor está configurado para: seleccionar y entregar una transmisión de vídeo en base a una solicitud del cliente, la transmisión de vídeo está asociada a un entorno; seleccionar una secuencia de audio y/o un elemento de audio y/o un conjunto de adaptación sobre la base de una solicitud del cliente, la solicitud se asocia a al menos la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales y a una escena de audio asociada al entorno; y entregar la transmisión de audio al cliente.

De acuerdo con un aspecto, las secuencias pueden encapsularse en conjuntos de adaptación, donde cada conjunto de adaptación incluye una pluralidad de secuencias asociadas a diferentes representaciones, con diferente tasa de bits y/o calidad, de un mismo contenido de audio, en donde el conjunto de adaptación seleccionado se selecciona en base a de la solicitud del cliente.

Según un aspecto, el sistema puede estar funcionando como un cliente y el servidor.

Según un aspecto el sistema puede incluir un servidor.

De acuerdo con un aspecto, se puede proporcionar un método para un entorno de realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR o vídeo de 360 grados configurado para recibir transmisiones de vídeo y/o audio que se reproducirán en un dispositivo de consumo de medios (por ejemplo, dispositivo de reproducción), que comprende: decodificar señales de vídeo de transmisiones de vídeo para la representación de escenas de entornos de vídeo VR, AR, MR o 360 grados para un usuario, y decodificar señales de audio de transmisiones de audio, solicitando y/u obteniendo de un servidor al menos una transmisión de audio en base a la ventana gráfica actual del usuario y/o datos posicionales y/u orientación de la cabeza y/o datos de movimiento y/o metadatos y/o datos virtuales y/o metadatos de posición.

De acuerdo con un aspecto, se puede proporcionar un programa informático que comprende instrucciones que, cuando son ejecutadas por un procesador, hacen que el procesador realice el método anterior.

Figuras

Las figuras 1.1–1.8 muestran ejemplos de la invención.

Las figuras 2-6 muestran escenarios de la invención.

Las figuras 7A-8B muestran métodos inventivos.

Aspectos inventivos

A continuación (por ejemplo, las figuras 1.1 y ss) se describen ejemplos de sistemas de acuerdo con aspectos inventivos.

Los ejemplos de un sistema de la invención (que se puede materializar mediante los diferentes ejemplos que se describen a continuación) se indican colectivamente con 102. Un sistema 102 puede ser un sistema de cliente, por ejemplo, como puede obtenerse de un sistema de servidor (por ejemplo, 120) de audio y/o transmisiones de vídeo para la representación de escenas de audio y/o entornos visuales a un usuario. El sistema de cliente 102 también puede recibir metadatos del sistema de servidor 120 que proporcionan, por ejemplo, información lateral y/o auxiliar con respecto a las transmisiones de audio y/o vídeo.

El sistema 102 puede estar asociado (o comprender en algunos ejemplos) a un dispositivo de consumo de medios (MCD) que en realidad reproduce señales de audio y/o vídeo a un usuario. En algunos ejemplos, el usuario puede usar el MCD.

El sistema 102 puede realizar solicitudes al servidor 120, donde las solicitudes están asociadas a al menos la ventana gráfica actual y/o la orientación de la cabeza (por ejemplo, la orientación angular) y/o los datos de movimiento y/o los metadatos de interacción y/o los datos virtuales de posición 110 (pueden proporcionarse varias métricas). Los datos de orientación y/u orientación de la cabeza y/o movimiento y/o metadatos de interacción y/o datos de posición virtuales 110 pueden proporcionarse en la retroalimentación desde el MCD al sistema de cliente 102 que, a su vez, puede proporcionar la solicitud al sistema de servidor 120 sobre la base de esta retroalimentación.

En algunos casos, la solicitud (que se indica con 112) puede contener la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o los datos de posición virtuales 110 (o una indicación o una versión procesada de ellos). Sobre la base de la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o los datos posicionales virtuales 110, el sistema de servidor 120 proporcionará las transmisiones de audio y/o vídeo y/o los metadatos necesarios. En este caso, el sistema de servidor 120 puede tener conocimiento de la posición del usuario (por ejemplo, en un entorno virtual) y puede asociar las transmisiones correctas a las posiciones del usuario.

En otros casos, la solicitud 112 del sistema de cliente 102 puede contener solicitudes explícitas de transmisiones de audio y/o vídeo en particular. En este caso, la solicitud 112 puede basarse en la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o los datos de posición virtual 110. El sistema de cliente 102 tiene el conocimiento de las señales de audio y vídeo que deben ser entregadas al usuario, incluso si el sistema de cliente 102 no tiene almacenados en el mismo las transmisiones necesarias. El sistema de cliente 102 puede tratar, en ejemplos, transmisiones particulares en el sistema de servidor 120

El sistema de cliente 102 puede ser un sistema para un entorno de realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR o vídeo de 360 grados configurado para recibir transmisiones de vídeo y audio para ser reproducidas en un dispositivo de consumo de medios, en el que el sistema 102 comprende:

al menos un decodificador de vídeo multimedia configurado para decodificar señales de vídeo de transmisiones de vídeo para la representación de escenas de entornos de vídeo VR, AR, MR o 360 grados para un usuario, y

al menos un decodificador de audio 104 configurado para decodificar señales de audio (108) de al menos una transmisión de audio 106,

en el que el sistema 102 está configurado para solicitar 112 al menos una transmisión de audio 106 y/o un elemento de audio de una transmisión de audio y/o un conjunto de adaptación a un servidor 120 sobre la base de al menos la ventana gráfica actual y/o la orientación del usuario y/o datos de movimiento y/o metadatos de interacción y/o datos posicionales virtuales 110.

Se debe tener en cuenta que en entornos VR, AR, MR, el usuario 140 puede significar que está en un entorno particular (por ejemplo, una habitación particular). El entorno se describe con señales de vídeo que están codificadas, por ejemplo, en el lado del servidor (lado del sistema del servidor 120, que no incluye necesariamente el sistema del servidor 120, pero que puede comprender un codificador diferente que ha codificado previamente transmisiones de vídeo que se han almacenado luego en un almacenamiento del servidor 120). En cada instante, en algunos ejemplos, el usuario puede disfrutar solo de algunas señales de vídeo (por ejemplo, la ventana gráfica).

En términos generales, cada entorno puede estar asociado a una escena de audio en particular. La escena de audio puede entenderse como la colección de todos los sonidos que deben reproducirse para el usuario en el entorno particular y durante un período de tiempo determinado.

Tradicionalmente, se ha entendido que los entornos están en un número discreto. En consecuencia, el número de ambientes ha sido entendido como finito. Por las mismas razones, el número de escenas de audio ha sido entendido como finito. Por lo tanto, en la técnica anterior, los sistemas VR, AR, MR se han diseñado de manera que:

- El usuario esté destinado a estar en un solo entorno en cada momento; por lo tanto, para cada entorno:

◦ El sistema de cliente 102 solicita al sistema de servidor 120 sólo las transmisiones de vídeo asociadas al entorno único;

◦ El sistema de cliente 102 solicita al sistema de servidor 120 sólo las transmisiones de audio asociadas a la escena única.

Este enfoque ha llevado a inconvenientes.

Por ejemplo, todas las transmisiones de audio se entregarán todas juntas al sistema de cliente 102 para cada escena/entorno, y las transmisiones de audio completamente nuevas deben entregarse cuando el usuario se traslada a un entorno diferente (por ejemplo, cuando el usuario pasa por un puerta, lo que implica una transmisión de ambientes/escenas).

Además, la experiencia antinatural ha sido causada en algunos casos: por ejemplo, cuando un usuario está cerca de una pared (por ejemplo, una pared virtual de una sala virtual), debería experimentar sonidos provenientes del otro lado de la pared. Sin embargo, esta experiencia es imposible con los entornos tradicionales: la colección de transmisiones

de audio asociadas a la escena actual obviamente no contiene ninguna secuencia asociada a los entornos/escenas adyacentes.

Por otro lado, la experiencia del usuario generalmente mejora cuando aumenta la tasa de bits de las transmisiones de audio. Esto puede causar problemas adicionales: cuanto mayor sea la tasa de bits, mayor será la carga útil que el sistema del servidor debe entregar al sistema de cliente 102. Por ejemplo, cuando una escena de audio contiene varias fuentes de audio (transportadas como elementos de audio), algunas de ellas ubicadas cerca para la posición del usuario y otras alejadas, las fuentes de sonido ubicadas lejos serían menos audibles. Por lo tanto, la entrega de todos los elementos de audio a la misma tasa de bits o nivel de calidad puede llevar a tasas de bits muy altas. Esto implica una transmisión de audio no eficaz. Si el sistema del servidor 120 entrega las transmisiones de audio a la tasa de bits más alta posible, se produce una entrega ineficiente, ya que los sonidos con un bajo nivel de audibilidad o poca relevancia para la escena de audio en general requerirían una tasa de bits alta, de manera similar a los sonidos relevantes generados más cerca del usuario. Por lo tanto, si todas las transmisiones de audio de una escena se entregan a la tasa de bits más alta, la comunicación entre el sistema del servidor 120 y el sistema de cliente 102 aumentaría innecesariamente la carga útil. Si todas las transmisiones de audio de una escena se entregan a una tasa de bits inferior, la experiencia del usuario no será satisfactoria.

Los problemas de comunicación exacerban el inconveniente descrito anteriormente: cuando un usuario pasa por una puerta, se supone que debe cambiar instantáneamente el entorno/la escena, lo que requeriría que, instantáneamente, el sistema 120 del servidor proporcione todas las corrientes al sistema de cliente. 102.

Por lo tanto, tradicionalmente no era posible resolver los problemas discutidos anteriormente.

Sin embargo, con la invención, es posible resolver estos problemas: el sistema de cliente 102 proporciona una solicitud al sistema de servidor 120 que también se puede basar en la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o metadatos de interacción y/o datos posicionales virtuales (y no sólo en función del entorno/escena). En consecuencia, el sistema de servidor 120 puede proporcionar, para cada instante, las transmisiones de audio que se van a representar para cada una, por ejemplo, la posición del usuario.

Por ejemplo, si el usuario nunca se acerca al muro, no es necesario que el sistema de cliente 102 solicite las transmisiones del entorno vecino (por ejemplo, el sistema de cliente 102 puede solicitarlo solo cuando el usuario se acerca al muro). Además, las corrientes provenientes del exterior de la pared pueden tener una tasa de bits reducida, ya que pueden escucharse a un volumen bajo. En particular, el sistema 120 del servidor puede entregar transmisiones más relevantes (por ejemplo, transmisiones provenientes de objetos de audio dentro del entorno actual) al sistema de cliente 102 con la mayor tasa de bits y/o el nivel de calidad más alto (como consecuencia del hecho de que las transmisiones menos relevantes tienen una tasa de bits y/o un nivel de calidad más bajos, lo que deja la banda libre para las transmisiones más relevantes).

Se puede obtener un nivel de calidad inferior, por ejemplo, reduciendo la tasa de bits o procesando los elementos de audio de tal manera que se reduzcan los datos requeridos para ser transmitidos, mientras que la tasa de bits utilizada por señal de audio se mantiene constante. Por ejemplo, si un número de 10 objetos de audio se ubican en diferentes posiciones muy alejadas del usuario, estos objetos se pueden mezclar en un número menor de señales según la posición del usuario:

- En las posiciones muy alejadas de la posición del usuario (por ejemplo, más alta que un primer umbral), los objetos se mezclan en 2 señales (otros números son posibles, en función de su posición espacial y semántica) y se entregan como 2 "objetos virtuales".

- En las posiciones más cercanas a la posición del usuario (por ejemplo, más bajo que el primer umbral pero más alto que un segundo umbral más pequeño que el primer umbral), los objetos se mezclan en 5 señales (según su posición espacial y semántica) y se entregan como 5 (otros números) son posibles) "objetos virtuales"

- En las posiciones muy cercanas a las posiciones del usuario (inferiores a los umbrales primero y segundo), los 10 objetos se entregan como 10 señales de audio que brindan la más alta calidad.

Mientras que para la más alta calidad las señales de audio pueden considerarse muy importantes y audibles, el usuario puede localizar individualmente cada objeto. Para los niveles de calidad más bajos en las posiciones más alejadas, algunos de los objetos de audio pueden volverse menos relevantes o menos audibles, por lo que el usuario no podría, de ninguna manera, localizar individualmente las señales de audio en el espacio y, por lo tanto, reducir el nivel de calidad para la entrega de estas señales de audio no conducirían a una reducción de la calidad de la experiencia para el usuario.

Otro ejemplo es cuando el usuario va más allá de una puerta: en la posición de transición (por ejemplo, en el límite entre dos entornos/escenas diferentes), el sistema 120 del servidor proporcionará las transmisiones de ambas

escenas/entorno, pero a velocidades de bits más bajas. Esto se debe a que el usuario experimentará sonidos provenientes de dos entornos diferentes (los sonidos pueden fusionarse a partir de diferentes transmisiones de audio originalmente asociadas a distintas escenas/entornos) y no existe la necesidad de contar con el nivel de calidad más alto de cada fuente de sonido (o elemento de audio).

En vista de lo anterior, la invención permite ir más allá del enfoque tradicional del número discreto de entornos visuales y escenas de audio, pero puede permitir una representación gradual de diferentes entornos/escenas, dando una experiencia más realista al usuario.

A continuación, se considera que cada entorno visual (por ejemplo, el entorno virtual) está asociado a una escena de audio (los atributos de los entornos también pueden ser atributos de la escena). Cada entorno/escena puede estar asociado, por ejemplo, a un sistema de coordenadas geométricas (que puede ser un sistema de coordenadas geométricas virtuales). El entorno/escena puede tener límites, de modo que, cuando la posición del usuario (por ejemplo, la posición virtual) va más allá de los límites, se alcanza un entorno/escena diferente. Los límites pueden basarse en el sistema de coordenadas utilizado. El entorno puede comprender objetos de audio (elementos de audio, fuentes de sonido) que pueden colocarse en algunas coordenadas particulares del entorno/escena. Con respecto, por ejemplo, a la posición relativa y/o la orientación del usuario con respecto a los objetos de audio (elementos de audio, fuentes de sonido), el sistema de cliente 102 puede solicitar diferentes transmisiones y/o el sistema de servidor 120 puede proporcionar diferentes transmisiones (por ejemplo, a tasas de bits y/o niveles de calidad más altos/bajos de acuerdo con la distancia y/o la orientación).

Más en general, el sistema de cliente 102 puede solicitar y/u obtener desde el sistema del servidor 120 transmisiones diferentes (por ejemplo, diferentes representaciones de los mismos sonidos a diferentes tasas de bits y/o niveles de calidad) sobre la base de su audibilidad y/o relevancia. La audibilidad y/o la relevancia se pueden determinar, por ejemplo, sobre la base de al menos la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o los datos virtuales de posición.

En varios ejemplos, existe la posibilidad de fusionar diferentes transmisiones. En varios casos, existe la posibilidad de componer, mezclar, multiplexar, superponer o combinar al menos dos escenas. Existe, por ejemplo, la posibilidad de usar un mezclador y/o renderizador (que puede, por ejemplo, usarse corriente abajo de varios decodificadores, cada uno decodificando al menos una transmisión de audio), o realizar una operación de transmisión de secuencias, por ejemplo, corriente arriba de la decodificación de las transmisiones. En otros casos, puede existir la posibilidad de decodificar diferentes transmisiones y renderizarlos con diferentes configuraciones de altavoces.

Cabe señalar que la presente invención no rechaza necesariamente el concepto de entorno visual y escena de audio. En particular, con la invención, las transmisiones de audio y vídeo asociadas a una escena/entorno particular pueden ser enviadas desde el sistema de servidor 120 al sistema de cliente 102 cuando el usuario ingresa a un entorno/escena. No obstante, dentro del mismo entorno/escena, se pueden solicitar, direccionar y/o entregar diferentes audios y/u objetos de audio y/o conjuntos de adaptación. En particular, puede existir la posibilidad de que:

- al menos algunos de los datos de vídeo asociados al entorno visual se envían desde el servidor 120 al cliente 102 en la entrada del usuario a una escena; y/o

- al menos algunos de los datos de audio (transmisiones, objetos, conjuntos de adaptación...) se envían al sistema de cliente 102 sólo sobre la base de la ventana gráfica actual (o futura) y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o posición virtual y/o selección/interacción de un usuario; y/o

-(en algunos casos): algunos datos de audio se envían al sistema de cliente 102 en función de la escena actual (independientemente de la posición actual o futura o de la ventana gráfica y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o la posición virtual y/o la selección de un usuario), mientras que los datos de audio restantes se entregan sobre la base de la orientación actual y futura o de la vista y/o la cabeza

- datos de orientación y/o movimiento y/o metadatos y/o posición virtual y/o selección de un usuario.

Debe notarse que los diversos elementos (sistema de servidor, sistema de cliente, MCD, etc.) pueden representar elementos en diferentes dispositivos de hardware o incluso en los mismos (por ejemplo, el cliente y el MCD pueden implementarse como parte del mismo teléfono móvil), o de modo similar, el cliente puede estar en una PC conectada a una pantalla secundaria que comprenda el MCD).

Ejemplos

Una forma de realización del sistema 102 (cliente) como se muestra en la **figura 1.1** está configurada para recibir transmisiones (audio) 106 sobre la base de una posición definida en un entorno (por ejemplo, un entorno virtual), que puede entenderse como asociado a una escena de vídeo y audio (en adelante, escena 150). Las diferentes posiciones

en la misma escena 150 en general implican que diferentes transmisiones 106 o diferentes metadatos asociados a las transmisiones 106 se proporcionen a un decodificador de audio 104 del sistema 102 (desde un servidor de medios 120, por ejemplo). El sistema 102 está conectado a un dispositivo de consumo de medios (MCD) desde el cual recibe retroalimentación asociada a la posición y/o posición virtual del usuario en el mismo entorno. A continuación, la posición del usuario en el entorno puede asociarse con la ventana gráfica particular que disfruta el usuario (la ventana gráfica se destina, por ejemplo, a la superficie, hipotética como una superficie rectangular proyectada en una esfera, que se representa para el usuario).

En un escenario de ejemplo, cuando el usuario se mueve en la escena 150 de VR, AR y/o MR, se puede imaginar que el contenido de audio está virtualmente generado por una o más fuentes de audio 152, que pueden cambiar. Las fuentes de audio 152 pueden entenderse como fuentes de audio virtuales, en el sentido de que pueden referirse a posiciones en el entorno virtual: la representación de cada fuente de audio se adapta a la posición del usuario (por ejemplo, en una ejemplificación simplificada, el nivel de la fuente de audio es más alta cuando el usuario está más cerca de la posición de la fuente de audio, y más baja cuando el usuario está más alejado de la fuente de audio). No obstante, cada elemento de audio (fuente de audio) está codificado en transmisiones de audio que se proporcionan al decodificador. Las transmisiones de audio pueden estar asociadas a varias posiciones y/o áreas en la escena. Por ejemplo, las fuentes de audio 152 que no son audibles en una escena pueden volverse audibles en la siguiente escena, por ejemplo, cuando se abre una puerta en la escena VR, AR y/o MR 150. El usuario puede entonces seleccionar ingresar a una nueva escena/entorno 150 (por ejemplo, una sala), y la escena de audio cambia. Con el fin de describir este escenario, el término de puntos de vista discretos en el espacio se puede utilizar, como una ubicación discreta en el espacio (o en el entorno VR), para el cual está disponible un contenido de audio diferente.

En términos generales, el servidor de medios 120 puede proporcionar transmisiones 106 asociadas a la escena particular 150 en base a la posición del usuario en la escena 150. Las transmisiones 106 pueden estar codificadas por al menos un codificador 154 y proporcionadas al servidor de medios 120. El servidor 120 de medios puede pasar las transmisiones 113 con comunicaciones 113 (por ejemplo, a través de una red de comunicación). La provisión de las transmisiones 113 puede basarse en las solicitudes 112 establecidas por el sistema 102 en base a la posición 110 del usuario (por ejemplo, en el entorno virtual). La posición 110 del usuario también puede entenderse como asociada a la ventana gráfica que disfruta el usuario (como para cada posición, hay un solo rectángulo que está representado) y al punto de vista (ya que el punto de vista es el centro de la ventana gráfica). Por lo tanto, la provisión de la ventana gráfica puede ser, en algunos ejemplos, la misma que la provisión de la posición.

El sistema 102, como se muestra en la **figura 1.2,** está configurado para recibir transmisiones (audio) 113 sobre la base de otra configuración en el lado del cliente. En esta implementación de ejemplo en el lado de la codificación, se proporciona una pluralidad de codificadores de medios 154 que se pueden usar para crear una o más secuencias 106 para cada escena disponible 150 asociada con una parte de escena de sonido de un punto de vista.

El servidor 120 de medios puede almacenar múltiples juegos de adaptación de audio y vídeo (no mostrado) que comprenden diferentes codificaciones de las mismas transmisiones de audio y vídeo a diferentes tasas de bits. Además, el servidor de medios puede contener información descriptiva de todos los conjuntos de adaptación, que puede incluir la disponibilidad de todos los conjuntos de adaptación creados. Los conjuntos de adaptación pueden incluir también información que describe una asociación de un conjunto de adaptación a una escena de audio y/o punto de vista en particular. De esta manera, cada conjunto de adaptación puede estar asociado con una de las escenas de audio disponibles.

Los conjuntos de adaptación pueden incluir, además, información que describa los límites de cada escena de audio y/o punto de vista que puede contener, por ejemplo, una escena de audio completa o sólo objetos de audio individuales. Los límites de una escena de audio se pueden definir, por ejemplo, como coordenadas geométricas de una esfera (por ejemplo, centro y radio).

El sistema 102 en el lado del cliente puede recibir información sobre la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o la posición virtual del usuario o cualquier información que caracterice los cambios activados por las acciones del usuario. Además, el sistema 102 puede recibir también información sobre la disponibilidad de todos los conjuntos de adaptación e información que describe una asociación de un conjunto de adaptación a una escena de audio y/o punto de vista; y/o información que describa los "límites" de cada escena de audio y/o punto de vista (que puede contener, por ejemplo, una escena de audio completa o sólo objetos individuales). Por ejemplo, dicha información se puede proporcionar como parte de la sintaxis XML de descripción de presentación de medios (MPD) en el caso de un entorno de entrega DASH.

El sistema 102 puede proporcionar una señal de audio al dispositivo de consumo de medios (MCD) utilizado para el consumo de contenido. El dispositivo de consumo de medios también es responsable de recopilar información sobre la ubicación y/o la orientación del usuario y/o la dirección del movimiento (o cualquier información que caracterice los cambios activados por las acciones del usuario) como datos de posición y transición 110.

Un procesador de visualización 1232 puede configurarse para recibir dichos datos de posición y transición 110 desde el lado del dispositivo de consumo de medios. El procesador de la ventana gráfica 1232 también puede recibir información y la ROI señalada en los metadatos y toda la información disponible en el extremo receptor (sistema 102). El procesador de la ventana gráfica 1232 puede entonces decidir en función de toda la información recibida y/o derivada de los metadatos recibidos y/o disponibles, qué punto de vista de audio se debe reproducir en un momento determinado. Por ejemplo, el procesador de la ventana gráfica 1232 puede decidir que se reproduzca una escena de audio completa, se debe crear una nueva escena de audio 108 de todas las escenas de audio disponibles, por ejemplo, sólo se reproducirán algunos elementos de audio de múltiples escenas de audio, mientras que otros elementos de audio restantes de estas escenas de audio no deben reproducirse. El procesador de la ventana gráfica 1232 también puede decidir si se debe reproducir una transición entre dos o más escenas de audio.

Puede proporcionarse una parte de selección 1230 para seleccionar, basándose en la información recibida desde el procesador de la ventana gráfica 1232 uno o más, conjuntos de adaptación de los conjuntos de adaptación disponibles como se indica en la información recibida por el extremo receptor; los conjuntos de adaptación seleccionados describen completamente la escena de audio que debe reproducirse en la ubicación actual del usuario. Esta escena de audio puede ser una escena de audio completa tal como se define en el lado de la codificación o puede crearse una nueva escena de audio a partir de todas las escenas de audio disponibles.

Además, en el caso de que una transición entre dos o más escenas de audio suceda según la indicación del procesador de la ventana gráfica 1232, la parte de selección puede configurarse para seleccionar uno o más conjuntos de adaptación de los conjuntos de adaptación disponibles como se señaló en la información recibida por el extremo receptor; los conjuntos de adaptación seleccionados describen completamente la escena de audio que puede ser requerida para ser reproducida en un futuro próximo (por ejemplo, si el usuario camina en la dirección de la siguiente escena de audio con cierta velocidad, se puede predecir que se requiere la siguiente escena de audio y se selecciona antes de la reproducción).

Además, algunos conjuntos de adaptación correspondientes a las ubicaciones vecinas se pueden seleccionar primero a una tasa de bits inferior y/o a un nivel de calidad inferior, por ejemplo, se elige una representación codificada a una tasa de bits inferior entre las representaciones disponibles en un conjunto de adaptación, y en función de los cambios de posición, la calidad aumenta al seleccionar una tasa de bits mayor para esos conjuntos de adaptación específicos, por ejemplo una representación codificada a una tasa de bits más alta se selecciona de las representaciones disponibles en un conjunto de adaptación.

Se puede proporcionar una parte de descarga y conmutación 1234 para solicitar, en función de la indicación recibida de la parte de selección, uno o más, conjuntos de adaptación de los conjuntos de adaptación disponibles del servidor de medios, que están configurados para recibir, uno o más, conjuntos de adaptación de los conjuntos de adaptación disponibles del servidor de medios y extraen información de metadatos de todas las transmisiones de audio recibidos.

Se puede proporcionar un procesador de metadatos 1236 para recibir de la descarga y la información de conmutación sobre las transmisiones de audio recibidas, información que puede incluir los metadatos de audio correspondientes a cada secuencia de audio recibida. El procesador de metadatos 1236 también se puede configurar para procesar y manipular los metadatos de audio asociados con cada transmisión de audio 113, en función de la información recibida del procesador de la ventana gráfica 1232 que puede incluir información sobre la ubicación y/u orientación del usuario y/o la dirección del movimiento 110, con el fin de seleccionar/habilitar los elementos de audio necesarios 152 que componen la nueva escena de audio según lo indicado por el procesador de la ventana gráfica 1232, permite la fusión de todas las transmisiones de audio 113 en una sola secuencia de audio 106.

Un mezclador/fusionador 1238 puede configurarse para fusionar todas las transmisiones de audio seleccionadas en una transmisión de audio 106 en función de la información recibida del procesador de metadatos 1236 que puede incluir los metadatos de audio procesados y modificados correspondientes a todas las transmisiones de audio 113 recibidos.

El decodificador de medios 104 configurado para recibir y decodificar al menos una transmisión de audio para la reproducción de la nueva escena de audio como lo indica el procesador de la ventana gráfica 1232 en función de la información sobre la ubicación del usuario y/o la orientación y/o la dirección del movimiento.

En otra forma de realización, el sistema 102 como se muestra en la **figura 1.7** puede configurarse para recibir transmisiones de audio 106 a diferentes tasas de bits de audio y/o niveles de calidad. La configuración del hardware de esta forma de realización es similar a la de la figura 1.2. Al menos una escena del entorno visual 150 puede estar asociada a al menos una pluralidad de N elementos de audio (N>=2), cada elemento de audio está asociado a una posición y/o área en el entorno visual. La al menos una pluralidad de N elementos de audio 152 se proporciona en al menos una representación a un alto nivel de velocidad de bits y/o nivel de calidad, y en donde la al menos una pluralidad de N elementos de audio 152 se proporciona en al menos una representación en baja tasa de bits y/o nivel de calidad, donde la al menos una representación se obtiene procesando los N elementos de audio 152 para obtener

un número menor de elementos de audio 152 (M<N) asociados a una posición o área cercana a la posición o área de los N elementos de audio 152.

El procesamiento de los N elementos de audio 152 podría ser, por ejemplo, una simple adición de las señales de audio o podría ser una mezcla activa basada en su posición espacial 110 o la representación de señales de audio usando su posición espacial a una nueva posición virtual ubicada entre las señales de audio. El sistema puede configurarse para solicitar la representación a un mayor nivel de velocidad de bits y/o calidad para los elementos de audio, en caso de que los elementos de audio sean más relevantes y/o más audibles en la posición virtual del usuario actual en la escena, en donde el sistema está configurado para solicitar la representación a menor tasa de bits y/o nivel de calidad para los elementos de audio, en caso de que los elementos de audio sean menos relevantes y/o menos audibles en la posición virtual del usuario actual en la escena.

La figura 1.8 muestra un ejemplo de un sistema (que puede ser el sistema 102) que muestra un sistema 102 para una realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR o entorno de vídeo de 360 grados configurado para recibir transmisiones de vídeo 1800 y transmisiones de audio 106 para ser reproducidas en un dispositivo de consumo de medios, en el que el sistema 102 puede comprender:

al menos un decodificador de vídeo de medios 1804 configurado para decodificar señales de vídeo 1808 de transmisiones de vídeo 1800 para la representación de un entorno de vídeo VR, AR, MR o 360 grados para un usuario, y

al menos un decodificador de audio 104 configurado para decodificar señales de audio 108 de al menos una transmisión de audio 106.

El sistema 102 puede configurarse para solicitar (112) al menos una transmisión de audio 106 y/o un elemento de audio de una transmisión de audio y/o un conjunto de adaptación a un servidor (por ejemplo, 120) sobre la base de al menos la ventana gráfica actual del usuario y/o datos de movimiento y/u orientación de la cabeza y/o metadatos de interacción y/o datos de posición virtuales 110 (por ejemplo, proporcionados como retroalimentación del dispositivo 180 de consumo de medios).

El sistema 102 puede ser el mismo del sistema 102 de las figuras 1.1-1.7 y/o obtener los escenarios de las figuras. 2a y ss.

Los ejemplos actuales también se refieren a un método para un entorno de realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR o vídeo de 360 grados configurado para recibir transmisiones de vídeo y/o audio para reproducirse en un dispositivo de consumo de medios [por ejemplo, dispositivo de reproducción], que comprende:

decodificar señales de vídeo de transmisiones de vídeo para la representación de escenas de entornos de vídeo VR, AR, MR o 360 grados a un usuario, y

decodificando señales de audio de transmisiones de audio,

solicitar y/u obtener de un servidor, al menos una transmisión de audio sobre la base de la ventana gráfica actual y/o los datos posicionales y/o los datos de movimiento y/u orientación de la cabeza y/o metadatos y/o datos virtuales posicionales y/o metadatos.

Caso 1

Diferentes escenas/entornos 150 en general implican la recepción de distintas transmisiones 106 de un servidor 120. Sin embargo, las transmisiones 106 recibidas por el decodificador de audio 104 también pueden estar condicionados por la posición del usuario en la misma escena 150.

En un primer momento (de inicio) ($t=t_1$) que se muestra en la **figura 2a**, un usuario está posicionado, por ejemplo, en una escena 150, que tiene una primera posición definida en un entorno VR (o entorno AR, o entorno MR). En el sistema cartesiano XYZ (por ejemplo, horizontal), una primera ventana gráfica (posición) 110' del usuario está asociada con las coordenadas x'_u e y'_u (el eje Z está orientado aquí para salir del papel). En esta primera escena 150, se ubican dos elementos de audio 152-1 y 152-1, que tienen las coordenadas respectivas x'_1 e y'_1 para el elemento de audio 1 (152-1) y x'_2 e y'_2 para el elemento de audio 2 (152-2). La distancia d'_1 del usuario al elemento de audio 1 (152-1) es menor que la distancia d'_2 (152-1) del usuario al elemento de audio 2. Todos los datos de la posición del usuario (vista) son transmitidos desde el MCD al sistema 102.

En un segundo momento de ejemplo ($t=t_2$) que se muestra en la **figura 2b**, el usuario está posicionado, por ejemplo, en la misma escena 150, pero en una segunda posición diferente. En el sistema cartesiano XY, una segunda ventana gráfica (posición) 110" del usuario está asociada con las nuevas coordenadas $x"_u$ e $y"_u$ (el eje Z está orientado aquí

para salir del papel). Ahora la distancia d"$_1$ del usuario desde el elemento de audio 1 (152-1) es mayor que la distancia d"$_2$ del usuario desde el elemento de audio 2 (152-2). Todos los datos de posición del usuario (ventana gráfica) se transmiten de nuevo desde el MCD al sistema 102.

El usuario, equipado con dicho MCD para visualizar una determinada ventana gráfica en un entorno de 360 grados, puede escuchar, por ejemplo, a través de auriculares. El usuario puede disfrutar de la reproducción de diferentes sonidos para diferentes posiciones representadas en las figura 2a y 2b de la misma escena 150.

Cualquier posición y/o cualquier transición y/o ventana gráfica y/o posición virtual y/u orientación de la cabeza y/o datos de movimiento dentro de la escena, por ejemplo, de la figura 2a a 2b puede transmitirse periódicamente (por ejemplo, en retroalimentación) desde el MCD al sistema 102 (cliente) como señal 110. El cliente puede retransmitir los datos de posición y transición 110' o 110" (por ejemplo, datos de la ventana gráfica) al servidor 120. El cliente 102 o el servidor 120 pueden decidir en función de la posición y los datos de transición 110' o 110" (por ejemplo, datos de la ventana gráfica) qué transmisiones de audio 106 se requieren para reproducir la escena de audio correcta en la posición actual del usuario. El cliente podría decidir y transmitir una solicitud 112 para una transmisión de audio 106 correspondiente, mientras que el servidor 120 puede configurarse para entregar las transmisiones 106 según la información de posición proporcionada por el cliente (sistema 102). Alternativamente, el servidor 120 podría decidir y entregar en consecuencia las transmisiones 106 según la información de posición proporcionada por el cliente (sistema 102).

El cliente (sistema 102) puede solicitar que se decodifique la transmisión de las transmisiones para representar la escena 150. En algunos ejemplos, el sistema 102 puede transmitir información sobre el nivel de calidad más alto que se reproducirá en el MCD (en otros ejemplos, es el servidor 120 que decide el nivel de calidad que se reproducirá en el MCD, sobre la base de la posición del usuario en la escena). En respuesta, el servidor 120 puede seleccionar una de una multitud de representaciones asociadas con la escena de audio a representar, para entregar al menos una transmisión 106 de acuerdo con la posición 110' o 110" del usuario. Por lo tanto, el cliente (sistema 102) puede estar configurado para entregar, por ejemplo, a través de un decodificador de audio 104, una señal de audio 108 al usuario para reproducir el sonido asociado con su posición real (efectiva) 110' o 110". (pueden utilizarse conjuntos de adaptación 113: pueden utilizarse diferentes variantes de las mismas transmisiones, por ejemplo, a diferentes velocidades binarias, para diferentes posiciones del usuario).

Las transmisiones 106 (que pueden preprocesarse o generarse sobre la marcha) pueden transmitirse al cliente (sistema 102) y pueden configurarse para una multitud de puntos de vista asociados con ciertas escenas de sonido.

Se ha observado que se pueden proporcionar diferentes calidades (por ejemplo, diferentes tasas de bits) para diferentes transmisiones 106 de acuerdo con la posición particular (por ejemplo, 110' o 110") del usuario en el entorno (por ejemplo, virtual). Por ejemplo: En el caso de una pluralidad de fuentes de audio 152-1 y 152-2, cada fuente de audio 152-1 y 152-2 puede asociarse a una posición particular dentro de la escena 150. Cuanto más cerca esté la posición del usuario 110' o 110' para la primera fuente de audio 152-1, mayor será la resolución y/o la calidad de la transmisión asociada a la primera fuente de audio 152-2. Este caso de ejemplo puede aplicarse al elemento de audio 1 (152-1) en la figura 2a, así como al elemento de audio 2 (152-2) en la figura 2b. Cuanto más alejada esté la posición 110 del usuario de la segunda fuente de audio 152-2, menor será la resolución necesaria de la transmisión 106 asociada a la segunda fuente de audio 152-2. Este caso de ejemplo puede aplicarse al elemento de audio 2 (152-2) en la figura 2a, así como al elemento de audio 1 (152-1) en la figura 2b.

De hecho, la primera, una fuente de audio cercana debe escucharse a un nivel más alto (y, por lo tanto, debe proporcionarse a una tasa de bits más alta), mientras que la segunda, una fuente de audio lejana debe escucharse a un nivel más bajo (lo que puede permitir que se requiera una menor resolución);

Por lo tanto, sobre la base de la posición 110' o 110" en el entorno provisto por el cliente 102, el servidor 120 puede proporcionar diferentes transmisiones 106 a diferentes tasas de bits (u otra calidad). Basado en el hecho de que los elementos de audio que están lejos no requieren altos niveles de calidad, la experiencia global de la experiencia del usuario se conserva incluso si se entregan a un nivel de bits o calidad inferior.

Por lo tanto, se pueden usar diferentes niveles de calidad para algunos elementos de audio en diferentes posiciones del usuario, al tiempo que se preserva la calidad de la experiencia.

Sin esta solución, todas las transmisiones 106 deberían ser proporcionadas por el servidor 120 al cliente a la tasa de bits más alta, lo que aumentaría la carga útil en el canal de comunicación desde el servidor 120 al cliente.

Caso 2

La **figura 3** (caso 2) muestra una forma de realización con otro escenario a modo de ejemplo (representado en un plano vertical XZ de un espacio XYZ, donde el eje Y se representa cuando entra en el papel), en donde el usuario se

mueve en un primer VR, AR y/o MR escena A (150A), abre una puerta y camina (transición 150AB) a través, lo que implica una transición de audio de la primera escena 150A en el momento t_1 sobre una posición transitoria (150AB) en el tiempo t_2 a la siguiente (segunda) escena B (150B) en el tiempo t_3.

En el punto de tiempo t_1 el usuario puede estar en la posición x_1 en la dirección x de una primera escena VR, AR y/o MR. En el punto de tiempo t_3 el usuario puede estar en una segunda escena VR, AR y/o MR (150B) diferente, en la posición x_3. En el instante t_2 el usuario puede estar en una posición de transición 150AB, mientras abre una puerta (por ejemplo, una puerta virtual) y camina hacia ella. Por lo tanto, la transición implica una transición de información de audio de la primera escena 150A a la segunda escena 150B.

En este contexto, el usuario está cambiando su posición 110, por ejemplo, desde el primer entorno VR (caracterizado por un primer punto de vista (A) como se muestra en la figura 1.1) hasta el segundo entorno VR (caracterizado por un segundo punto de vista (B) como se muestra en la figura 1.1). En un caso particular, por ejemplo, durante la transición a través de la puerta ubicada en una posición x_2 en la dirección x, algunos elementos de audio 152A y 152B pueden estar presentes en ambos puntos de vista (posiciones A y B).

El usuario (estando equipado con el MCD) está cambiando su posición 110 (x_1-x_3) hacia la puerta, lo que puede implicar que, en la posición de transición x_2, los elementos de audio pertenecen a la primera escena 150A y la segunda escena 150B. El MCD transmite la nueva posición y los datos de transición 110 al cliente, que lo retransmite al servidor de medios 120. El usuario puede estar habilitado para escuchar las fuentes de audio apropiadas definidas por la posición intermedia x_2 entre la primera y la segunda posición x_1 y x_3.

Cualquier posición y cualquier transición desde la primera posición (x_1) a la segunda posición (x_3) ahora se transmite periódicamente (por ejemplo, de manera continua) desde el MCD al cliente. El cliente 102 puede retransmitir los datos de posición y transición 110 (x_1-x_3) al servidor de medios 120, que está configurado para entregar en consecuencia un elemento dedicado de, por ejemplo, un nuevo conjunto de transmisiones preprocesadas 106 en forma de un conjunto de adaptación actualizado 113', dependiendo de la posición recibida y los datos de transición 110 (x_1-x_3).

El servidor de medios 120 puede seleccionar una de una multitud de representaciones asociadas con la información mencionada anteriormente, no sólo con respecto a la capacidad del MCD para mostrar la tasa de bits más alta, sino también con respecto a la posición y los datos de transición 110 (x_1-x_3) del usuario durante su movimiento de una posición a otra. (En este contexto, es posible utilizar conjuntos de adaptación: el servidor 120 de medios puede decidir qué conjunto de adaptación 113 representa de manera óptima la transición virtual del usuario, sin interferir con la capacidad de representación del MCD).

El servidor de medios 120 puede entregar, por lo tanto, una transmisión dedicada 106 (por ejemplo, como un nuevo conjunto de adaptación 113) según la transición de las posiciones. El cliente 102 puede configurarse para entregar de manera correspondiente una señal de audio 108 al usuario 140, por ejemplo, a través del decodificador de audio de medios 104.

Las transmisiones 106 (generadas sobre la marcha y/o preprocesados) pueden transmitirse en un conjunto de adaptación 113 actualizado periódicamente (por ejemplo, continuamente) al cliente 102.

Cuando el usuario camina por la puerta, el servidor 120 puede transmitir tanto las transmisiones 106 de la primera escena 150A como las transmisiones 106 de la segunda escena 150B. Esto es para mezclar o multiplexación o componer o reproducir simultáneamente estas transmisiones 106, para dar una impresión real al usuario. Por lo tanto, sobre la base de la posición 110 del usuario (por ejemplo, "posición correspondiente a la puerta"), el servidor 120 transmite diferentes transmisiones 106 al cliente.

Incluso en este caso, como las diferentes transmisiones 106 deben escucharse simultáneamente, pueden tener diferentes resoluciones y pueden transmitirse desde el servidor 120 al cliente en diferentes resoluciones. Cuando el usuario haya completado la transición y se encuentre en la segunda escena (posición) 150A (y haya cerrado la puerta detrás de él), habrá una posibilidad para el servidor 120 de reducir o abstenerse de transmitir las transmisiones 106 de la primera escena 150. (en caso de que el servidor 120 ya haya proporcionado al cliente 102 las transmisiones, el cliente 102 puede decidir no usarlos).

Caso 3

La **figura 4** (caso 3) muestra una forma de realización con otro escenario de ejemplo (representado en un plano vertical XZ de un espacio XYZ, donde el eje Y se representa al ingresar al papel), donde el usuario se mueve en un VR, AR y/o la escena MR 150A implica una transición de audio de una primera posición en el tiempo t_1 a una segunda posición también en la primera escena 150A en el tiempo t_2. El usuario en la primera posición puede estar lejos de una pared en el tiempo t_1 a una distancia d_1 de la pared; y puede estar cerca de la pared en el tiempo t_2, a una distancia d_2 de la pared. Aquí, $d_1 > d_2$. Mientras que a la distancia d_1 el usuario sólo escucha la fuente 152A de la escena 150A, también

puede escuchar la fuente 152B de la escena 150B más allá de la pared.

Cuando el usuario está en la segunda posición (d_2), el cliente 102 envía al servidor 120 los datos relativos a la posición del usuario 110 (d_2), y recibe, desde el servidor 120, no solo las transmisiones de audio 106 de la primera escena 150A, pero también las transmisiones de audio 106 de la segunda escena 150B. Sobre la base de los metadatos proporcionados por el servidor 120, por ejemplo, el cliente 102 causará la reproducción, por ejemplo, a través del decodificador 104, de las transmisiones 106 de la segunda escena 150B (más allá de la pared) a un volumen bajo.

Incluso en este caso, la tasa de bits (calidad) de las transmisiones 106 de la segunda escena 150B puede ser baja, por lo que requiere una carga útil de transmisión reducida desde el servidor 120 al cliente. En particular, la posición 110 (d_1, d_2) del cliente (y/o la ventana gráfica) define las transmisiones de audio 106 que proporciona el servidor 120.

Por ejemplo, el sistema 102 puede configurarse para obtener las transmisiones asociadas a una primera escena actual (150A) asociada al primer entorno actual y, en caso de que la distancia de la posición del usuario o la posición virtual desde un límite (por ejemplo, a la pared) de la escena está por debajo de un umbral predeterminado (por ejemplo, cuando $d_2<d_{umbral}$), el sistema 102 obtiene además transmisiones de audio asociadas al segundo entorno adyacente y/o adyacente asociado a la segunda escena (150B).

Caso 4

Las figuras 5a y 5b muestran una forma de realización con otro escenario ejemplar (representado en un plano horizontal XY de un espacio XYZ, donde el eje Z se representa como que sale del papel), donde el usuario está posicionado en uno y el mismo VR, AR y/o MR escena 150 pero en diferentes instantes a diferentes distancias, por ejemplo, dos elementos de audio.

En el primer instante $t=t_1$ que se muestra en la **figura 5a**, un usuario está posicionado, por ejemplo, en una primera posición. En esta primera posición, un primer elemento de audio 1 (152-1) y un segundo elemento de audio 2 (152-2) están ubicados (por ejemplo, virtualmente) a distancias d_1 y d_2 respectivas del usuario equipado con MCD. Ambas distancias d_1 y d_2 pueden ser mayores en este caso que un umbral de distancia d_{umbral}, definido, y por lo tanto el sistema 102 está configurado para agrupar ambos elementos de audio en una sola fuente virtual 152-3. La posición y las propiedades (por ejemplo, la extensión espacial) de la fuente virtual única se pueden calcular basándose, por ejemplo, en las posiciones de las dos fuentes originales de manera que imite lo mejor posible el campo de sonido original generado por las dos fuentes (por ejemplo, dos fuentes puntuales bien localizadas pueden reproducirse en medio de la distancia entre ellas como una sola fuente). Los datos de posición del usuario 110 (d_1, d_2) pueden transmitirse desde el MCD al sistema 102 (cliente) y, posteriormente, al servidor 120, que puede decidir enviar una transmisión de audio apropiado 106 para ser procesado por el sistema de servidor 120 (en otras formas de realización, es el cliente 102 el que decide qué la ventana gráfica actual del usuario y/o la orientación se transmitirán desde el servidor 120). Al agrupar ambos elementos de audio en una sola fuente virtual 152-3, el servidor 120 puede seleccionar una de las múltiples representaciones asociadas con la información mencionada anteriormente. (Por ejemplo, es posible entregar en consecuencia una transmisión dedicada 106 y un conjunto de adaptación 113' en consecuencia asociado con, por ejemplo, un solo canal). Por consiguiente, el usuario puede recibir a través del MCD una señal de audio que se transmite desde el único elemento de audio virtual 152- 3 posicionados entre los elementos de audio reales 1 (152-1) y 2 (152-2).

En un segundo instante, $t=t_2$ se muestra en la **figura 5b**, un usuario está posicionado, por ejemplo, en la misma escena 150, que tiene una segunda posición definida en el mismo entorno VR que en la figura 5a. En esta segunda posición, los dos elementos de audio 152-1 y 152-2 están ubicados (por ejemplo, virtualmente) a distancias d_3 y d_4 respectivas del usuario. Ambas distancias d_3 y d_4 pueden ser más pequeñas que la distancia umbral d_{umbral}, y por lo tanto, la agrupación de los elementos de audio 152-1 y 152-2 en una sola fuente virtual 152-3 ya no se usa. Los datos de posición del usuario se transmiten desde el MCD al sistema 102 y, posteriormente, al servidor 120, que puede decidir enviar otra transmisión de audio apropiado 106 para que sea procesado por el servidor del sistema 120 (en otras formas de realización, esta decisión la toma el cliente 102). Al evitar agrupar los elementos de audio, el servidor 120 puede seleccionar una representación diferente asociada con la información mencionada anteriormente para entregar en consecuencia una transmisión dedicada 106 con un conjunto de adaptación 113 asociado de manera correspondiente con diferentes canales para cada elemento de audio. En consecuencia, el usuario puede recibir a través del MCD una señal de audio 108 que se transmite desde dos elementos de audio diferentes 1 (152-1) y 2 (152-2). Por lo tanto, cuanto más cerca esté la posición 110 del usuario de las fuentes de audio 1 (152-1) y 2 (152-2), más alto será el nivel de calidad necesario della ventana gráfica actual del usuario y/o la orientación asociado a las fuentes de audio que se seleccionará.

De hecho, cuanto más cerca estén las fuentes de audio 1 (152-1) y 2 (152-2) con respecto al usuario, como se muestra en la figura 5b, más alto debe ser el nivel de ajuste y, por lo tanto, las señales de audio 108 pueden traducirse a un nivel de calidad superior. Por el contrario, las fuentes de audio ubicadas a distancia 1 y 2 representadas en la figura 5a deben escucharse a un nivel inferior, como lo reproduce la única fuente virtual, por lo que se representan, por

ejemplo, en un nivel de calidad inferior.

En una configuración similar, una multitud de elementos de audio puede ubicarse frente a un usuario, todos ellos posicionados a distancias mayores que la distancia umbral del usuario. En una realización, dos grupos de cinco elementos de audio pueden combinarse en dos fuentes virtuales. Los datos de posición del usuario se transmiten desde el MCD al sistema 102 y, posteriormente, al servidor 120, que puede decidir enviar una transmisión de audio apropiado 106 para que sea procesado por el servidor del sistema 120. Al agrupar los 10 elementos de audio en sólo dos fuentes virtuales únicas en las que el servidor 120 puede seleccionar una de una multitud de representaciones asociadas con la información mencionada anteriormente para entregar en consecuencia una transmisión dedicada 106 con un conjunto de adaptación 113' asociado de forma correspondiente con, por ejemplo, dos elementos de audio únicos. En consecuencia, el usuario puede recibir a través del MCD una señal de audio que se transmite desde dos elementos de audio virtuales distintos ubicados en la misma área de posicionamiento con los elementos de audio reales.

En un instante posterior de tiempo, un usuario se acerca a la multitud de (diez) elementos de audio. En esta escena subsiguiente, todos los elementos de audio se ubican a distancias que son más pequeñas que la distancia umbral d_{umbral}, y por lo tanto el sistema 102 está configurado para terminar la agrupación de elementos de audio. Los nuevos datos de posición del usuario se transmiten desde el MCD al sistema 102 y, posteriormente, al servidor 120, que puede decidir enviar otra transmisión de audio apropiado 106 a ser procesada por el sistema del servidor 120. Al no agrupar los elementos de audio, el servidor 120 puede seleccionar una representación diferente asociada con la información antes mencionada para entregar en consecuencia una transmisión dedicada 106 con un conjunto de adaptación 113' asociado en consecuencia con diferentes canales para cada elemento de audio. En consecuencia, el usuario puede recibir a través del MCD una señal de audio que se transmite desde diez elementos de audio diferentes. Por lo tanto, cuanto más cerca esté la posición 110 del usuario de las fuentes de audio, mayor será la resolución necesaria de la transmisión asociada a las fuentes de audio que se seleccionará.

Caso 5

La **figura 6** (caso 5) representa a un usuario 140 ubicado en una posición de una sola escena 150 usando un dispositivo de consumo de medios (MCD) que puede dirigirse en tres direcciones diferentes a modo de ejemplo (cada una asociada a una vista diferente 160-1, 160- 2, 160-3). Estas direcciones, como se muestra en la figura 6, pueden tener una orientación (por ejemplo, orientación angular) en un sistema de coordenadas polares y/o en el sistema cartesiano XY que apunta a un primer punto de vista 801 ubicado, por ejemplo, a 180° en la parte inferior de la figura 6, en un segundo punto de vista 802 ubicado, por ejemplo, 90° en el lado derecho de la figura 6 y en un tercer punto de vista 803 ubicado, por ejemplo, a 0° en la parte superior de la figura 6. Cada uno de estos puntos de vista está asociado a la orientación del usuario 140 que lleva el dispositivo de consumidor de medios (MCD), y al usuario que se encuentra en el centro se le ofrece una vista específica que muestra el renderizado del MCD La correspondiente señal de audio 108 según la orientación del MCD.

En este entorno de realidad virtual particular, un primer elemento de audio s1 (152) está ubicado en la primera ventana gráfica 160-1, en la vecindad del punto de vista ubicado, por ejemplo, a las 180° y un segundo elemento de audio s2 (152) está ubicado en la tercera ventana 160-3, la vecindad del punto de vista ubicado, por ejemplo, en 180°. Antes de cambiar su orientación, el usuario 140 experimenta en la primera orientación hacia el punto de vista 801 (ventana gráfica 160-1) un sonido asociado con su posición real (efectiva) siendo más alta desde el elemento de audio s1 que el elemento de audio s2.

Al cambiar su orientación, el usuario 140 puede experimentar en la segunda orientación hacia el punto de vista 802, un sonido asociado con su posición real 110 es casi la misma sonoridad que proviene de ambos elementos de audio s1 y s2.

Finalmente, al cambiar su orientación, el usuario 140 puede experimentar en la tercera orientación hacia el punto de vista 801 (vista 160-3) que un sonido asociado con el elemento de audio 2 sea más alto que el sonido asociado al elemento de audio s1 (de hecho, el sonido el elemento de audio 2 llega desde la parte frontal, mientras que el sonido del elemento de audio 1 llega desde la parte posterior).

Por lo tanto, diferentes ventanas gráficas y/u orientaciones y/o datos de posición virtual pueden asociarse a diferentes tasas de bits y/o calidades.

Otros casos y ejemplos

La **figura 7a** muestra una forma de realización del método para recibir transmisiones de audio de un sistema en forma de una secuencia de etapas de operación en un diagrama. En cualquier momento, un usuario del sistema 102 está asociado con su ventana gráfica actual y/o con la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o la posición virtual. En cierto momento, el sistema puede determinar en la etapa 701 de

la figura 7A los elementos de audio que se reproducirán en la base de la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o la posición virtual. Por lo tanto, en una próxima etapa 703 se puede determinar la relevancia y el nivel de audibilidad para cada elemento de audio. Como se describe anteriormente en la figura 6, un entorno VR puede tener diferentes elementos de audio ubicados en una escena particular 150 ya sea en la vecindad del usuario o más lejos, pero también tiene una orientación específica en los 360 grados que lo rodean. Todos estos factores determinan la relevancia y el nivel de audibilidad de cada uno de dichos elementos de audio.

En una siguiente etapa 705, el sistema 102 puede solicitar las transmisiones de audio de acuerdo con la relevancia y el nivel de audibilidad determinados para cada uno de los elementos de audio del servidor 120 de medios.

En una siguiente etapa 707, el sistema 102 puede recibir las transmisiones de audio 113 preparadas en consecuencia por el servidor de medios 120, en donde las secuencias con diferentes tasas de bits pueden reflejar la relevancia y el nivel de audibilidad según lo determinado en las etapas anteriores.

En una siguiente etapa 709, el sistema 102 (por ejemplo, el decodificador de audio) puede decodificar las transmisiones de audio 113 recibidos, de modo que en la etapa 711 se reproduzca la escena particular 150 (por ejemplo, mediante el MCD), de acuerdo con la ventana gráfica actual y/u orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o posición virtual.

La**figura 7b** representa una interacción entre un servidor de medios 120 y un sistema 102 de acuerdo con la secuencia de operación descrita anteriormente. En un momento determinado, el servidor de medios puede transmitir una transmisión de audio 750 a una tasa de bits inferior, de acuerdo con lo anterior, se determinó una menor relevancia y un nivel de audibilidad de los elementos de audio relevantes de una escena 150 anterior. El sistema puede determinar en un momento posterior 752 que una interacción o un cambio se produzca en los datos de posición. Tal interacción puede resultar, por ejemplo, de cualquiera de los cambios en los datos de posición en la misma escena 150 o, por ejemplo, activando una manija de la puerta mientras el usuario intenta ingresar a una segunda escena separada de la primera escena por una puerta provista por la manija de la puerta.

Un cambio en la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o la posición virtual pueden resultar en una solicitud 754 enviada por el sistema 102 al servidor de medios 120. Esta solicitud puede reflejar una mayor relevancia y el nivel de audibilidad de los elementos de audio relevantes determinados para esa escena 150 posterior. Como respuesta a la solicitud 754, el servidor de medios puede transmitir una transmisión 756 a una tasa de bits más alta, permitiendo una reproducción plausible y realista de la escena 150 por el sistema 102 en cualquier posición virtual del usuario actual.

La**figura 8a** muestra otra forma de realización del método para recibir transmisiones de audio de un sistema también en forma de una secuencia de etapas de operación en un diagrama. En un determinado momento 801 se puede realizar una determinación de una primera ventana gráfica y/u orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o posición virtual. Deduciendo un caso afirmativo, una petición de transmisiones asociadas a la primera posición definida por una tasa de bits baja puede ser preparada y transmitida por el sistema 102 en la etapa 803.

Una etapa determinante 805 que tiene tres resultados diferentes se puede realizar en un momento posterior. Uno o dos umbral(es) definido(s) puede(n) ser relevante(s) en esta etapa para determinar, por ejemplo, una decisión predictiva con respecto a una vista posterior y/u orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o posición virtual. Por lo tanto, se puede realizar una comparación con un primer y/o segundo umbral, con respecto a la probabilidad de un cambio a una segunda posición, lo que resulta en, por ejemplo, tres diferentes etapas posteriores por realizar.

En un resultado que refleje, por ejemplo, una probabilidad muy baja (por ejemplo, asociada a la comparación anterior con un primer umbral predeterminado), se realizará una nueva etapa de comparación 801.

En un resultado que refleja una baja probabilidad (por ejemplo, más alto que el primer umbral predeterminado pero, en los ejemplos, más bajo que un segundo umbral predeterminado más alto que el primer umbral) puede resultar en una solicitud en la etapa 809 para transmisiones de audio 113 a una tasa de bits baja.

En un resultado que refleja una alta probabilidad (por ejemplo, más alto que el segundo umbral predeterminado), se puede realizar una solicitud, en la etapa 807, para transmisiones de audio 113 a una alta tasa de bits. Por lo tanto, una etapa posterior que se realizará después de ejecutar las etapas 807 u 809 podría ser nuevamente la etapa determinante 801.

La **figura 8B** representa una interacción entre un servidor de medios 120 y un sistema 102 de acuerdo con sólo uno de los diagramas de operación de secuencia descritos anteriormente. En un momento determinado, el servidor de

medios puede transmitir una transmisión de audio 850 a una tasa de bits baja, según un nivel de relevancia y audibilidad determinado anterior de los elementos de audio de una escena 150 anterior. El sistema puede determinar en un momento posterior 852 que una interacción predeciblemente va a ocurrir. Un cambio predictivo de la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o la posición virtual pueden resultar en una solicitud apropiada 854 enviada por el sistema 102 al servidor de medios 120. Esta solicitud puede reflejar uno de los casos descritos anteriormente con respecto a una alta probabilidad de alcanzar una segunda posición asociada con una tasa de bits alta de acuerdo con el nivel de audibilidad de los elementos de audio según se requiere para la escena 150 respectiva. Como respuesta, el servidor de medios puede transmitir una transmisión 856 a una mayor tasa de bits, permitiendo una reproducción plausible y realista de la escena 150 por el sistema 102 en la posición virtual de cualquier usuario actual.

El sistema 102 como se muestra en la **figura 1.3** está configurado para recibir transmisiones de audio 113 en base a otra configuración en el lado del cliente, en donde la arquitectura del sistema puede usar puntos de vista discretos basados en una solución que usa múltiples decodificadores de audio 1320, 1322. En el lado del cliente, el sistema 102 puede incorporar, por ejemplo, partes del sistema descritas en la figura 1.2 que adicional o alternativamente comprende múltiples decodificadores de audio 1320, 1322 que pueden configurarse para decodificar las transmisiones de audio individuales como lo indica el procesador de metadatos 1236, por ejemplo con una serie de elementos de audio desactivados.

Puede proporcionarse un mezclador/renderizador 1238 en el sistema 102 que está configurado para reproducir la escena de audio final en base a la información sobre la ubicación y/u orientación del usuario y/o la dirección de movimiento, es decir, por ejemplo, algunos de los elementos de audio que no son audibles en esa ubicación específica deben ser deshabilitados o no representados.

Las siguientes realizaciones mostradas en las **figuras 1.4, 1.5 y** 1.6 se basan en conjuntos de adaptación independientes para puntos de vista discretos con conjuntos de adaptación flexibles. En el caso de que el usuario se mueva en un entorno VR, la escena de audio puede cambiar de manera continua. Para garantizar una buena experiencia de audio, todos los elementos de audio que componen una escena de audio en un momento determinado pueden tener que estar disponibles para un decodificador de medios que puede utilizar la información de posición para crear la escena de audio final.

Si el contenido está precodificado, para una serie de ubicaciones predefinidas, el sistema puede proporcionar una reproducción precisa de las escenas de audio en estas ubicaciones específicas bajo el supuesto de que estas escenas de audio no se superponen y el usuario puede "saltar/cambiar" de un lugar a otro.

Pero en los casos en que el usuario "camina" de una ubicación a la siguiente, los elementos de audio de dos (o más) escenas de audio pueden ser audibles al mismo tiempo. Se proporcionó una solución para estos casos de uso en los ejemplos de sistemas anteriores, donde independientemente de los mecanismos provistos para decodificar múltiples transmisiones de audio (ya sea utilizando un mezclador con un solo decodificador de medios o con varios mezcladores/renderizadores adicionales), las transmisiones de audio que describan escenas de audio completas deben proporcionarse al cliente.

A continuación se proporciona una optimización al introducir la noción de elementos de audio comunes entre múltiples transmisiones de audio.

Análisis de aspectos y ejemplos

Solución 1: Conjuntos de adaptación independientes para ubicaciones discretas (puntos de vista).

Una forma de resolver el problema descrito es usar conjuntos de adaptación independientes completos para cada ubicación. Para una mejor comprensión de la solución, la figura 1.1 se utiliza como un ejemplo de escenario. En este ejemplo, se utilizan tres puntos de vista distintos (que comprenden tres escenas de audio diferentes) para crear un entorno VR completo, en el que el usuario debería poder moverse. Por lo tanto:

- Las varias escenas de audio independientes o superpuestas se codifican en una serie de transmisiones de audio. Para cada escena de audio, se puede usar una transmisión principal dependiendo del caso de uso, una transmisión principal y las transmisiones auxiliares adicionales (por ejemplo, algunos objetos de audio que contienen diferentes idiomas pueden codificarse en transmisiones independientes para una entrega eficiente). En el ejemplo proporcionado, la escena de audio A se codifica en dos secuencias (A1 y A2), la escena de audio B se codifica en tres secuencias (B1, B2 y B3) mientras que la escena de audio C se codifica en tres secuencias (C1, C2 y C3). Debe notarse que la escena de audio A y la escena de audio B comparten una serie de elementos comunes (en este ejemplo, dos objetos de audio). Dado que cada escena debe ser completa e independiente (para la reproducción independiente, por ejemplo, en dispositivos de reproducción sin VR), los elementos comunes deben codificarse dos veces para cada escena.

• Todas las transmisiones de audio están codificadas a diferentes tasas de bits (es decir, diferentes representaciones), que permiten una adaptación eficiente de las tasas de bits dependiendo de la conexión de la red (es decir, para los usuarios que usan una conexión de alta velocidad, la versión codificada de alta tasa de bits se entrega mientras que para los usuarios con una conexión de red de menor velocidad se entrega una versión de tasa de bits más baja).

• Las transmisiones de audio se almacenan en un servidor de medios, donde, para cada transmisión de audio, las diferentes codificaciones a diferentes tasas de bits (es decir, diferentes representaciones) se agrupan en un conjunto de adaptación con los datos adecuados que indican la disponibilidad de todos los conjuntos de adaptación creados.

• Además, en los conjuntos de adaptación, el servidor de medios recibe información sobre los "límites" de ubicación de cada escena de audio y su relación con cada conjunto de adaptación (que puede contener, por ejemplo, una escena de audio completa o sólo objetos individuales). De esta manera, cada conjunto de adaptación puede estar asociado con una de las escenas de audio disponibles. Los límites de una escena de audio pueden definirse, por ejemplo, como coordenadas geométricas de una esfera (por ejemplo, centro y radio).

◦ Cada conjunto de adaptación también contiene información descriptiva sobre las ubicaciones en las que la escena de sonido o los elementos de audio están activos. Por ejemplo, si una transmisión auxiliar contiene uno o varios objetos, el conjunto de adaptación podría contener información como las ubicaciones donde los objetos son audibles (por ejemplo, las coordenadas del centro de una esfera y el radio).

• El servidor de medios proporciona información sobre los "límites" de ubicación asociados con cada conjunto de adaptación al cliente, por ejemplo, un cliente DASH. Por ejemplo, esto se puede incrustar en la sintaxis XML de Descripción de Presentación de Medios (MPD) en el caso de un entorno de entrega DASH.

• El cliente recibe información sobre la ubicación y/o la orientación del usuario y/o la dirección del movimiento (o cualquier información que caracterice los cambios provocados por las acciones del usuario)

• El cliente recibe la información sobre cada conjunto de adaptación, y en base a esto y la ubicación y/o la orientación y/o dirección del movimiento del usuario (o cualquier información que caracterice los cambios provocados por las acciones del usuario, por ejemplo, que comprenden las coordenadas x, y, z y/o valores de desvío, inclinación, balanceo), el cliente selecciona uno o más conjuntos de adaptación que describen completamente una escena de audio que debe reproducirse en la ubicación actual del usuario.

• El cliente solicita uno o más conjuntos de adaptación.

◦ Además, el cliente puede seleccionar más conjuntos de adaptación que describan más de una escena de audio, y usar las transferencias de audio correspondientes a más de una escena de audio para crear una nueva escena de audio que debe reproducirse en la ubicación actual del usuario. Por ejemplo, si el usuario camina en el entorno VR, y en un momento en el tiempo se encuentra entre (o en una ubicación situada en un lugar donde dos escenas de audio tienen efectos audibles).

◦ Una vez que las transmisiones de audio estén disponibles, se pueden usar múltiples decodificadores de medios para decodificar las transmisiones de audio individuales y un mezclador/renderizador 1238 adicional para reproducir la escena de audio final en función de la información sobre la ubicación y/u orientación del usuario y/o la dirección de movimiento (es decir, por ejemplo, algunos de los elementos de audio que no son audibles en esa ubicación específica deben desactivarse o no representarse)

◦ Alternativamente, se puede usar un procesador de metadatos 1236 para manipular los metadatos de audio asociados con todas las transmisiones de audio, en función de la información sobre la ubicación del usuario y/o la orientación y/o la dirección del movimiento, para:

Seleccionar/habilitar los elementos de audio 152 necesarios para componer la nueva escena de audio;

▪ Y para permitir la fusión de todas las transmisiones de audio en una sola transmisión de audio.

• El servidor de medios entrega los conjuntos de adaptación requeridos

• De forma alternativa, el cliente proporciona la información sobre el posicionamiento del usuario al servidor de medios y el servidor de medios proporciona una indicación sobre los conjuntos de adaptación requeridos.

La figura 1.2 muestra otro ejemplo de implementación de dicho sistema que comprende:

• en el lado de la codificación

◦ una pluralidad de codificadores de medios que pueden usarse para crear una o más transmisiones de audio para cada escena de audio disponible asociada con una parte de escena de sonido de un punto de vista

◦ una pluralidad de codificadores de medios que pueden usarse para crear una o más transmisiones de vídeo para cada escena de vídeo disponible asociada con una parte de escena de vídeo de un punto de vista. Los codificadores de vídeo no están representados en la figura por simplicidad

◦ un servidor de medios que almacena múltiples conjuntos de adaptación de audio y vídeo que comprenden diferentes codificaciones de las mismas transmisiones de audio y vídeo a diferentes tasas de bits (es decir, diferentes representaciones). Además, el servidor de medios contiene información descriptiva de todos los conjuntos de adaptación, que pueden incluir

▪ disponibilidad de todos los conjuntos de adaptación creados;

▪ información que describe una asociación de un conjunto de adaptación a una escena de audio y/o punto de vista; de esta manera, cada conjunto de adaptación puede estar asociado con una de las escenas de audio disponibles;

▪ información que describe los "límites" de cada escena de audio y/o punto de vista (que puede contener, por ejemplo, una escena de audio completa o sólo objetos individuales). Los límites de una escena de audio pueden definirse, por ejemplo, como coordenadas geométricas de una esfera (por ejemplo, centro y radio).

• en el lado del cliente un sistema (sistema de cliente) que puede comprender en cualquiera de:

◦ un extremo receptor, que puede recibir:

▪ información sobre la ubicación y/o la orientación del usuario y/o la dirección del movimiento (o cualquier información que caracterice los cambios provocados por las acciones del usuario)

▪ información sobre la disponibilidad de todos los conjuntos de adaptación e información que describe una asociación de un conjunto de adaptación a una escena de audio y/o punto de vista; y/o información que describa los "límites" de cada escena de audio y/o punto de vista (que puede contener, por ejemplo, una escena de audio completa o sólo objetos individuales). Por ejemplo, dicha información puede proporcionarse como parte de la sintaxis XML de descripción de presentación de medios (MPD) en el caso de un entorno de entrega DASH.

◦ un lado del dispositivo de consumo de medios utilizado para el consumo de contenido (por ejemplo, basado en un HMD). El dispositivo de consumo de medios también es responsable de recopilar información sobre la ubicación y/o la orientación del usuario y/o la dirección del movimiento (o cualquier información que caracterice los cambios provocados por las acciones del usuario)

◦ un procesador de ventana gráfica 1232, que se puede configurar para

▪ recibir información sobre la ventana gráfica actual que puede contener la ubicación y/u orientación del usuario y/o la dirección del movimiento (o cualquier información que caracterice los cambios provocados por las acciones del usuario) desde el lado del dispositivo de consumo de medios.

▪ recibir información sobre y la ROI señalada en los metadatos (ventanas gráficas de vídeo señaladas como en la especificación OMAF).

▪ recibir toda la información disponible en el extremo receptor;

▪ decidir en función de toda la información recibida y/o derivada de los metadatos recibidos y/o disponibles, qué punto de vista de audio/ vídeo debe reproducirse en un momento determinado. Por ejemplo, el procesador de ventanas gráficas 1232 puede decidir que:

• se reproduce una escena de audio completa

• se debe crear una nueva escena de audio a partir de todas las escenas de audio disponibles (por ejemplo, solo se reproducirán algunos elementos de audio de múltiples escenas de audio, mientras que otros elementos de audio restantes de estas escenas de audio no se reproducirán)

• se debe reproducir una transición entre dos o más escenas de audio

◦una parte de selección 1230 configurada para seleccionar, en función de la información recibida del procesador de

ventanas gráficas 1232 uno o más, los conjuntos de adaptación fuera de los conjuntos de adaptación disponibles como se indica en la información recibida por el extremo receptor; los ajustes de adaptación seleccionados describen completamente la escena de audio que debe reproducirse en la ubicación actual del usuario. Esta escena de audio puede ser una escena de audio completa como se define en el lado de la codificación o debe crearse una nueva escena de audio a partir de todas las escenas de audio disponibles.

▪ Además, en el caso de que se produzca una transición entre dos o más escenas de audio en función de la indicación del procesador de ventanas gráficas 1232, la parte de selección 1230 puede estar configurada para seleccionar uno o más conjuntos de adaptación de entre los conjuntos de adaptación disponibles, tal como se indica en la información recibida por el extremo receptor; los conjuntos de adaptación seleccionados describen completamente la escena de audio que puede ser necesario reproducir en un futuro próximo (por ejemplo, si el usuario camina en la dirección de una escena de audio siguiente con una cierta velocidad, se puede predecir que se requiere la escena de audio siguiente y se selecciona antes de la reproducción).

▪ **Además, algunos conjuntos de adaptación correspondientes a las ubicaciones vecinas pueden seleccionarse primero a una tasa de bits inferior (es decir, una representación codificada a una tasa de bits más baja se elige entre las representaciones disponibles en un conjunto de adaptación),**y en función de los cambios de posición, la calidad aumenta seleccionando una tasa de bits más alta para esos conjuntos de adaptación específicos **(es decir, una representación codificada a una tasa de bits más alta se selecciona de las representaciones disponibles en un conjunto de adaptación).**

◦ una parte de descarga y conmutación que se puede configurar para:

▪ solicitud, basada en la indicación recibida de la parte de selección 1230, uno o más, conjuntos de adaptación fuera de los conjuntos de adaptación disponibles del servidor de medios 120;

▪ recibir, uno o más, conjuntos de adaptación (es decir, una representación de todas las representaciones disponibles dentro de cada conjunto de adaptación) fuera de los conjuntos de adaptación disponibles del servidor de medios 120;

▪ extraer información de metadatos de todas las transmisiones de audio recibidos

◦ un procesador de metadatos 1236 que puede configurarse para:

▪ recibir de la información de descarga y cambio acerca de las transmisiones de audio recibidos, información que puede incluir los metadatos de audio correspondientes a cada transmisión de audio recibido

▪ procesar y manipular los metadatos de audio asociados con cada transmisión de audio, en función de la información recibida del procesador de ventanas gráficas 1232 que puede incluir información sobre la ubicación y/u orientación del usuario y/o la dirección del movimiento, con el fin de:

seleccionar/habilitar los elementos de audio requeridos 152 que componen la nueva escena de audio como lo indica el procesador de ventanas gráficas 1232;

▪ permitir la combinación de todas las transmisiones de audio en una sola transmisión de audio.

◦ un multiplexador de transmisiones/fusionador 1238 que puede configurarse para fusionar todas las transmisiones de audio seleccionadas en una transmisión de audio en función de la información recibida del procesador de metadatos 1236, que puede incluir los metadatos de audio modificados y procesados correspondientes a todas las transmisiones de audio recibidos

◦ un decodificador de medios configurado para recibir y decodificar al menos una transmisión de audio para la reproducción de la nueva escena de audio, según lo indicado por el procesador de ventanas gráficas 1232, según la información sobre la ubicación y/o la orientación del usuario y/o la dirección del movimiento

La figura 1.3 muestra un sistema que comprende, en el lado del cliente, un sistema (sistema de cliente) que puede incorporar, por ejemplo, partes del sistema descrito en la figura 1.2 que adicional o alternativamente comprende:

• múltiples decodificadores de medios que pueden configurarse para decodificar las transmisiones de audio individuales según lo indicado por el procesador de metadatos 1236 (por ejemplo, con una serie de elementos de audio desactivados).

• un mezclador/renderizador 1238 que puede configurarse reproduce la escena de audio final en base a la información sobre la ubicación y/o la orientación del usuario y/o la dirección del movimiento (es decir, algunos de los elementos de audio que no son audibles en esa ubicación específica deben estar deshabilitados o no representados)

Solución 2

Las figuras 1.4, 1.5 y 1.6 se refieren a ejemplos de acuerdo con una solución 2 de la invención (que pueden ser formas de realización de los ejemplos de las figuras 1.1 y/o 1.2 y/o 1.3): conjuntos de adaptación independientes para ubicaciones discretas (puntos de vista) con conjuntos de adaptación flexibles.

En el caso de que el usuario se mueva en un entorno VR, la escena de audio 150 puede cambiar de forma continua. Para garantizar una buena experiencia de audio, todos los elementos de audio 152 que componen una escena de audio 150 en un determinado momento pueden tener que estar disponibles para un decodificador de medios que puede utilizar la información de posición para crear la escena de audio final.

Si el contenido está precodificado, para una serie de ubicaciones predefinidas, el sistema puede proporcionar una reproducción precisa de las escenas de audio en estas ubicaciones específicas bajo el supuesto de que estas escenas de audio no se superponen y el usuario puede "saltar/cambiar" de un lugar a otro.

Pero en los casos en que el usuario "camina" de una ubicación a la siguiente, los elementos de audio 152 de dos (o más) escenas de audio 150 pueden ser audibles al mismo tiempo. Se proporcionó una solución para estos casos de uso en los ejemplos de sistemas anteriores, donde, independientemente de los mecanismos provistos para decodificar múltiples transmisiones de audio (ya sea utilizando un mezclador con un solo decodificador de medios o un decodificador de medios múltiples con un mezclador/renderizador adicional 1238), las transmisiones de audio que describen escenas de audio completas 150 deben proporcionarse al cliente/sistema 102.

A continuación, se proporciona una optimización al introducir la noción de elementos de audio comunes 152 entre múltiples transmisiones de audio.

La figura 1.4 muestra un ejemplo en el que diferentes escenas comparten al menos un elemento de audio (objeto de audio, fuente de sonido...). Por lo tanto, el cliente 102 puede recibir, por ejemplo, una transmisión principal 106A asociada sólo a una escena A (por ejemplo, asociado al entorno donde se encuentra el usuario actualmente) y asociada a los objetos 152A, y una transmisión auxiliar 106B compartida por una escena diferente B (por ejemplo, una transmisión en el límite entre la escena A en la que el usuario se encuentra actualmente y una transmisión adyacente o adyacente B que comparte los objetos 152B) y está asociado a los objetos 152B.

Por lo tanto, como se muestra en la figura 1.4:

- Las varias escenas de audio independientes o superpuestas se codifican en una serie de transmisiones de audio. Las transmisiones de audio 106 se crean de tal manera que:

◦ para cada escena de audio 150, se puede crear una transmisión principal que contiene sólo los elementos de audio 152 que forman parte de la escena de audio respectiva, pero que no forman parte de ninguna otra escena de audio; y/o

◦ para todas las escenas de audio 150 que comparten elementos de audio 152, los elementos de audio comunes 152 pueden codificarse solo en transmisiones de audio auxiliares asociadas solo con una de las escenas de audio y se crea la información de metadatos apropiada que indica la asociación con otras escenas de audio. O, dicho de otro modo, los metadatos adicionales indican la posibilidad de que algunas transmisiones de audio puedan utilizarse junto con varias escenas de audio; y/o

◦ dependiendo del caso de uso, se pueden crear transmisiones auxiliares adicionales (por ejemplo, algunos objetos de audio que contienen diferentes idiomas pueden codificarse en la ventana gráfica actual del usuario y/o la orientación s independientes para una entrega eficiente).

◦ En la realización provista:

▪ La escena de audio A está codificada en:

• una transmisión de audio principal (A1, 106A),

• una transmisión de audio auxiliar (A2, 106B)

• información de metadatos que puede indicar que algunos elementos de audio 152B de la escena de audio A no están codificados en estas transmisiones de audio A, sino en una secuencia auxiliar A2 (106B) que pertenece a diferentes escenas de audio (escena de audio B)

La escena de audio B está codificada en:

- una transmisión de audio principal (B1, 106C),
- una transmisión de audio auxiliar (B2),
- una transmisión de audio auxiliar (B3),
- información de metadatos que puede indicar que los elementos de audio 152B de la transmisión de audio B2 son elementos de audio comunes 152B que también pertenecen a la escena de audio A.

▪ La escena de audio C se codifica en tres secuencias (C1, C2 y C3).

• Las transmisiones de audio 106 (106A, 106B, 106C...) pueden codificarse a diferentes tasas de bits (es decir, diferentes representaciones), que permiten una adaptación eficiente de las tasas de bits, por ejemplo, dependiendo de la conexión de la red (es decir, para usuarios que utilizan una conexión de alta velocidad la versión codificada de tasa de bits se entrega, mientras que para los usuarios con conexión de red de menor velocidad se entrega una versión de tasa de bits más baja).

• Las transmisiones de audio 106 se almacenan en un servidor de medios 120, donde, para cada secuencia de audio, las diferentes codificaciones a diferentes velocidades de bits (es decir, diferentes representaciones) se agrupan en un conjunto de adaptación con los datos adecuados que indican la disponibilidad de todos los conjuntos de adaptación creados. (Múltiples representaciones de secuencias asociadas a las mismas señales de audio, pero a diferentes tasas de bits y/o calidades y/o resoluciones pueden estar presentes en el mismo conjunto de adaptación).

• Además, en los conjuntos de adaptación, el servidor de medios 120 puede recibir información sobre los "límites" de ubicación de cada escena de audio y su relación con cada conjunto de adaptación (que puede contener, por ejemplo, una escena de audio completa o sólo objetos individuales). De esta manera, cada conjunto de adaptación puede estar asociado con una o más de las escenas de audio 150 disponibles. Los límites de una escena de audio pueden definirse, por ejemplo, como coordenadas geométricas de una esfera (por ejemplo, centro y radio).

◦ Cada conjunto de adaptación puede contener también información descriptiva sobre las ubicaciones en las que la escena de sonido o los elementos de audio 152 están activos. Por ejemplo, si una transmisión auxiliar (por ejemplo, A2, 106B) contiene uno o varios objetos, el conjunto de adaptación podría contener información como las ubicaciones donde los objetos son audibles (por ejemplo, coordenadas del centro de una esfera y radio).

• Adicional o alternativamente, cada conjunto de adaptación (por ejemplo, el conjunto de adaptación asociado a la escena B) puede contener información descriptiva (por ejemplo, metadatos) que puede indicar que los elementos de audio (por ejemplo, 152B) de una escena de audio (por ejemplo, B) están (también o además) codificados en transmisiones de audio (por ejemplo, 106B) que pertenecen a una escena de audio diferente (por ejemplo, A).

• el servidor de medios 120 puede proporcionar información sobre los "límites" de ubicación asociados con cada conjunto de adaptación al sistema 102 (cliente), por ejemplo, un cliente DASH. Por ejemplo, esto se puede incrustar en la sintaxis XML de descripción de presentación de medios (MPD) en el caso de un entorno de entrega DASH.

• el sistema 102 (cliente) puede recibir información sobre la ubicación y/o la orientación del usuario y/o la dirección del movimiento (o cualquier información que caracterice los cambios provocados por las acciones del usuario).

• el sistema 102 (cliente) puede recibir información sobre cada conjunto de adaptación, y en función de esto y/o la ubicación y/u orientación del usuario y/o la dirección del movimiento (o cualquier información que caracterice los cambios activados por las acciones del usuario, por ejemplo, que comprende coordenadas x, y, z y/o valores de desvío, inclinación, balanceo), el sistema 102 (cliente) puede seleccionar uno o más Conjuntos de Adaptación que describen completa o parcialmente una escena de audio 150 que debe reproducirse en la ubicación actual del usuario 140.

• El sistema 102 (cliente) puede solicitar uno o más conjuntos de adaptación:

◦ Además, el sistema 102 (cliente) puede seleccionar uno o más conjuntos de adaptación que describan completa o parcialmente más de una escena de audio 150, y usar las transmisiones de audio 106 correspondientes a más de una escena de audio 150 para crear una nueva escena de audio 150 para reproducir en la ubicación actual del usuario 140.

◦ Sobre la base de los metadatos que indican que los elementos de audio 152 son parte de múltiples escenas de audio 150, los elementos de audio comunes 152 pueden solicitarse sólo una vez para crear la nueva escena de audio, en lugar de solicitarlos dos veces, una por cada escena de audio completa.

◦ Una vez que las transmisiones de audio están disponibles para el sistema cliente 102, por ejemplo, uno o varios decodificador(es) de medios (104) pueden utilizarse para decodificar las transmisiones de audio individuales y/o un mezclador/renderizador adicional para reproducir la escena de audio final basándose en la información sobre la ubicación y/o orientación y/o dirección de movimiento del usuario (es decir, por ejemplo, algunos de los elementos de audio que no son audibles en esa ubicación específica deberían desactivarse o no renderizarse).

◦ De forma alternativa o adicional, se puede utilizar un procesador de metadatos para manipular los metadatos de audio asociados a todas las transmisiones de audio, basándose en la información sobre la ubicación y/o orientación y/o dirección de movimiento del usuario, con el fin de:

▪ Seleccionar/habilitar los elementos de audio 152 necesarios (152A–152c) que componen la nueva escena de audio; y/o

▪ Y permitir la fusión de todas las transmisiones de audio en una sola transmisión de audio.

• El servidor de medios 120 puede entregar los conjuntos de adaptación requeridos

• Alternativamente, el sistema 102 (cliente) proporciona la información sobre el posicionamiento 140 del usuario al servidor de medios 120 y el servidor de medios proporciona • una indicación sobre los conjuntos de adaptación requeridos.

La figura 1.5 muestra otra implementación de ejemplo de dicho sistema que comprende:

• en el lado de la codificación

◦ una pluralidad de codificadores de medios 154 que se pueden usar para crear uno o más transmisiones de audio 106 que incorporan elementos de audio 152 de uno o más de la escena de audio 150 disponible asociada con una parte de escena de sonido de un punto de vista.

▪ para cada escena de audio 150, se puede crear una transmisión principal conteniendo solo los Elementos de audio 152 que forman parte de la escena de audio 150 respectiva, pero no parte de ninguna otra escena de audio

▪ Pueden crearse transmisiones auxiliares adicionales para la misma escena de audio (por ejemplo, algunos objetos de audio que contengan diferentes idiomas pueden codificarse en transmisiones independientes para una entrega eficiente).

▪ se pueden crear transmisiones auxiliares adicionales que contienen:

• elementos de audio 152 comunes a más de una escena de audio 150

•información de metadatos que indica la asociación de estela ventana gráfica actual del usuario y/o la orientación auxiliar con todas las demás escenas de audio 150 que comparten los elementos de audio comunes 152. O dicho de otra manera, los metadatos indican la posibilidad de que algunas transmisiones de audio se puedan usar junto con múltiples escenas de audio.

◦ una pluralidad de codificadores de medios que pueden usarse para crear una o más transmisiones de vídeo para cada escena de vídeo disponible asociada con una parte de escena de vídeo de un punto de vista. Los codificadores de vídeo no están representados en la figura por simplicidad

◦ un servidor de medios 120 que almacena múltiples conjuntos de adaptación de audio y vídeo que comprenden diferentes codificaciones de las mismas transmisiones de audio y vídeo a diferentes tasas de bits (es decir, diferentes representaciones). Además, el servidor de medios 120 contiene información descriptiva de todos los conjuntos de adaptación, que pueden incluir

▪ disponibilidad de todos los conjuntos de adaptación creados;

▪ información que describe una asociación de un conjunto de adaptación a una escena de audio y/o punto de vista; de esta manera, cada conjunto de adaptación puede estar asociado con una de las escenas de audio disponibles;

▪ información que describe los "límites" de cada escena de audio y/o punto de vista (que puede contener, por ejemplo, una escena de audio completa o sólo objetos individuales). Los límites de una escena de audio pueden definirse, por ejemplo, como coordenadas geométricas de una esfera (por ejemplo, centro y radio).

▪ información que indica la asociación de un conjunto de adaptación con más de una escena de audio, que comparte al menos un elemento de audio común.

• en el lado del cliente un sistema (sistema de cliente) que puede comprender en cualquiera de:

◦ un extremo receptor, que puede recibir:

▪ información sobre la ubicación y/o la orientación del usuario y/o la dirección del movimiento (o cualquier información que caracterice los cambios provocados por las acciones del usuario)

▪ información sobre la disponibilidad de todos los conjuntos de adaptación e información que describe una asociación de un conjunto de adaptación a una escena de audio y/o punto de vista; y/o información que describa los "límites" de cada escena de audio y/o punto de vista (que puede contener, por ejemplo, una escena de audio completa o sólo objetos individuales). Por ejemplo, dicha información puede proporcionarse como parte de la sintaxis XML de descripción de presentación de medios (MPD) en el caso de un entorno de entrega DASH.

▪ información que indica la asociación de un conjunto de adaptación con más de una escena de audio, que comparte al menos un elemento de audio común.

◦ un lado del dispositivo de consumo de medios utilizado para el consumo de contenido (por ejemplo, basado en un HMD). El dispositivo de consumo de medios también es responsable de recopilar información sobre la ubicación y/o la orientación del usuario y/o la dirección del movimiento (o cualquier información que caracterice los cambios provocados por las acciones del usuario)

◦ un procesador de ventana gráfica 1232, que se puede configurar para

▪ recibir información sobre la ventana gráfica actual que puede contener la ubicación y/u orientación del usuario y/o la dirección del movimiento (o cualquier información que caracterice los cambios provocados por las acciones del usuario) desde el lado del dispositivo de consumo de medios.

▪ recibir información sobre y la ROI señalada en los metadatos (ventanas gráficas de vídeo señaladas como en la especificación OMAF).

▪ recibir toda la información disponible en el extremo receptor;

▪ decidir en función de toda la información recibida y/o derivada de los metadatos recibidos y/o disponibles, qué punto de vista de audio/ vídeo debe reproducirse en un momento determinado. Por ejemplo, el procesador de ventanas gráficas 1232 puede decidir que:

• se reproduce una escena de audio completa

• se debe crear una nueva escena de audio a partir de todas las escenas de audio disponibles (por ejemplo, solo se reproducirán algunos elementos de audio de múltiples escenas de audio, mientras que otros elementos de audio restantes de estas escenas de audio no se reproducirán)

• se debe reproducir una transición entre dos o más escenas de audio

o una parte de selección 1230 configurada para seleccionar, en función de la información recibida del procesador de ventanas gráficas 1232 uno o más, los conjuntos de adaptación fuera de los conjuntos de adaptación disponibles como se indica en la información recibida por el extremo receptor; los ajustes de adaptación seleccionados describen completa o parcialmente la escena de audio que debe reproducirse en la ubicación actual del usuario. Esta escena de audio puede ser una escena de audio completa o parcialmente completa como se define en el lado de la codificación o se debe crear una nueva escena de audio a partir de todas las escenas de audio disponibles.

▪ Además, en el caso de que los elementos de audio 152 que pertenecen a más de una escena de audio, al menos un conjunto de adaptación se seleccione en función de la información que indique la asociación de al menos un conjunto de adaptación con más de una escena de audio, que contengan el mismo elementos de audio 152.

▪ Además, en el caso de que se produzca una transición entre dos o más escenas de audio en función de la indicación del procesador de ventanas gráficas 1232, la parte de selección 1230 puede estar configurada para seleccionar uno o más conjuntos de adaptación de entre los conjuntos de adaptación disponibles, tal como se indica en la información recibida por el extremo receptor; los conjuntos de adaptación seleccionados describen completamente la escena de audio que puede ser necesario reproducir en un futuro próximo (por ejemplo, si el usuario camina en la dirección de una escena de audio siguiente con una cierta velocidad, se puede predecir que se requiere la escena de audio

siguiente y se selecciona antes de la reproducción).

▪ Además, algunos conjuntos de adaptación correspondientes a las ubicaciones vecinas pueden seleccionarse primero a una tasa de bits inferior (es decir, una representación codificada a una tasa de bits más baja se elige entre las representaciones disponibles en un conjunto de adaptación), y en función de los cambios de posición, la calidad aumenta, seleccionando una tasa de bits más alta para esos conjuntos de adaptación específicos (es decir, una representación codificada a una tasa de bits más alta se selecciona de las representaciones disponibles en un conjunto de adaptación).

◦ una parte de descarga y conmutación que se puede configurar para:

▪ solicitud, basada en la indicación recibida de la parte de selección 1230, uno o más, conjuntos de adaptación fuera de los conjuntos de adaptación disponibles del servidor de medios 120;

▪ recibir, uno o más, conjuntos de adaptación (es decir, una representación de todas las representaciones disponibles dentro de cada conjunto de adaptación) fuera de los conjuntos de adaptación disponibles del servidor de medios 120;

▪ extraer información de metadatos de todas las transmisiones de audio recibidos

o un procesador de metadatos 1236 que puede configurarse para:

▪ recibir de la información de descarga y cambio acerca de las transmisiones de audio recibidos, información que puede incluir los metadatos de audio correspondientes a cada transmisión de audio recibido

▪ procesar y manipular los metadatos de audio asociados con cada transmisión de audio, en función de la información recibida del procesador de ventanas gráficas 1232 que puede incluir información sobre la ubicación y/u orientación del usuario y/o la dirección del movimiento, con el fin de:

seleccionar/habilitar los elementos de audio requeridos 152 que componen la nueva escena de audio como lo indica el procesador de ventanas gráficas 1232;

• permitir la combinación de todas las transmisiones de audio en una sola transmisión de audio.

◦ un multiplexador de transmisiones/fusionador 1238 que puede configurarse para fusionar todas las transmisiones de audio seleccionadas en una transmisión de audio en función de la información recibida del procesador de metadatos 1236, que puede incluir los metadatos de audio modificados y procesados correspondientes a todas las transmisiones de audio recibidos

◦ un decodificador de medios configurado para recibir y decodificar al menos una transmisión de audio para la reproducción de la nueva escena de audio, según lo indicado por el procesador de ventanas gráficas 1232, según la información sobre la ubicación y/o la orientación del usuario y/o la dirección del movimiento

La figura 1.6 muestra un sistema que comprende, en el lado del cliente, un sistema (sistema de cliente) que puede incorporar, por ejemplo, partes del sistema descrito en la figura 5 que adicional o alternativamente comprende:

• múltiples decodificadores de medios que pueden configurarse para decodificar las transmisiones de audio individuales según lo indicado por el procesador de metadatos 1236 (por ejemplo, con una serie de elementos de audio desactivados).

• un mezclador/renderizador 1238 que puede configurarse reproduce la escena de audio final en base a la información sobre la ubicación y/o la orientación del usuario y/o la dirección del movimiento (es decir, algunos de los elementos de audio que no son audibles en esa ubicación específica deben estar deshabilitados o no representados)

Actualizaciones del formato de archivo para reproducción de archivos

Para el caso de uso del formato de archivo, se pueden encapsular múltiples transmisiones principales y auxiliares como Pistas separadas en un único archivo ISOBMFF. Una sola pista de dicho archivo representaría un solo elemento de audio como se mencionó anteriormente. Dado que no hay un MPD disponible, que contenga la información necesaria para una reproducción correcta, la información debe proporcionarse en el nivel de formato de archivo, por ejemplo, al proporcionar/introducir un cuadro de formato de archivo específico o cuadros de formato de archivo específico en el nivel de pista y película. Dependiendo del caso de uso, hay diferente información necesaria para permitir una representación correcta de las escenas de audio encapsuladas, sin embargo, el siguiente conjunto de información es fundamental y, por lo tanto, siempre debe estar presente:

• Información sobre las escenas de audio incluidas, por ejemplo, "límites de ubicación"

• Información sobre todos los elementos de audio disponibles, especialmente qué elemento de audio está encapsulado en qué pista

• Información sobre la ubicación de los elementos de audio encapsulados

• Una lista de todos los elementos de audio que pertenecen a una escena de audio, un elemento de audio puede pertenecer a varias escenas de audio

Con esta información, todos los casos de uso mencionados, incluido el que tiene el procesador de metadatos adicional y la codificación compartida, también deberían funcionar en un entorno basado en archivos.

Otras consideraciones acerca de los ejemplos anteriores

En los ejemplos (por ejemplo, al menos uno entre las figuras 1.1–6), al menos una escena puede asociarse con al menos un elemento de audio (fuente de audio 152), cada elemento de audio está asociado a una posición y/o área en el visual entorno en el que el elemento de audio es audible, de modo que se proporcionan diferentes transmisiones de audio desde el sistema 120 del servidor al sistema 102 del cliente para diferentes posiciones del usuario y/o visores y/u orientaciones de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos posicionales virtuales en la escena.

En los ejemplos, el sistema de cliente 102 puede configurarse para decidir si al menos un elemento de audio 152 de una transmisión de audio (por ejemplo, A1, A2) y/o un conjunto de adaptación se reproducirá en presencia de la ventana gráfica del usuario actual y/u orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o posición virtual en la escena, en donde el sistema 102 está configurado para solicitar y/o recibir al menos un elemento de audio en la posición virtual del usuario actual.

En los ejemplos, el sistema de cliente (por ejemplo, 102) puede configurarse para decidir de manera predecible si al menos un elemento de audio (152) de una transmisión de audio y/o un conjunto de adaptación se volverán relevantes y/o audibles en función de al menos la ventana gráfica actual del usuario y/u orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales (110), y en el que el sistema está configurado para solicitar y/o recibir al menos un elemento de audio y/o transmisión de audio y/o adaptación establecidos en la posición virtual de un usuario particular antes del movimiento y/o interacción predichos del usuario en la escena, en donde el sistema está configurado para reproducir al menos en el elemento de audio y/o la transmisión de audio, cuando se recibe, en la posición virtual del usuario particular después del movimiento y/o interacción del usuario en la escena. Véanse, por ejemplo, las figuras 8A y 8B. En algunos ejemplos, al menos una de las operaciones del sistema 102 o 120 puede realizarse sobre la base de datos predictivos y/o estadísticos y/o agregados.

En los ejemplos, el sistema de cliente (por ejemplo, 102) puede configurarse para solicitar y/o recibir el al menos un elemento de audio (por ejemplo, 152) a una tasa de bits y/o nivel de calidad más bajos, en la posición virtual del usuario antes de que el usuario movimiento y/o interacción en la escena, en donde el sistema está configurado para solicitar y/o recibir el al menos un elemento de audio a una tasa de bits y/o nivel de calidad más altos, en la posición virtual del usuario después del movimiento y/o interacción del usuario en la escena. Véase, por ejemplo, la figura 7B.

En los ejemplos, al menos un elemento de audio puede estar asociado a al menos una escena, al menos un elemento de audio está asociado a una posición y/o área en el entorno visual asociado a la escena. El sistema está configurado para solicitar diferentes transmisiones a diferentes tasas de bits y/o niveles de calidad para elementos de audio en función de su relevancia y/o nivel de audibilidad en la posición virtual de cada usuario en la escena. El sistema está configurado para solicitar una transmisión de audio a una tasa de bits y/o nivel de calidad más altos para elementos de audio que son más relevantes y/o más audibles en la posición virtual del usuario actual, y/o una transmisión de audio a una tasa de bits inferior y/o nivel de calidad para los elementos de audio que son menos relevantes y/o menos audibles en la posición virtual del usuario actual. Véase, en términos generales, la figura 7A. Véase también las figuras 2a y 2b (en donde las fuentes más relevantes y/o audibles pueden ser las más cercanas al usuario), figura 3 (en donde la fuente más relevante y/o audible es la fuente de la escena 150a cuando el usuario está en la posición x_1, y la fuente más relevante y/o audible es la fuente de la escena 150b cuando el usuario está en la posición x_3), figura 4 (en donde, en el instante t_2, las fuentes más relevantes y/o audibles pueden ser las de la primera escena), figura 6 (donde las fuentes más audibles pueden ser aquellas que el usuario ve de frente).

En los ejemplos, al menos un elemento de audio (152) está asociado a una escena, cada elemento de audio está asociado a una posición y/o área en el entorno visual asociado a la escena, en donde el sistema de cliente 102 está configurado para enviar periódicamente al sistema de servidor 120 es la ventana gráfica actual del usuario y/o los datos de orientación y/o movimiento de la cabeza y/o los metadatos de interacción y/o los datos de posición virtuales (110), de modo que: para una posición más cercana a al menos un elemento de audio (152), se proporciona una

transmisión a mayor tasa de bits y/o calidad, desde el servidor, y para una posición más alejada de al menos un elemento de audio (152), se proporciona una transmisión a una menor tasa de bits y/o calidad, desde el servidor. Véase, por ejemplo, las figuras 2a y 2b.

Una pluralidad de escenas (por ejemplo, 150A, 150B) para múltiples entornos visuales, como entornos adyacentes y/o vecinos, de modo que las primeras secuencias se proporcionen asociadas a una primera escena actual (por ejemplo, 150A) y, en caso de transición del usuario (150AB) a una segunda escena adicional (por ejemplo, 150B), para proporcionar tanto las secuencias asociadas a la primera escena como las segundas secuencias asociadas a la segunda escena. Véase la figura 3.

Se definen una pluralidad de escenas para un primer y un segundo entornos visuales, siendo el primero y el segundo entornos adyacentes y/o entornos adyacentes, en donde se proporcionan las primeras secuencias asociadas a la primera escena, desde el servidor, para la reproducción de la primera escena en caso de que la posición virtual del usuario esté en un primer entorno asociado a la primera escena, se proporcionan las segundas transmisiones asociadas a la segunda escena, desde el servidor, para la reproducción de la segunda escena en caso de que la posición virtual del usuario sea en un segundo entorno asociado a la segunda escena, y tanto las primeras secuencias asociadas a la primera escena como las segundas secuencias asociadas a la segunda escena se proporcionan en caso de que la posición virtual del usuario esté en una posición de transición entre la primera escena y la segunda escena. Véase, por ejemplo, la figura 3

Las primeras transmisiones asociadas a la primera escena se obtienen a una tasa de bits y/o calidad superiores cuando el usuario está en el primer entorno asociado a la primera escena, mientras que las segundas transmisiones asociadas al segundo entorno de escena asociadas al segundo entorno son obtenidas a una tasa de bits y/o calidad inferior cuando el usuario se encuentra en el comienzo de una posición de transición desde la primera escena a la segunda escena, y las primeras transmisiones asociadas a la primera escena se obtienen a una tasa de bits y/o calidad inferior y la las segundas transmisiones asociadas a la segunda escena se obtienen a una tasa de bits y/o calidad superior cuando el usuario se encuentra en el final de una posición de transición de la primera escena a la segunda escena. Este es el caso de figura 3.

Una pluralidad de escenas (por ejemplo, 150A, 150B) se define para múltiples entornos visuales (por ejemplo, entornos adyacentes), de modo que el sistema 102 puede solicitar y/u obtener las transmisiones asociadas a la escena actual a una tasa de bits más alta y/o la calidad y las transmisiones asociadas a la segunda escena a una tasa de bits y/o calidad inferior. Véase, por ejemplo, la figura 4.

En los ejemplos, se define una pluralidad de N elementos de audio y, en caso de que la distancia del usuario a la posición o área de estos elementos de audio sea mayor que un umbral predeterminado, los N elementos de audio se procesan para obtener un número menor de elementos de audio M (M<N) asociado a una posición o área cercana a la posición o área de los N elementos de audio, a fin de proporcionar al sistema al menos una transmisión de audio asociada a los N elementos de audio, en caso de que la distancia del usuario a la posición o el área de los N elementos de audio sea más pequeña que un umbral predeterminado, o para proporcionar al sistema al menos una transmisión de audio asociada a los M elementos de audio, en caso de que la distancia del usuario a la posición o área de los N elementos de audio sea mayor que un umbral predeterminado. Véase, por ejemplo, la figura 1.7.

En los ejemplos, al menos una escena de entorno visual está asociada a al menos una pluralidad de N elementos de audio (N>=2), cada elemento de audio está asociado a una posición y/o área en el entorno visual, en donde al menos al menos se puede proporcionar una pluralidad de N elementos de audio en al menos una representación a un nivel de calidad de bits y/o de alta tasa de bits, y en la que al menos una pluralidad de elementos de audio N se proporciona en al menos una representación a una tasa de bits y/o de calidad baja nivel, donde la al menos una representación se obtiene al procesar los N elementos de audio para obtener un número M más pequeño de elementos de audio (M<N) asociados a una posición o área cercana a la posición o área de los N elementos de audio, en donde el sistema está configurado para solicitar la representación a una tasa de bits y/o nivel de calidad más altos para los elementos de audio, en caso de que los elementos de audio sean más relevantes y/o más audibles en la posición virtual del usuario actual en la escena, en donde el sistema se configura para solicitar la representación a menor tasa de bits y/o nivel de calidad para los elementos de audio, en caso de que los elementos de audio sean menos relevantes y/o menos audibles en la posición virtual del usuario actual en la escena. Véase, por ejemplo, la figura 1.7.

En algunos ejemplos, en caso de que la distancia del usuario y/o la relevancia y/o el nivel de audibilidad y/o la orientación angular sean inferiores al umbral predeterminado, se obtienen transmisiones diferentes para los distintos elementos de audio. Véase, por ejemplo, la figura 1.7.

En los ejemplos, se proporcionan diferentes elementos de audio en diferentes ventanas gráficas, de modo que, en caso de que un primer elemento de audio se encuentre dentro de una ventana gráfica actual, el primer elemento de audio se obtiene a una tasa de bits más alta que un segundo elemento de audio que no entra dentro de la ventana gráfica. Véase, por ejemplo, la figura 6.

En los ejemplos, se definen al menos dos escenas de entorno visual, en donde al menos uno de los elementos de audio primero y segundo está asociado a una primera escena asociada a un primer entorno de visión, y al menos un tercer elemento de audio está asociado a una segunda escena asociada a una segundo entorno visual, en el que el sistema 102 está configurado para obtener metadatos que describen que el al menos un segundo elemento de audio está asociado adicionalmente con la segunda escena del entorno visual, y en el que el sistema está configurado para solicitar y/o recibir el al menos primero y segundo elementos de audio, en caso de que la posición virtual del usuario se encuentre en el primer entorno visual, y en el que el sistema esté configurado para solicitar y/o recibir al menos el segundo y tercer elementos de audio, en caso de que la posición virtual del usuario se encuentre en la segunda escena del entorno visual, y en el que el sistema está configurado para solicitar y/o recibir al menos el primer y el segundo y el tercer elementos de audio, en caso de que la posición virtual del usuario esté en transición entre la primera escena del entorno visual y la segunda escena del entorno visual. Véase, por ejemplo, la figura 1.4. Esto también puede aplicarse a la figura 3.

En los ejemplos, al menos un primer elemento de audio puede proporcionarse en al menos una transmisión de audio y/o conjunto de adaptación, y el al menos un segundo elemento de audio se proporciona en al menos una segunda transmisión de audio y/o conjunto de adaptación, y al menos un tercer elemento de audio se proporciona en al menos una tercera transmisión de audio y/o conjunto de adaptación, y en el que la escena del entorno visual al menos la primera es descrita por metadatos como una escena completa que requiere al menos la primera y la segunda transmisiones de audio y/o conjuntos de adaptación, y en donde la escena del segundo entorno visual se describe por metadatos como una escena incompleta que requiere al menos un tercer conjunto de adaptación y/o transmisión de audio y al menos un segundo conjunto de adaptación y/o transmisión de audio asociados con al menos la primera escena del entorno visual, en donde el sistema comprende un procesador de metadatos configurado para manipular los metadatos, para permitir la fusión de la segunda transmisión de audio que pertenece al primer entorno visual y al tercer audio la ventana gráfica actual del usuario y/o la orientación asociado con el segundo entorno visual en una nueva transmisión única, en caso de que la posición virtual del usuario se encuentre en el segundo entorno visual. Véase Por ejemplo, las figuras 1.2-1.3, 1.5 y 1.6.

En los ejemplos, el sistema 102 puede comprender un procesador de metadatos (por ejemplo, 1236) configurado para manipular los metadatos en al menos una transmisión de audio antes del al menos un decodificador de audio, en función de la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o metadatos de interacción y/o datos posicionales virtuales.

En los ejemplos, el procesador de metadatos (por ejemplo, 1236) puede configurarse para habilitar y/o deshabilitar al menos un elemento de audio en al menos una transmisión de audio antes de al menos un decodificador de audio, en función de la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales, en donde el procesador de metadatos puede configurarse para deshabilitar al menos un elemento de audio en al menos una transmisión de audio antes del al menos un decodificador de audio, en caso de que el sistema decida que el elemento de audio ya no debe reproducirse como consecuencia de una ventana gráfica actual y/o una orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales, y en el que el procesador de metadatos puede configurarse para permitir al menos un elemento de audio en al menos una transmisión de audio antes de al menos un decodificador de audio, en caso de que el sistema decida que el elemento de audio debe reproducirse como consecuencia de la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales.

Lado del servidor

Aquí anteriormente también se hace referencia a un servidor (120) para entregar transmisiones de audio y vídeo a un cliente para una realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR o entorno de vídeo de 360 grados, las transmisiones de vídeo y audio para reproducirse en un dispositivo de consumo de medios, en el que el servidor (120) comprende un codificador para codificar y/o un almacenamiento para almacenar transmisiones de vídeo para describir un entorno visual, estando el entorno visual asociado a una escena de audio; en el que el servidor comprende además un codificador para codificar y/o un almacenamiento para almacenar una pluralidad de transmisiones y/o elementos de audio y/o conjuntos de adaptación para ser entregados al cliente, siendo las transmisiones y/o elementos de audio y/o conjuntos de adaptación asociados a al menos una escena de audio, en donde el servidor está configurado para:

seleccionar y entregar una transmisión de vídeo en base a una solicitud del cliente, donde la transmisión de vídeo se asocia a un entorno;

seleccionar una secuencia de audio y/o un elemento de audio y/o un conjunto de adaptación sobre la base de una solicitud del cliente, la solicitud se asocia a al menos la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales y a una escena de

audio asociada al entorno; y

entregar la transmisión de audio al cliente.

Otras formas de realización y variantes

Dependiendo de ciertos requisitos de implementación, los ejemplos pueden implementarse en hardware. La implementación se puede realizar utilizando un medio de almacenamiento digital, por ejemplo, un disquete, un disco versátil digital (DVD), un disco Blu–Ray, un disco compacto (CD), una memoria de solo lectura (ROM), una lectura programable de solo memoria (PROM), una memoria de solo lectura borrable y programable (EPROM), una memoria de solo lectura programable borrable eléctricamente (EEPROM) o una memoria flash, que tiene señales de control legibles electrónicamente almacenadas en ella, que cooperan (o son capaces de cooperar) con un sistema informático programable tal que se realice el método respectivo. Por lo tanto, el medio de almacenamiento digital puede ser legible por ordenador.

En general, los ejemplos pueden implementarse como un producto de programa informático con instrucciones de programa, las instrucciones de programa son operativas para realizar uno de los métodos cuando el producto de programa informático se ejecuta en un ordenador. Las instrucciones del programa pueden, por ejemplo, almacenarse en un medio legible por máquina.

Otros ejemplos comprenden el programa informático para realizar uno de los métodos descritos en este documento, almacenado en un soporte legible por máquina. En otras palabras, un ejemplo de método es, por lo tanto, un programa informático que tiene instrucciones de un programa para realizar uno de los métodos descritos aquí, cuando el programa informático se ejecuta en un ordenador.

Un ejemplo adicional de los métodos es, por lo tanto, un medio de soporte de datos (o un medio de almacenamiento digital, o un medio legible por ordenador) que comprende, grabado en el mismo, el programa informático para realizar uno de los métodos descritos en el presente documento. El medio portador de datos, el medio de almacenamiento digital o el medio grabado son tangibles y/o no transitorios, en lugar de señales que son intangibles y transitorias.

Un ejemplo adicional comprende una unidad de procesamiento, por ejemplo, un ordenador, o un dispositivo lógico programable que realiza uno de los métodos descritos en este documento.

Un ejemplo adicional comprende un ordenador que tiene instalado en ella el programa informático para realizar uno de los métodos descritos en este documento.

Un ejemplo adicional comprende un aparato o un sistema que transfiere (por ejemplo, electrónica u ópticamente) un programa informático para realizar uno de los métodos descritos en el presente documento a un receptor. El receptor puede ser, por ejemplo, un ordenador, un dispositivo móvil, un dispositivo de memoria o similar. El aparato o sistema puede comprender, por ejemplo, un servidor de archivos para transferir el programa informático al receptor.

En algunos ejemplos, se puede usar un dispositivo lógico programable (por ejemplo, una matriz de puertas programable de campo) para realizar algunas o todas las funcionalidades de los métodos descritos en este documento. En algunos ejemplos, una matriz de puerta programable de campo puede cooperar con un microprocesador para realizar uno de los métodos descritos en este documento. En general, los métodos pueden ser realizados por cualquier aparato de hardware apropiado.

Los ejemplos descritos anteriormente son ilustrativos de los principios discutidos anteriormente. Se entiende que las modificaciones y variaciones de las disposiciones y los detalles descritos en este documento serán evidentes. Por lo tanto, la intención es estar limitada por el alcance de acuerdo con las reivindicaciones de patentes inminentes y no por los detalles específicos presentados a modo de descripción y explicación de los ejemplos en el presente documento.

De acuerdo con un primer aspecto, se proporciona un sistema para una realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR o entorno de vídeo de 360 grados configurado para recibir transmisiones de vídeo y audio para ser reproducidas en un dispositivo de consumo de medios, en donde el sistema comprende: al menos un decodificador de vídeo multimedia configurado para decodificar señales de vídeo de transmisiones de vídeo para la representación de escenas de entornos de vídeo VR, AR, MR o 360 grados para un usuario, y al menos un decodificador de audio configurado para decodificar señales de audio de al menos una transmisión de audio, en el que el sistema puede configurarse para solicitar al menos una transmisión de audio y/o un elemento de audio de una transmisión de audio y/o un conjunto de adaptación a un servidor sobre la base de al menos la ventana gráfica actual del usuario y/u orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales.

De acuerdo con un segundo aspecto se proporciona el sistema del primer aspecto configurado para proporcionar al servidor la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de

interacción y/o los datos de posición virtuales para obtener al menos una secuencia de audio y/o un elemento de audio de una transmisión de audio y/o un conjunto de adaptación del servidor.

De acuerdo con un tercer aspecto, se proporciona el sistema según el primer o segundo aspecto, en el que al menos una escena está asociada a al menos un elemento de audio, donde cada elemento de audio está asociado a una posición y/o área en el entorno visual donde el elemento de audio es audible, de modo que se proporcionan diferentes transmisiones de audio para diferentes posiciones del usuario y/o ventanas gráficas y/u orientaciones de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos virtuales de posición en la escena.

De acuerdo con un cuarto aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores configurado para decidir si al menos un elemento de audio de una transmisión de audio y/o un conjunto de adaptación se reproducirán para la ventana del usuario actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o posición virtual en la escena, y donde el sistema está configurado para solicitar y/o recibir el al menos un elemento de audio en la posición virtual del usuario actual.

De acuerdo con un quinto aspecto, se proporciona el sistema de acuerdo con cualquiera de los aspectos anteriores, en el que el sistema está configurado para decidir de manera predecible si al menos un elemento de audio de una transmisión de audio y/o un conjunto de adaptación se volverán relevantes y/o audibles en función de al menos la vista actual y la orientación de la cabeza del usuario y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales, y en el que el sistema puede configurarse para solicitar y/o recibir al menos un elemento de audio y/o transmisión de audio y/o conjunto de adaptación en una determinada posición virtual del usuario antes del movimiento y/o interacción predichos del usuario en la escena, en donde el sistema está configurado para reproducir al menos en el elemento de audio y/o transmisión de audio, cuando se recibe, en la posición virtual del usuario particular después del movimiento del usuario y/o la interacción en la escena.

De acuerdo con un sexto aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores configurado para solicitar y/o recibir el al menos un elemento de audio a una tasa de bits y/o nivel de calidad más bajos, en la posición virtual del usuario ante el movimiento y/o interacción de un usuario en la escena, en donde el sistema está configurado para solicitar y/o recibir el al menos un elemento de audio a una tasa de bits y/o nivel de calidad más altos, en la posición virtual del usuario después del movimiento y/o interacción del usuario en la escena.

De acuerdo con un séptimo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores en el que al menos un elemento de audio está asociado a al menos una escena en donde cada elemento de audio está asociado a una posición y/o área en el entorno visual asociado a la escena, en donde el sistema puede ser configurado para solicitar y/o recibir transmisiones a mayor tasa de bits y/o calidad para elementos de audio más cercanos al usuario que para elementos de audio más alejados del usuario.

De acuerdo con un octavo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en donde al menos un elemento de audio está asociado a una posición y/o área en el entorno visual asociado a la escena, en donde el sistema está configurado para solicitar diferentes transmisiones a diferentes velocidades de bits y/o niveles de calidad para elementos de audio según su relevancia y/o nivel de audibilidad en la posición virtual de cada usuario en la escena, en el que el sistema está configurado para solicitar una transmisión de audio a una tasa de bits mayor y/o nivel de calidad para los elementos de audio que son más relevantes y/o más audibles en la posición virtual del usuario actual, y/o una transmisión de audio a una tasa de bits más baja y/o nivel de calidad para los elementos de audio que son menos relevantes y/o menos audibles en la posición virtual del usuario actual.

De acuerdo con un noveno aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en el que al menos un elemento de audio está asociado a una escena, cada elemento de audio asociado a una posición y/o área en el entorno visual asociado a la escena, en donde el sistema está configurado para enviarse periódicamente al servidor la ventana gráfica actual y/o los datos de movimiento y/u orientación de la cabeza y/o metadatos de interacción y/o datos posicionales virtuales, de modo que: para una primera posición, se proporciona una transmisión a mayor tasa de bits y/o calidad, desde el servidor, y para una segunda posición, se proporciona una transmisión a menor tasa de bits y/o calidad, desde el servidor, en donde la primera posición está más cerca del al menos un elemento de audio que la segunda posición.

De acuerdo con un décimo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en donde una pluralidad de escenas están definidas para múltiples entornos visuales, como entornos adyacentes y/o vecinos, de modo que se proporcionen primeras secuencias asociadas a una primera escena actual y, en caso de que el usuario realice la transición a una segunda escena posterior, se proporcionen tanto las secuencias asociadas a la primera escena como las segundas secuencias asociadas a la segunda escena.

De acuerdo con un undécimo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en el que se definen una pluralidad de escenas para un primer y un segundo entornos visuales, siendo el primero y el segundo entornos adyacentes y/o entornos adyacentes, en donde se proporcionan las primeras secuencias asociadas

a la primera escena, desde el servidor, para la reproducción de la primera escena en caso de que la posición del usuario o la posición virtual se encuentre en un primer entorno asociado a la primera escena, se proporcionan las segundas transmisiones asociadas a la segunda escena, desde el servidor, para la reproducción de la segunda escena en caso de la posición del usuario o la posición virtual en un segundo entorno asociado a la segunda escena, y tanto las primeras transmisiones asociadas a la primera escena como las segundas transmisiones asociadas a la segunda escena se proporcionan en caso de que la posición del usuario o la posición virtual estén en una posición de transición entre la primera escena y la segunda escena.

De acuerdo con un duodécimo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en donde se definen una pluralidad de escenas para un primer y un segundo entornos visuales, que son entornos adyacentes y/o vecinos, en donde el sistema está configurado para solicitar y/o recibir primeras transmisiones asociadas a una primera escena asociada al primer entorno, para la reproducción de la primera escena en caso de que la posición virtual del usuario se encuentre en el primer entorno, en donde el sistema está configurado para solicitar y/o recibir segundas transmisiones asociadas a la segunda escena asociada al segundo entorno, para la reproducción de la segunda escena en caso de que la posición virtual del usuario se encuentre en el segundo entorno, y donde el sistema está configurado para solicitar y/o recibir tanto primeras transmisiones asociadas a la primera escena como segundas transmisiones asociadas a la segunda escena en caso de que la posición virtual del usuario se encuentre en una posición de transición entre el primer entorno y el segundo entorno.

De acuerdo con un cuarto aspecto, se proporciona el sistema según los aspectos décimo a duodécimo, donde las primeras transmisiones asociadas a la primera escena se obtienen a una mayor velocidad de bits y/o calidad cuando el usuario se encuentra en el primer entorno asociado a la primera escena, mientras que las segundas transmisiones asociadas a la segunda escena asociada al segundo entorno se obtienen a una tasa de bits y/o calidad inferior cuando el usuario se encuentra al principio de una posición de transición desde la primera escena a la segunda escena, y las primeras transmisiones asociadas a la primera escena se obtienen a una tasa de bits y/o calidad inferior y las segundas transmisiones asociadas a la segunda escena se obtienen a una tasa de bits y/o calidad superior cuando el usuario se encuentra al final de una posición de transición desde la primera escena a la segunda escena, en la que la tasa de bits y/o calidad inferior es inferior a la tasa de bits y/o calidad superior.

De acuerdo con un decimocuarto aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en el que se define una pluralidad de escenas para múltiples entornos tales como entornos adyacentes y/o vecinos, de modo que el sistema está configurado para obtener las transmisiones asociadas a una primera escena actual asociada a un primer entorno actual, y en caso de que la distancia de la posición del usuario o la posición virtual desde un límite de la escena esté por debajo de un umbral predeterminado, el sistema obtiene transmisiones de audio asociadas a un segundo entorno adyacente y/o adyacente asociado a la segunda escena.

De acuerdo con un decimoquinto aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en el que se define una pluralidad de escenas para múltiples entornos visuales, de modo que el sistema solicite y/u obtenga las transmisiones asociadas a la escena actual a una tasa de bits y/o calidad superior y las transmisiones asociadas a la segunda escena a una tasa de bits y/o calidad inferior, en donde la tasa de bits y/o calidad inferior es menor que la tasa de bits y/o calidad superior.

De acuerdo con un decimosexto aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en donde se definen una pluralidad de N elementos de audio, y, en caso de que la distancia del usuario a la posición o área de estos elementos de audio sea mayor que un umbral predeterminado, los N elementos de audio se procesan para obtener un número M más pequeño de elementos de audio (M<N) asociados a una posición o área cercana a la posición o área de los N elementos de audio, a fin de proporcionar al sistema al menos una transmisión de audio asociada a los N elementos de audio, en caso de que la distancia del usuario a la posición o área de los N elementos de audio sea menor que un umbral predeterminado, o para proporcionar al sistema al menos una transmisión de audio asociada a los M elementos de audio, en caso de que la distancia del usuario a la posición o el área de los N elementos de audio es mayor que un umbral predeterminado.

De acuerdo con un aspecto decimoséptimo, se proporciona el sistema según cualquiera de los aspectos anteriores en el que al menos una escena de entorno visual se asocia a al menos una pluralidad de N elementos de audio (N>=2), en donde cada elemento de audio está asociado a una posición y/o área en el entorno visual, en el que al menos al menos una pluralidad de N elementos de audio se proporciona en al menos una representación a un alto nivel de velocidad de bits y/o calidad, y en el que al menos una pluralidad de N elementos de audio se proporciona en al menos una representación a baja tasa de bits y/o nivel de calidad, donde la al menos una representación se obtiene procesando los N elementos de audio para obtener un número menor de elementos de audio (M <N) asociados a una posición o área cercana a la posición o área de los N elementos de audio, en donde el sistema puede configurarse para solicitar la representación a una tasa de bits y/o nivel de calidad más altos para los elementos de audio, en caso de que los elementos de audio sean más relevantes y/o más audibles en la posición virtual del usuario actual en la escena, en donde el sistema está configurado para solicitar la representación a menor tasa de bits y/o nivel de calidad para los elementos de audio, en caso de que los elementos de audio sean menos relevantes y/o menos audibles en

la posición virtual del usuario actual en la escena.

De acuerdo con un decimoctavo aspecto, se proporciona el sistema según los aspectos decimosexto y decimoséptimo, en el que, en caso de que la distancia del usuario y/o la relevancia y/o el nivel de audibilidad y/o la orientación angular sean inferiores a un umbral predeterminado, se obtienen diferentes transmisiones para los diferentes elementos de audio.

De acuerdo con un decimonoveno aspecto, se proporciona el sistema de acuerdo con cualquiera de los aspectos anteriores, en el que el sistema está configurado para solicitar y/u obtener las transmisiones sobre la base de la orientación del usuario y/o la dirección de movimiento del usuario y/o las interacciones del usuario en la escena.

De acuerdo con un vigésimo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en donde la ventana está asociada a la posición y/o posición virtual y/o datos de movimiento y/o cabeza

De acuerdo con un vigésimo primer aspecto, se proporciona el sistema de acuerdo con cualquiera de los aspectos anteriores, en el que se proporcionan diferentes elementos de audio en diferentes ventanas gráficas, en donde el sistema está configurado para solicitar y/o recibir, en caso de que un primer elemento de audio (S1) caiga dentro de una ventana gráfica (160-1), el primer elemento de audio a una tasa de bits más alta que un segundo elemento de audio (S2) que no cae dentro de la ventana gráfica.

De acuerdo con un vigésimo segundo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores configurado para solicitar y/o recibir primeras transmisiones de audio y segundas transmisiones de audio, en donde los primeros elementos de audio en las primeras transmisiones de audio son más relevantes y/o más audibles que los segundos elementos de audio en las segundas transmisiones de audio, en donde las primeras transmisiones de audio se solicitan y/o reciben a una tasa de bits y/o calidad superior a la velocidad de bits y/o calidad de las segundas transmisiones de audio.

De acuerdo con un vigésimo tercer aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en donde se definen al menos dos escenas de entorno visual, en donde al menos uno de los elementos de audio primero y segundo está asociado a una primera escena asociada a un primer entorno de visión, y al menos un tercer elemento de audio está asociado a una segunda escena asociada a una segundo entorno visual, en el que el sistema está configurado para obtener metadatos que describen que el al menos un segundo elemento de audio está asociado adicionalmente con la segunda escena del entorno visual, y en el que el sistema está configurado para solicitar y/o recibir el al menos primero y segundo elementos de audio, en caso de que la posición virtual del usuario se encuentre en el primer entorno visual, y en el que el sistema esté configurado para solicitar y/o recibir al menos el segundo y tercer elementos de audio, en caso de que la posición virtual del usuario se encuentre en la segunda escena del entorno visual, y en el que el sistema está configurado para solicitar y/o recibir al menos el primer y el segundo y el tercer elementos de audio, en caso de que la posición virtual del usuario esté en transición entre la primera escena del entorno visual y la segunda escena del entorno visual.

De acuerdo con un aspecto vigésimo cuarto, se proporciona el sistema de acuerdo con el aspecto vigésimo tercero, en donde el al menos un primer elemento de audio se proporciona en al menos una transmisión de audio y/o conjunto de adaptación, y el al menos un segundo elemento de audio se proporciona en al menos una segunda transmisión de audio y/o conjunto de adaptación, y el al menos un tercer elemento de audio se proporciona en al menos una tercera transmisión de audio y/o conjunto de adaptación, y en donde la al menos primera escena de entorno visual se describe mediante metadatos como una escena completa que requiere las al menos primera y segunda transmisiones de audio y/o conjuntos de adaptación, y en donde la segunda escena de entorno visual es descrita por metadatos como una escena incompleta que requiere el al menos la tercera transmisión de audio y/o conjunto de adaptación y la al menos segunda transmisión de audio y/o conjuntos de adaptación asociados con la al menos primera escena de entorno visual, en donde el sistema comprende un procesador de metadatos configurado para manipular los metadatos, para permitir la fusión de la segunda transmisión de audio perteneciente al primer entorno visual y la tercera transmisión de audio asociada con el segundo entorno visual en una nueva transmisión única, en caso de que la posición virtual del usuario se encuentre en el segundo entorno visual.

De acuerdo con un vigésimo quinto aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en donde el sistema comprende un procesador de metadatos configurado para manipular los metadatos en al menos una transmisión de audio antes del al menos un decodificador de audio, en función de la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción. y/o datos posicionales virtuales.

De acuerdo con un vigésimo sexto aspecto, se proporciona el sistema según el aspecto vigésimo quinto, en el que el procesador de metadatos está configurado para habilitar y/o deshabilitar al menos un elemento de audio en al menos una transmisión de audio antes del al menos un decodificador de audio, basado en la ventana gráfica actual y/o la orientación y/o los datos del movimiento de la cabeza y/o metadatos de interacción y/o datos posicionales virtuales,

en donde el procesador de metadatos está configurado para deshabilitar al menos un elemento de audio en al menos una transmisión de audio antes de al menos un decodificador de audio, en caso de que el sistema decida que el elemento de audio ya no debe reproducirse como consecuencia de una vista actual y/o una orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales, y en el que el procesador de metadatos puede configurarse para permitir al menos un elemento de audio en al menos una transmisión de audio antes de al menos un decodificador de audio, en caso de que el sistema decida que el elemento de audio se va a reproducir como consecuencia de la vista actual de un usuario y/o la orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales.

De acuerdo con un vigésimo séptimo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, configurado para deshabilitar la decodificación de elementos de audio seleccionados en función de la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o la posición virtual.

De acuerdo con un vigésimo octavo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, configurado para fusionar al menos una primera transmisión de audio asociada a la escena de audio actual a al menos una transmisión asociada a una escena de audio vecina, adyacente y/o futura.

De acuerdo con un vigésimo noveno aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, configurado para obtener y/o recopilar datos estadísticos o agregados en la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o datos virtuales de posición, para transmitir una solicitud al servidor asociado a los datos estadísticos o agregados.

De acuerdo con un trigésimo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores configurado para desactivar la decodificación y/o la reproducción de al menos una transmisión sobre la base de metadatos asociados a al menos una transmisión y sobre la base de la ventana gráfica actual y/o la orientación de la cabeza del usuario y/o datos de movimiento y/o metadatos y/o datos posicionales virtuales.

De acuerdo con un trigésimo primer aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, configurado además para: manipular los metadatos asociados con un grupo de transmisiones de audio seleccionadas, en función de al menos la vista actual o estimada del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o los datos de posición virtuales para: seleccionar y/o habilitar y/o activar elementos de audio que componen la escena de audio a reproducir; y/o habilite la combinación de todas las transmisiones de audio seleccionadas en una sola transmisión de audio.

De acuerdo con un trigésimo segundo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, configurado para controlar la solicitud de al menos una transmisión al servidor sobre la base de la distancia de la posición del usuario desde los límites de entornos vecinos y/o adyacentes asociados a diferentes escenas u otras métricas asociadas a la posición del usuario en el entorno actual o las predicciones sobre el entorno futuro.

De acuerdo con un trigésimo tercer aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en donde la información se proporciona desde el sistema servidor (120), para cada elemento de audio u objeto de audio, en donde la información incluye información descriptiva sobre las ubicaciones en las que la escena de sonido o los elementos de audio están activos.

De acuerdo con un trigésimo cuarto aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, configurado para elegir entre reproducir una escena y componer o mezclar o multiplexar o superponer o combinar al menos dos escenas sobre la base de la orientación actual o futura o de la ventana gráfica y/o la orientación y/o movimiento, y/o los metadatos y/o la posición virtual y/o la selección de un usuario, en donde las dos escenas están asociadas a diferentes entornos vecinos y/o adyacentes.

De acuerdo con un trigésimo quinto aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, configurado para crear o utilizar al menos los conjuntos de adaptación de modo que: Se asocian varios conjuntos de adaptación con una escena de audio; y/o se proporciona información adicional que relaciona cada conjunto de adaptación con un punto de vista o una escena de audio; y/o

Se facilita información adicional que puede incluir

- Información sobre los límites de una escena de audio y/o

- Información sobre la relación entre un conjunto de adaptación y una escena de audio (por ejemplo, la escena de audio se codifica en tres transmisiones que se encapsulan en tres conjuntos de adaptación) y/o

- Información sobre la conexión entre los límites de la escena de audio y los múltiples conjuntos de adaptación.

De acuerdo con un trigésimo sexto aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, configurado para: recibir una transmisión para una escena asociada a un entorno vecino o adyacente; comience a decodificar y/o reproducir ella ventana gráfica actual del usuario y/o la orientación para el entorno vecino o adyacente en la detección de la transición de un límite entre dos entornos.

De acuerdo con un trigésimo séptimo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores configurado para funcionar como un cliente y un servidor configurado para entregar transmisiones de vídeo y/o audio para reproducirse en un dispositivo de consumo de medios.

De acuerdo con un trigésimo octavo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, en el que el sistema está configurado además para: solicitar y/o recibir al menos un primer conjunto de adaptación que comprende al menos una transmisión de audio asociada con al menos una primera escena de audio; solicitar y/o recibir al menos un segundo conjunto de adaptación que comprenda al menos una segunda transmisión de audio asociada con al menos dos escenas de audio, incluida la al menos una primera escena de audio; y permitir una fusión de la al menos una primera transmisión de audio y la de al menos una segunda transmisión de audio en una nueva transmisión de audio que se decodificará, en función de los metadatos disponibles con respecto a la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o metadatos y/o datos de posición virtual y/o información que describe una asociación de al menos un primer conjunto de adaptación a al menos una primera escena de audio y/o una asociación de al menos un segundo conjunto de adaptación a al menos una primera escena de audio.

De acuerdo con un trigésimo noveno aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, configurado para recibir información sobre la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o los datos de posición virtuales y/o cualquier información que caracterice los cambios activados por las acciones del usuario; y recibir información sobre la disponibilidad de conjuntos de adaptación e información que describe una asociación de al menos un conjunto de adaptación a al menos una escena y/o punto de vista y/o ventana gráfica y/o posición y/o posición virtual y/o datos de movimiento y/u orientación.

De acuerdo con un cuadragésimo aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, configurado para decidir si al menos un elemento de audio de al menos una escena de audio incrustada en al menos una transmisión y al menos un elemento de audio adicional de al menos una escena de audio adicional incrustada en al menos una transmisión adicional deben ser reproducidos; y causar, en caso de una decisión positiva, una operación de fusionar, componer, multiplexar, superponer o combinar al menos una transmisión adicional de la escena de audio adicional a la al menos una transmisión de la al menos una escena de audio.

De acuerdo con un cuadragésimo primer aspecto, se proporciona el sistema según cualquiera de los aspectos anteriores, configurado para manipular los metadatos de audio asociados con las transmisiones de audio seleccionadas, en función de al menos la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o los datos de posición virtuales, para: seleccionar y/o habilitar y/o activar los elementos de audio que componen la escena de audio que se decidió reproducir; y activar la fusión de todas las transmisiones de audio seleccionadas en una sola secuencia de audio.

De acuerdo con un aspecto cuadragésimo segundo, se proporciona un servidor para entregar transmisiones de audio y vídeo a un cliente para un entorno de realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR, o vídeo de 360 grados, las transmisiones de vídeo y audio serán reproducidos en un dispositivo de consumo de medios, en el que el servidor comprende un codificador para codificar y/o un almacenamiento para almacenar transmisiones de vídeo para describir un entorno visual, estando asociado el entorno visual a una escena de audio; en el que el servidor comprende además un codificador para codificar y/o un almacenamiento para almacenar una pluralidad de secuencias y/o elementos de audio y/o conjuntos de adaptación para ser entregados al cliente, las secuencias y/o elementos de audio y/o conjuntos de adaptación estar asociado a al menos una escena de audio, en donde el servidor está configurado para: seleccionar y entregar una transmisión de vídeo en base a una solicitud del cliente, la transmisión de vídeo está asociada a un entorno; seleccionar una secuencia de audio y/o un elemento de audio y/o un conjunto de adaptación sobre la base de una solicitud del cliente, la solicitud se asocia a al menos la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o metadatos de interacción y/o datos de posición virtuales y a una escena de audio asociada al entorno; y entregar la transmisión de audio al cliente.

De acuerdo con un cuadragésimo tercer aspecto, se proporciona el servidor del aspecto cuadragésimo segundo, en el que las transmisiones se encapsulan en conjuntos de adaptación, cada conjunto de adaptación incluye una pluralidad de secuencias asociadas a diferentes representaciones, con diferente tasa de bits y/o calidad, de un mismo contenido de audio, en donde el conjunto de adaptación seleccionado se selecciona en base a de la solicitud del cliente.

De acuerdo con un cuadragésimo cuarto aspecto, se proporciona el sistema según cualquiera de los aspectos uno a cuarenta y uno, operando como un cliente y el servidor.

De acuerdo con un aspecto cuadragésimo quinto, el sistema del aspecto cuadragésimo cuarto, incluyendo el servidor de los aspectos cuadragésimo segundo o cuadragésimo tercero.

De acuerdo con un aspecto cuadragésimo sexto, se proporciona un método para una realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR o vídeo de 360 grados configurado para recibir transmisiones de vídeo y/o audio que se reproducirán en un dispositivo de consumo de medios que comprende: decodificar señales de vídeo de transmisiones de vídeo para la representación de escenas de entornos de vídeo VR, AR, MR o 360 grados para un usuario, y decodificar señales de audio de transmisiones de audio, solicitando y/u obteniendo de un servidor al menos una transmisión de audio en base a la ventana gráfica actual del usuario y/o datos posicionales y/u orientación de la cabeza y/o datos de movimiento y/o metadatos y/o datos virtuales y/o metadatos de posición.

De acuerdo con un cuadragésimo séptimo aspecto, se proporciona un programa informático que comprende instrucciones que, cuando son ejecutadas por un procesador, hacen que el procesador realice el método del aspecto cuadragésimo sexto.

REIVINDICACIONES

1. Un sistema (102) para un entorno de realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR, o vídeo de 360 grados configurado para recibir transmisiones de vídeo y audio que se reproducirán en un dispositivo de consumo de medios,

en donde el sistema (102) comprende:

al menos un decodificador de vídeo multimedia configurado para decodificar señales de vídeo de transmisiones de vídeo (1800) para la representación de entornos de RV, RA, RM o vídeo de 360 grados a un usuario, y

al menos un decodificador de audio (104) configurado para decodificar señales de audio (108) de al menos una transmisión de audio (106) para la representación de una escena de audio,

en donde el sistema (102) está configurado para solicitar (112) al menos una transmisión de audio (106) desde un servidor (120) sobre la base de al menos la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos posicionales virtuales (110), y

en donde el sistema (102) está configurado para solicitar el al menos una transmisión de audio al servidor (120) en función de una distancia de la posición del usuario con respecto a los límites de entornos de vídeo vecinos y/o adyacentes asociados a diferentes escenas de audio,

en donde se define una pluralidad de escenas de audio (150A, 150B) para múltiples entornos de vídeo, tales como entornos de vídeo adyacentes y/o vecinos,

de modo que se proporcionen las primeras transmisiones asociadas a una primera escena de audio actual y, en caso de que el usuario pase a una segunda escena de audio posterior, se proporcionen tanto las transmisiones de audio asociadas a la primera escena de audio como las segundas transmisiones de audio asociadas a la segunda escena de audio,

en donde las primeras transmisiones asociadas a la primera escena de audio se obtienen a una tasa de bits más alta cuando el usuario se encuentra en el primer entorno asociado a la primera escena de audio, mientras que las segundas transmisiones asociadas a la segunda escena de audio asociada al segundo entorno se obtienen a una tasa de bits más baja cuando el usuario se encuentra al principio de una posición de transición desde la primera escena de audio a la segunda escena de audio, y

las primeras transmisiones asociadas a la primera escena de audio se obtienen a una tasa de bits más baja y las segundas transmisiones asociadas a la segunda escena de audio se obtienen a una tasa de bits más alta cuando el usuario se encuentra al final de una posición de transición de la primera escena de audio a la segunda escena de audio,

en donde la tasa de bits más baja es menor que la tasa de bits más alta.

2. El sistema de la reivindicación 1, configurado para proporcionar al servidor (120) la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos posicionales virtuales (110) para obtener la al menos una transmisión de audio (106) y/o un elemento de audio de una transmisión de audio del servidor (120).

3. El sistema de la reivindicación 1, en donde al menos una escena de audio está asociada a al menos un elemento de audio (152), cada elemento de audio está asociado a una posición y/o área en el entorno de vídeo donde el elemento de audio es audible, de modo que se proporcionan diferentes transmisiones de audio para diferentes posiciones del usuario y/o ventanas gráficas y/u orientaciones de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos virtuales de posición en la escena de audio.

4. El sistema de la reivindicación 1, configurado para decidir si al menos un elemento de audio de una transmisión de audio debe reproducirse para la ventana gráfica y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o la posición virtual del usuario actual en una escena de audio, y

en donde el sistema está configurado para solicitar y recibir el al menos un elemento de audio en la posición virtual del usuario actual.

5. El sistema de la reivindicación 1, configurado para decidir predictivamente si al menos un elemento de audio

(152) de una transmisión de audio será relevante y/o audible basándose en al menos la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos posicionales virtuales (110),

en donde el sistema está configurado para solicitar y recibir el al menos un elemento de audio y/o transmisión de audio en la posición virtual de un usuario concreto antes del movimiento y/o interacción previstos del usuario en una escena de audio, y

en donde el sistema está configurado para reproducir el al menos un elemento de audio y/o transmisión de audio, cuando se recibe, en la posición virtual del usuario particular después del movimiento y/o interacción del usuario en la escena de audio.

6. El sistema de la reivindicación 1, configurado para solicitar y/o recibir el al menos un elemento de audio (152) a una tasa de bits inferior, en la posición virtual del usuario antes de una interacción del usuario, siendo la interacción el resultado de un cambio de datos posicionales en la misma escena de audio (150) o de la entrada en una escena siguiente separada de la escena actual,

en donde el sistema está configurado para solicitar y recibir el al menos un elemento de audio a una tasa de bits más alta, en la posición virtual del usuario después de la interacción del usuario en una escena de audio.

7. El sistema de la reivindicación 1, en donde al menos un elemento de audio (152) asociado a al menos una escena de audio está asociado a una posición y/o área en el entorno de vídeo asociado a una escena de audio,

en donde el sistema está configurado para solicitar y recibir transmisiones a una tasa de bits más alta para elementos de audio más cercanos al usuario que para elementos de audio más distantes del usuario.

8. El sistema de la reivindicación 1, en donde al menos un elemento de audio (152) está asociado a al menos una escena de audio, estando el último elemento de audio asociado a una posición y/o área en el entorno de vídeo asociado a una escena de audio,

en donde el sistema está configurado para solicitar diferentes transmisiones a diferentes tasas de bits para elementos de audio basados en su relevancia y/o nivel de audibilidad en la posición virtual de cada usuario en una escena de audio,

en donde el sistema está configurado para solicitar una transmisión de audio a una tasa de bits más alta para elementos de audio que son más relevantes y/o más audibles en la posición virtual del usuario actual, y/o

una transmisión de audio a menor velocidad de bits para los elementos de audio que son menos relevantes y/o menos audibles en la posición virtual del usuario actual.

9. El sistema de la reivindicación 1, en donde al menos un elemento de audio (152) está asociado a una escena de audio, estando cada elemento de audio asociado a una posición y/o área en el entorno de vídeo asociado a una escena de audio,

en donde el sistema está configurado para enviar periódicamente al servidor la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos posicionales virtuales (110), de modo que:

para una primera posición, se proporciona una transmisión a mayor velocidad de bits, desde el servidor, y

para una segunda posición, se proporciona una transmisión con una tasa de bits inferior, desde el servidor,

en donde la primera posición está más cerca del al menos un elemento de audio (152) que la segunda posición.

10. El sistema de la reivindicación 1, en donde una pluralidad de escenas de audio (150A, 150B) se definen para un primer y un segundo entorno de vídeo, siendo el primer y el segundo entorno de vídeo entornos de vídeo adyacentes y/o vecinos,

en el que las primeras transmisiones asociadas a la primera escena de audio se proporcionan, desde el servidor, para la reproducción de la primera escena de audio en caso de que la posición del usuario o la posición virtual se encuentren en un primer entorno de vídeo asociado a la primera escena de audio,

se proporcionan segundas transmisiones asociadas a la segunda escena de audio, desde el servidor, para la reproducción de la segunda escena de audio en caso de que la posición o posición virtual del usuario se encuentre en un segundo entorno de vídeo asociado a la segunda escena de audio, y

se proporcionan tanto primeras transmisiones asociadas a la primera escena de audio como segundas transmisiones asociadas a la segunda escena de audio en caso de que la posición del usuario o la posición virtual se encuentre en una posición de transición entre la primera escena de audio y la segunda escena de audio.

11. El sistema de la reivindicación 1, en donde una pluralidad de escenas de audio (150A, 150B) se definen para un primer y un segundo entorno de vídeo, que son entornos adyacentes y/o vecinos,

en donde el sistema está configurado para solicitar y recibir primeras transmisiones asociadas a una primera escena de audio asociada (150A) al primer entorno, para la reproducción de la primera escena de audio en caso de que la posición virtual del usuario se encuentre en el primer entorno,

en donde el sistema está configurado para solicitar y recibir segundas transmisiones asociadas a la segunda escena de audio (150B) asociada al segundo entorno, para la reproducción de la segunda escena de audio en caso de que la posición virtual del usuario se encuentre en el segundo entorno, y

en donde el sistema está configurado para solicitar y recibir tanto primeras transmisiones asociadas a la primera escena de audio como segundas transmisiones asociadas a la segunda escena de audio en caso de que la posición virtual del usuario se encuentre en una posición de transición (150AB) entre el primer entorno y el segundo entorno.

12. El sistema de la reivindicación 1, en donde una pluralidad de escenas de audio (150A, 150B) se define para múltiples entornos tales como entornos adyacentes y/o vecinos,

de modo que el sistema esté configurado para obtener las transmisiones de audio asociadas a una primera escena de audio actual asociada a un primer entorno actual, y,

en caso de que la distancia de la posición o posición virtual del usuario respecto a un límite de una escena de audio sea inferior a un umbral predeterminado, el sistema obtiene además transmisiones de audio asociadas a un segundo entorno, adyacente y/o vecino, asociado a la segunda escena de audio.

13. El sistema de la reivindicación 1, en donde una pluralidad de escenas de audio (150A, 150B) se define para múltiples entornos de vídeo,

para que el sistema solicite y obtenga las transmisiones de audio asociadas a una escena de audio actual a una velocidad de bits superior y las transmisiones de audio asociadas a la segunda escena de audio a una velocidad de bits inferior.

14. El sistema de la reivindicación 1, en donde al menos un entorno de vídeo está asociado a al menos una pluralidad de N elementos de audio, estando cada elemento de audio asociado a una posición y/o área en el entorno de vídeo,

en donde la al menos una pluralidad de N elementos de audio se proporciona en al menos una representación a alta tasa de bits, y

en donde la al menos una pluralidad de N elementos de audio se proporciona en al menos una representación a baja tasa de bits, en donde la al menos una representación se obtiene procesando los N elementos de audio para obtener un número menor M de elementos de audio asociados a una posición o área cercana a la posición o área de los N elementos de audio,

en donde el sistema está configurado para solicitar la representación a una tasa de bits más alta para los elementos de audio, en caso de que los elementos de audio sean más relevantes y/o más audibles en la posición virtual del usuario actual en una escena de audio,

en donde el sistema está configurado para solicitar la representación a una tasa de bits inferior para los elementos de audio, en caso de que los elementos de audio sean menos relevantes y/o menos audibles en la posición virtual del usuario actual en una escena de audio.

15. El sistema de la reivindicación 14, en donde se obtienen diferentes transmisiones de audio para los diferentes elementos de audio.

16. El sistema de la reivindicación 1, configurada para solicitar y/u obtener las transmisiones en función de la orientación del usuario y/o la dirección del movimiento del usuario y/o las interacciones del usuario en la escena.

17. El sistema de cualquiera de las reivindicaciones anteriores, en donde la ventana gráfica está asociada a la posición y/o posición virtual y/o datos de movimiento y/o cabeza.

18. El sistema de la reivindicación 1 o 17, en donde diferentes elementos de audio se proporcionan en diferentes ventanas gráficas, en donde el sistema está configurado para solicitar y recibir, en caso de que un primer elemento de audio (S1) caiga dentro de una ventana gráfica (160-1), el primer elemento de audio a una tasa de bits más alta que un segundo elemento de audio (S2) que no cae dentro de la ventana gráfica.

19. El sistema (102) de la reivindicación 1,

en donde al menos un primer y segundo elementos de audio (152A, 152B) están asociados a una primera escena de audio asociada a un primer entorno de vídeo, y al menos un tercer elemento de audio (152C) está asociado a una segunda escena de audio asociada a un segundo entorno de vídeo,

en donde el sistema está configurado para obtener metadatos de interacción que describen que el al menos un segundo elemento de audio (152B) está asociado adicionalmente con el segundo entorno de vídeo,

en donde el sistema está configurado para solicitar y recibir los al menos un primer y segundo elementos de audio (152A, 152B), en caso de que la posición virtual del usuario se encuentre en el primer entorno de vídeo,

en donde el sistema está configurado para solicitar y recibir los al menos un segundo y tercer elementos de audio (152B, 152C), en caso de que la posición virtual del usuario se encuentre en el segundo entorno de vídeo, y

en donde el sistema está configurado para solicitar y recibir los al menos un primer y segundo y tercer elementos de audio (152A, 152B, 152C), en caso de que la posición virtual del usuario se encuentre en transición entre el primer entorno de vídeo y el segundo entorno de vídeo,

en donde el al menos un primer elemento de audio (152) se proporciona en al menos una transmisión de audio (A1, 106A), y el al menos un segundo elemento de audio (152B) se proporciona en al menos una segunda transmisión de audio (A2, 106B), y el al menos un tercer elemento de audio (152C) se proporciona en al menos una tercera transmisión de audio (B1, 10C), y en el que al menos un primer entorno de vídeo se describe mediante metadatos de interacción como una escena de audio que requiere al menos una primera y segunda transmisiones de audio (A1, A2, 106A, 106B), y en donde el segundo entorno de vídeo se describe mediante metadatos de interacción como una escena de audio que requiere al menos una tercera transmisión de audio (B1, 106C) y al menos una segunda transmisión de audio (A2, 152B) asociada con al menos un primer entorno de vídeo,

en donde el sistema comprende un procesador de metadatos (1236) configurado para manipular los metadatos de interacción, para fusionar la segunda transmisión de audio (A2, 152B) perteneciente al primer entorno de vídeo y la tercera transmisión de audio (B1, 152C) asociada al segundo entorno de vídeo en una nueva transmisión única, en caso de que la posición virtual del usuario se encuentre en el segundo entorno de vídeo.

20. El sistema de la reivindicación 1, en donde el sistema comprende un procesador de metadatos (1236) configurado para manipular metadatos en al menos una transmisión de audio antes del al menos un decodificador de audio (104), basándose en la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos posicionales virtuales.

21. El sistema de la reivindicación 20, en donde el procesador de metadatos (1236) está configurado para habilitar y/o deshabilitar al menos un elemento de audio (152A-152C) en al menos una transmisión de audio (106A-106C) antes del al menos un decodificador de audio (104), basándose en la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos de interacción y/o los datos posicionales virtuales, en donde

el procesador de metadatos (1236) está configurado para desactivar al menos un elemento de audio (152A-152C) en al menos una transmisión de audio (106A-106C) antes del al menos un decodificador de audio (104), en caso de que el sistema decida que el elemento de audio (152A-152C) ya no se va a reproducir como consecuencia de una ventana gráfica actual y/o de la orientación de la cabeza y/o de datos de

movimiento y/o de metadatos de interacción y/o de datos posicionales virtuales, y en donde

el procesador de metadatos (1236) está configurado para habilitar al menos un elemento de audio (152A-152C) en al menos una transmisión de audio antes del al menos un decodificador de audio, en caso de que el sistema decida que el elemento de audio (152A-152C) debe reproducirse como consecuencia de la ventana gráfica actual de un usuario y/o la orientación de la cabeza y/o datos de movimiento y/o metadatos de interacción y/o datos posicionales virtuales.

22. El sistema de la reivindicación 1 o 19, configurado para desactivar la decodificación de los elementos de audio (152A-152C) en función de la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o la posición virtual.

23. El sistema de la reivindicación 1, configurado para fusionar al menos una primera transmisión de audio (106A) asociada a una escena de audio actual con al menos una transmisión (106C) asociada a una escena de audio vecina, adyacente y/o futura.

24. El sistema de la reivindicación 1 o 19, configurado para obtener y/o recopilar datos estadísticos o agregados en la ventana gráfica actual y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o datos virtuales de posición, para transmitir la solicitud al servidor (120) asociado a los datos estadísticos o agregados.

25. El sistema de la reivindicación 1 o 19, configurado para desactivar la decodificación y/o la reproducción de al menos una transmisión sobre la base de metadatos asociados a al menos una transmisión y sobre la base de la ventana gráfica actual y/o la orientación de la cabeza del usuario y/o datos de movimiento y/o metadatos y/o datos posicionales virtuales.

26. El sistema de la reivindicación 1 o 19, configurado además para:

manipular los metadatos asociados a un grupo de secuencias de audio seleccionadas (106A-106C), basándose en al menos el punto de vista actual o estimado del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o los datos posicionales virtuales, con el fin de:

seleccionar y/o activar los elementos de audio (152A-152C) que componen una escena de audio a reproducir; y/o

fusionar todas las transmisiones de audio seleccionadas en una única transmisión de audio.

27. El sistema de la reivindicación 1 o 19, en donde se proporciona información desde el servidor (120), para cada elemento de audio (152A-152C) u objeto de audio, en donde la información incluye información descriptiva sobre las ubicaciones en las que una escena de audio o los elementos de audio están activos.

28. El sistema de la reivindicación 1 o 19, configurado para elegir entre reproducir una escena y componer o mezclar o multiplexar o superponer o combinar al menos dos escenas de audio sobre la base de la orientación actual o futura o de la ventana gráfica y/o la orientación y/o movimiento, y/o los metadatos y/o la posición virtual y/o la selección de un usuario, las dos escenas de audio están asociadas a diferentes entornos vecinos y/o adyacentes.

29. El sistema según cualquiera de las reivindicaciones anteriores, configurado para:

recibir una transmisión para una escena de audio asociada a un entorno vecino o adyacente;

iniciar la decodificación y reproducción de la transmisión de audio para el entorno vecino o adyacente en la detección de la transición de un límite entre dos entornos.

30. El sistema de la reivindicación 1 o 19, configurado para

decidir si se van a reproducir al menos un elemento de audio (152) de al menos una escena de audio incrustada en al menos una transmisión de audio (152A) y al menos un elemento de audio adicional (152B) de al menos una escena de audio adicional incrustada en al menos una transmisión de audio adicional (106B); y

provocar, en caso de decisión positiva, una operación de fusión o composición o multiplexar o superposición o combinación de al menos una transmisión adicional (106B) de la escena de audio adicional a la al menos una transmisión (106A) de la al menos una escena de audio.

31. El sistema de la reivindicación 1 o 19, configurado para

manipular los metadatos de audio asociados a las transmisiones de audio seleccionadas, basándose al menos en la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o los datos de movimiento y/o los metadatos y/o los datos posicionales virtuales, con el fin de:

seleccionar y/o habilitar y/o activar los elementos de audio que componen una escena de audio decidida a reproducir; y

habilitar la fusión de todas las transmisiones de audio seleccionadas en una única transmisión de audio.

32. Un método para un entorno de realidad virtual, VR, realidad aumentada, AR, realidad mixta, MR, o vídeo de 360 grados configurado para recibir transmisiones de vídeo y / audio que se reproducirán en un dispositivo de consumo de medios, que comprende:

decodificación de señales de vídeo a partir de transmisiones de vídeo para la representación de entornos de RV, RA, RM o vídeo de 360 grados a un usuario,

decodificación de señales de audio a partir de transmisiones de audio para la representación de escenas de audio,

solicitar y/u obtener de un servidor (120) al menos una transmisión de audio sobre la base de la ventana gráfica actual y/o los datos posicionales y/o los datos de movimiento y/u orientación de la cabeza y/o metadatos y/o datos virtuales posicionales y/o metadatos, y

solicitar la al menos una transmisión al servidor (120) en función de una distancia de la posición del usuario a los límites de entornos de vídeo vecinos y/o adyacentes asociados a diferentes escenas de audio,

en donde se define una pluralidad de escenas de audio (150A, 150B) para múltiples entornos de vídeo, tales como entornos de vídeo adyacentes y/o vecinos,

de modo que se proporcionen las primeras transmisiones asociadas a una primera escena de audio actual y, en caso de que el usuario pase a una segunda escena de audio posterior, se proporcionen tanto las transmisiones de audio asociadas a la primera escena de audio como las segundas transmisiones de audio asociadas a la segunda escena de audio,

en donde las primeras transmisiones asociadas a la primera escena de audio se obtienen a una tasa de bits más alta cuando el usuario se encuentra en el primer entorno asociado a la primera escena de audio, mientras que las segundas transmisiones asociadas a la segunda escena de audio asociada al segundo entorno se obtienen a una tasa de bits más baja cuando el usuario se encuentra al principio de una posición de transición desde la primera escena de audio a la segunda escena de audio, y

las primeras transmisiones asociadas a la primera escena de audio se obtienen a una tasa de bits más baja y las segundas transmisiones asociadas a la segunda escena de audio se obtienen a una tasa de bits más alta cuando el usuario se encuentra al final de una posición de transición de la primera escena de audio a la segunda escena de audio,

en donde la tasa de bits más baja es menor que la tasa de bits más alta.

33. Un programa informático que comprende instrucciones que, cuando son ejecutadas por un procesador, hacen que el procesador realice el método de la reivindicación 32.

34. El sistema de la reivindicación 1 configurada para:

solicitar y recibir al menos un primer conjunto de adaptación que comprenda la primera transmisión de audio (106A) asociada a la primera escena de audio;

solicitar y recibir al menos un segundo conjunto de adaptación que comprenda la segunda transmisión de audio (106B) asociada con al menos dos escenas de audio, incluida la primera escena de audio; y

fusionar la primera transmisión de audio (106A) y la segunda transmisión de audio (106B) en una nueva transmisión de audio a decodificar, basándose en los metadatos disponibles relativos a la ventana gráfica actual del usuario y/o la orientación de la cabeza y/o datos de movimiento y/o metadatos y/o datos

posicionales virtuales y/o información que describa una asociación del al menos un primer conjunto de adaptación a la al menos una primera escena de audio y/o una asociación del al menos un segundo conjunto de adaptación a la al menos una primera escena de audio.

5

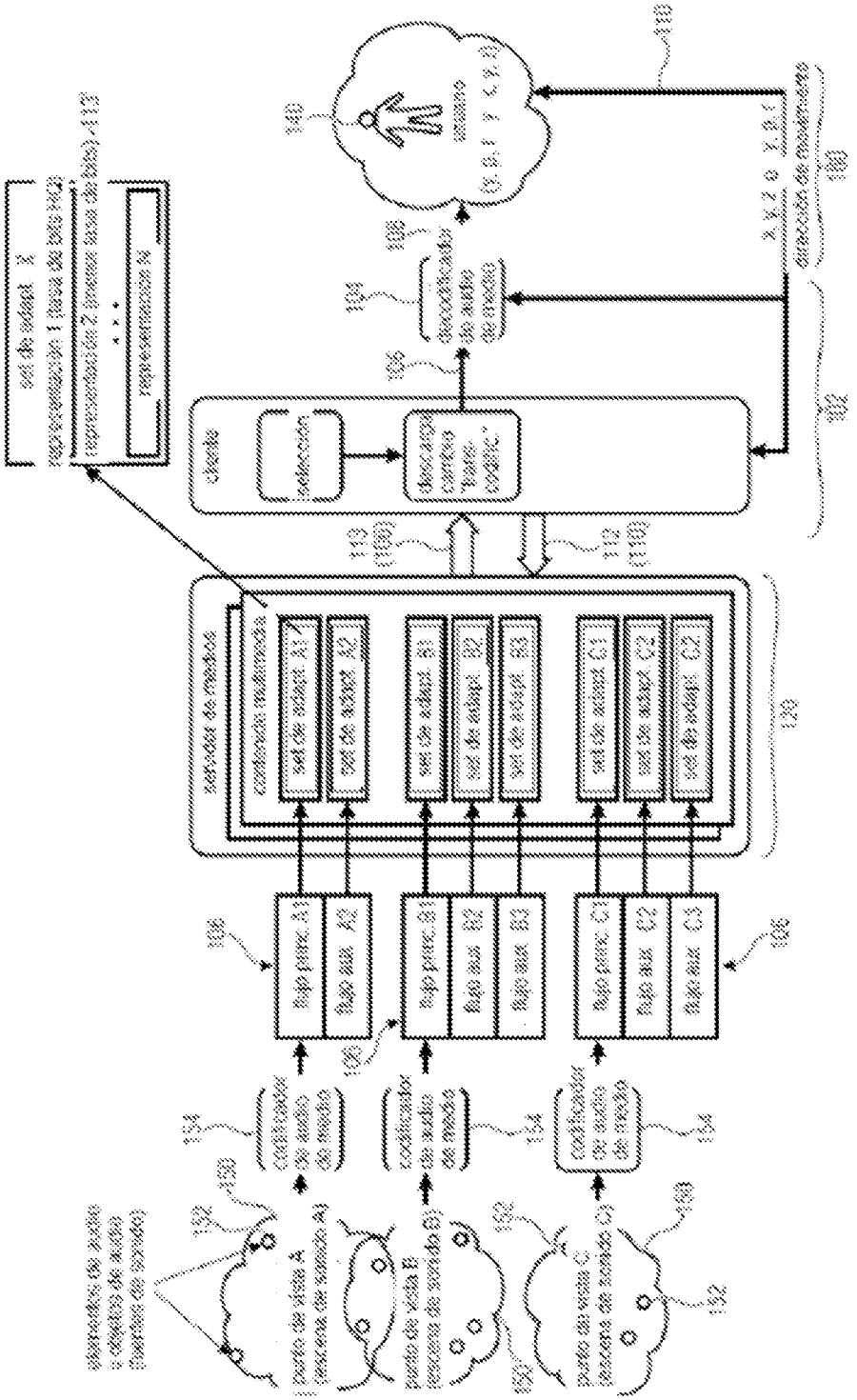

Fig. 1.1

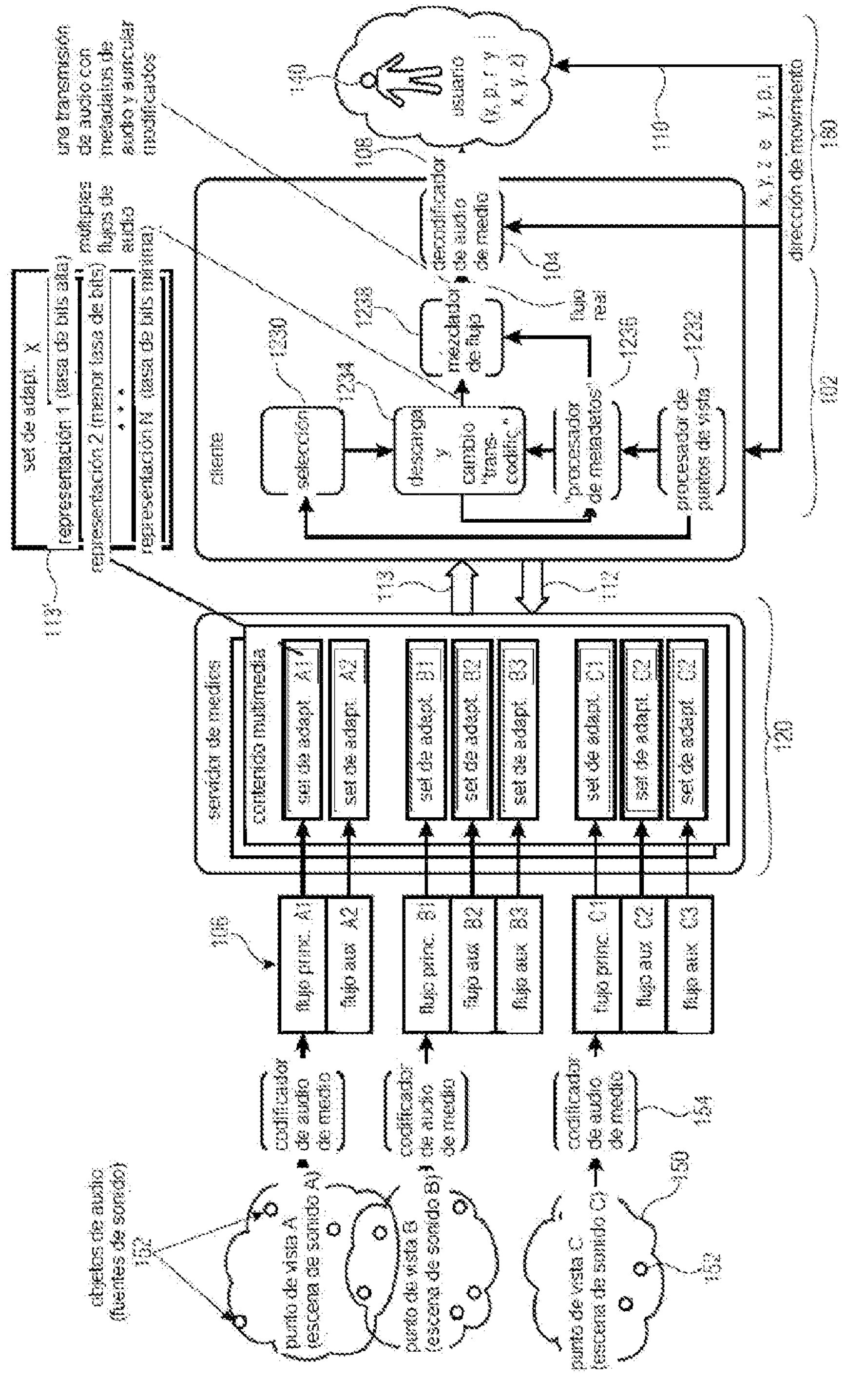
objetos de audio (fuentes de sonido)
152
punto de vista A (escena de sonido A)
punto de vista B (escena de sonido B)
punto de vista C (escena de sonido C)
150
152
codificador de audio de medio
codificador de audio de medio
codificador de audio de medio
154
106
flujo princ. A1
flujo aux A2
flujo princ. B1
flujo aux B2
flujo aux B3
flujo princ. C1
flujo aux C2
flujo aux C3
servidor de medios
contenido multimedia
set de adapt. A1
set de adapt. A2
set de adapt. B1
set de adapt. B2
set de adapt. B3
set de adapt. C1
set de adapt. C2
set de adapt. C3
120
113
112
113'
set de adapt. X
representación 1 (tasa de bits alta)
representación 2 (menor tasa de bits)
representación N (tasa de bits mínima)
múltiples flujos de audio
una transmisión de audio con metadatos de audio y auricular modificados
cliente
selección
1230
descarga y cambio "trans-codific."
1234
"procesador de metadatos"
1236
procesador de puntos de vista
1232
mezclador de flujo
1238
flujo real
decodificador de audio de medio
104
108
102
140
usuario
110
x, y, z e y, θ, ı
dirección de movimiento
180

Fig. 1.2

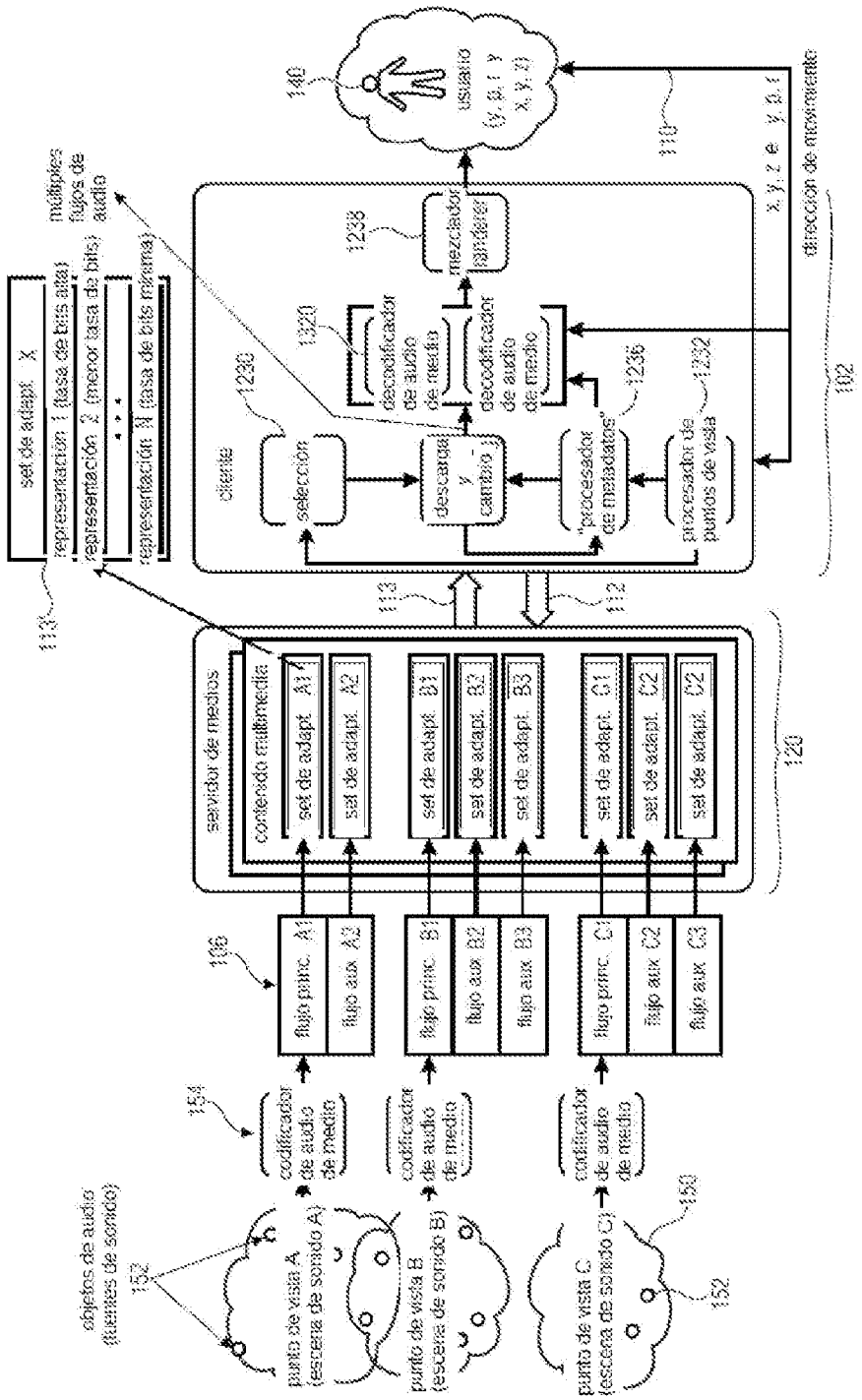
objetos de audio (fuentes de sonido)
152
punto de vista A (escena de sonido A)
punto de vista B (escena de sonido B)
punto de vista C (escena de sonido C)
154
codificador de audio de medio
codificador de audio de medio
codificador de audio de medio
156
flujo princ. A1
flujo aux A2
flujo princ. B1
flujo aux B2
flujo aux B3
flujo princ. C1
flujo aux C2
flujo aux C3
120
servidor de medios
contenido multimedia
set de adapt. A1
set de adapt. A2
set de adapt. B1
set de adapt. B2
set de adapt. B3
set de adapt. C1
set de adapt. C2
set de adapt. C3
113
112
cliente
1230
selección
descarga y cambio
"procesador de metadatos"
procesador de puntos de vista
1232
1236
1220
decodificador de audio de medio
decodificador de audio de medio
1238
mezclador / renderer
múltiples flujos de audio
set de adapt. X
representación 1 (tasa de bits alta)
representación 2 (menor tasa de bits)
representación N (tasa de bits mínima)
113
140
usuario (y, p, r) y (x, y, z)
110
x, y, z e y, p, r
dirección de movimiento
102

Fig. 1.3

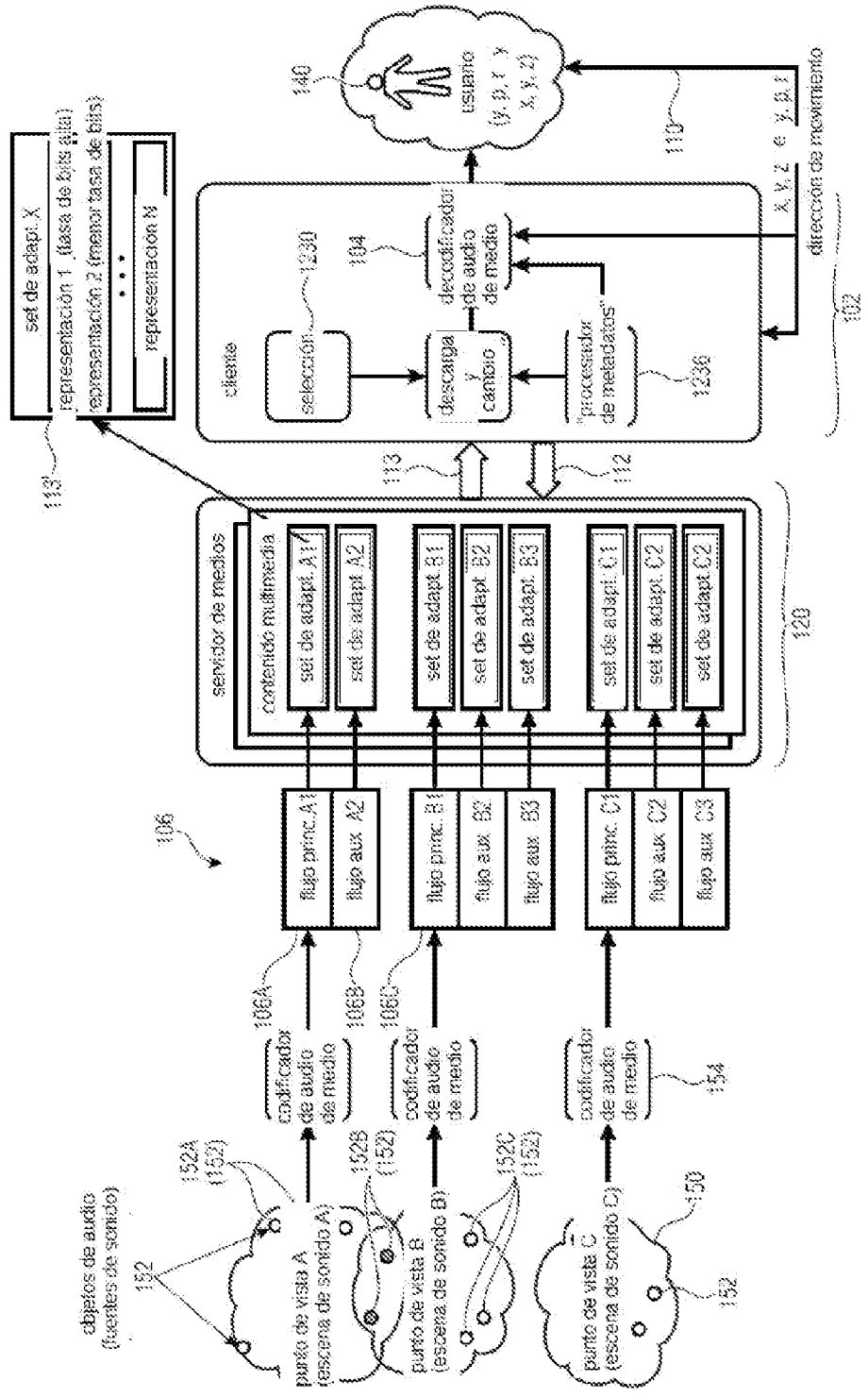
objetos de audio
(fuentes de sonido)
152
152A (152)
152B (152)
152C (152)
punto de vista A
(escena de sonido A)
punto de vista B
(escena de sonido B)
punto de vista C
(escena de sonido C)
150
152
codificador de audio de medio
154
106A
106B
106C
flujo princ. A1
flujo aux A2
flujo princ. B1
flujo aux B2
flujo aux B3
flujo princ. C1
flujo aux C2
flujo aux C3
106
servidor de medios
contenido multimedia
set de adapt. A1
set de adapt. A2
set de adapt. B1
set de adapt. B2
set de adapt. B3
set de adapt. C1
set de adapt. C2
set de adapt. C3
120
113
112
cliente
selección
1230
descarga y cambio
104
decodificador de audio de medio
"procesador de metadatos"
1236
140
usuario
(y, θ, r, x, y, z)
110
x, y, z e y, θ, r
dirección de movimiento
102
113'
set de adapt. X
representación 1 (tasa de bits alta)
representación 2 (menor tasa de bits)
representación N

Fig. 1.4

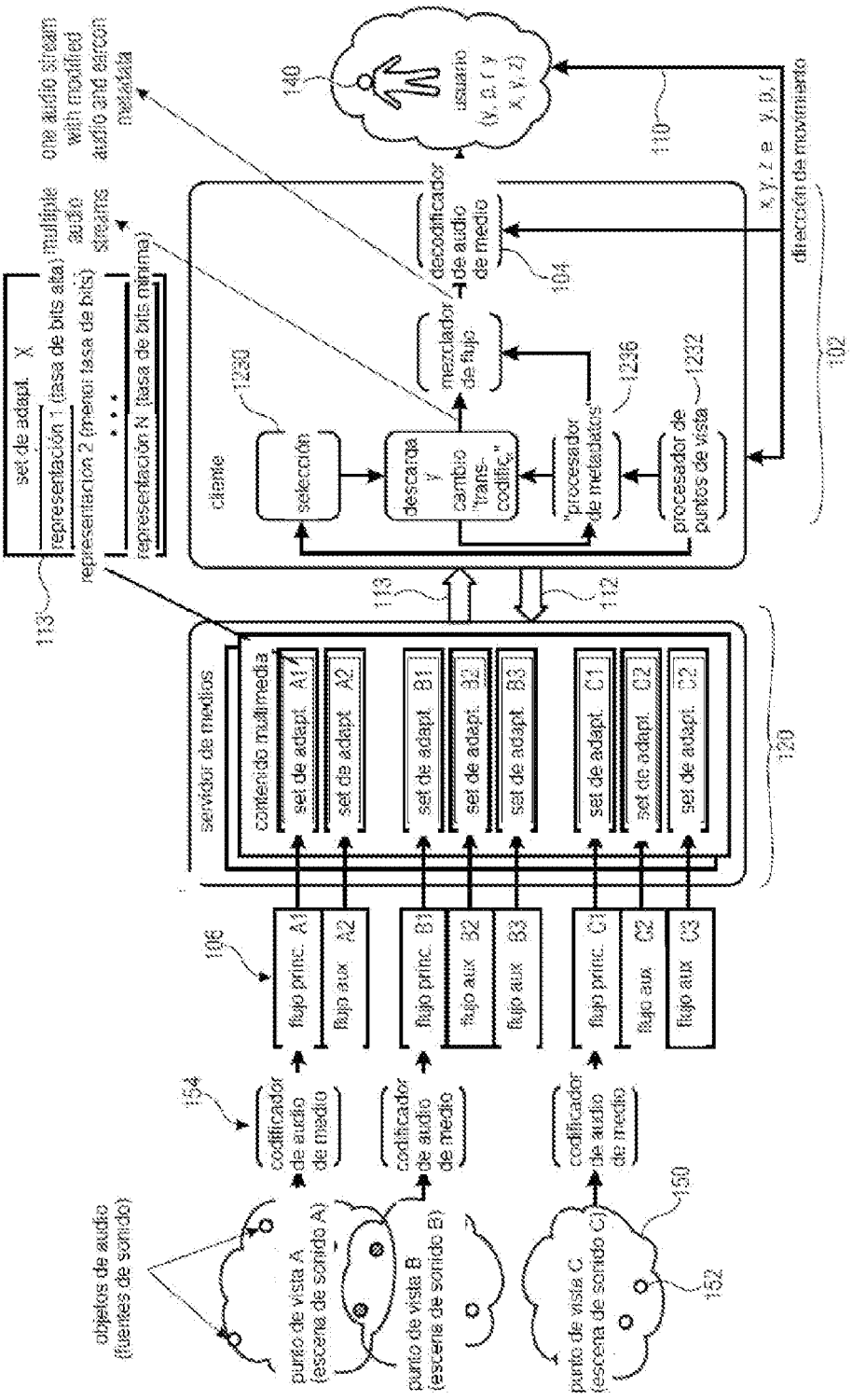
objetos de audio (fuentes de sonido)
punto de vista A (escena de sonido A)
punto de vista B (escena de sonido B)
punto de vista C (escena de sonido C)
150
152
codificador de audio de medio
154
flujo princ. A1
flujo aux A2
flujo princ. B1
flujo aux B2
flujo aux B3
flujo princ. C1
flujo aux C2
flujo aux C3
156
servidor de medios
contenido multimedia
set de adapt. A1
set de adapt. A2
set de adapt. B1
set de adapt. B2
set de adapt. B3
set de adapt. C1
set de adapt. C2
set de adapt. C3
120
113
112
cliente
1230
selección
descarga y cambio "trans-codific."
"procesador de metadatos"
1236
procesador de puntos de vista
1232
mezclador de flujo
decodificador de audio de medio
104
140
usuario (x, y, z y x, y, z)
110
x, y, z e y, p, r
dirección de movimiento
102
115
set de adapt. X
representación 1 (tasa de bits alta)
representación 2 (menor tasa de bits)
representación N (tasa de bits mínima)
multiple audio streams
one audio stream with modified audio and certain metadata

Fig. 1.5

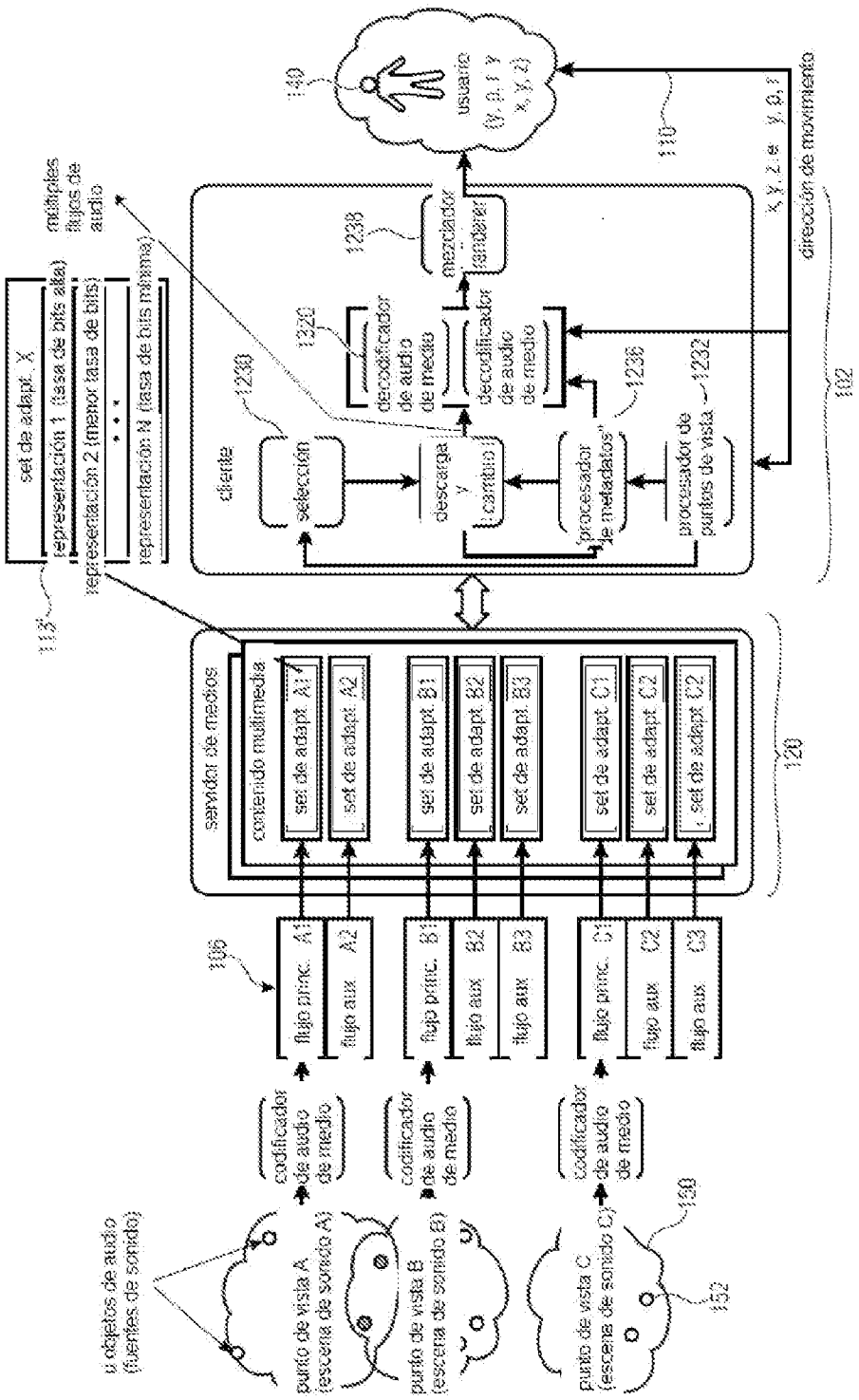
u objetos de audio (fuentes de sonido)
punto de vista A (escena de sonido A)
punto de vista B (escena de sonido B)
punto de vista C (escena de sonido C)
150
152
codificador de audio de medio
codificador de audio de medio
codificador de audio de medio
106
flujo princ. A1
flujo aux A2
flujo princ. B1
flujo aux B2
flujo aux B3
flujo princ. C1
flujo aux C2
flujo aux C3
servidor de medios
contenido multimedia
set de adapt. A1
set de adapt. A2
set de adapt. B1
set de adapt. B2
set de adapt. B3
set de adapt. C1
set de adapt. C2
set de adapt. C3
120
118
set de adapt. X
representación 1 (tasa de bits alta)
representación 2 (menor tasa de bits)
representación N (tasa de bits mínima)
múltiples flujos de audio
cliente
1230
selección
descarga y control
"procesador de metadatos"
procesador de puntos de vista
1236
1232
1220
decodificador de audio de medio
decodificador de audio de medio
1238
mezclador (renderer)
102
dirección de movimiento
x, y, z o α, β, γ
110
140
usuario
(y, p, r y x, y, z)

Fig. 1.6

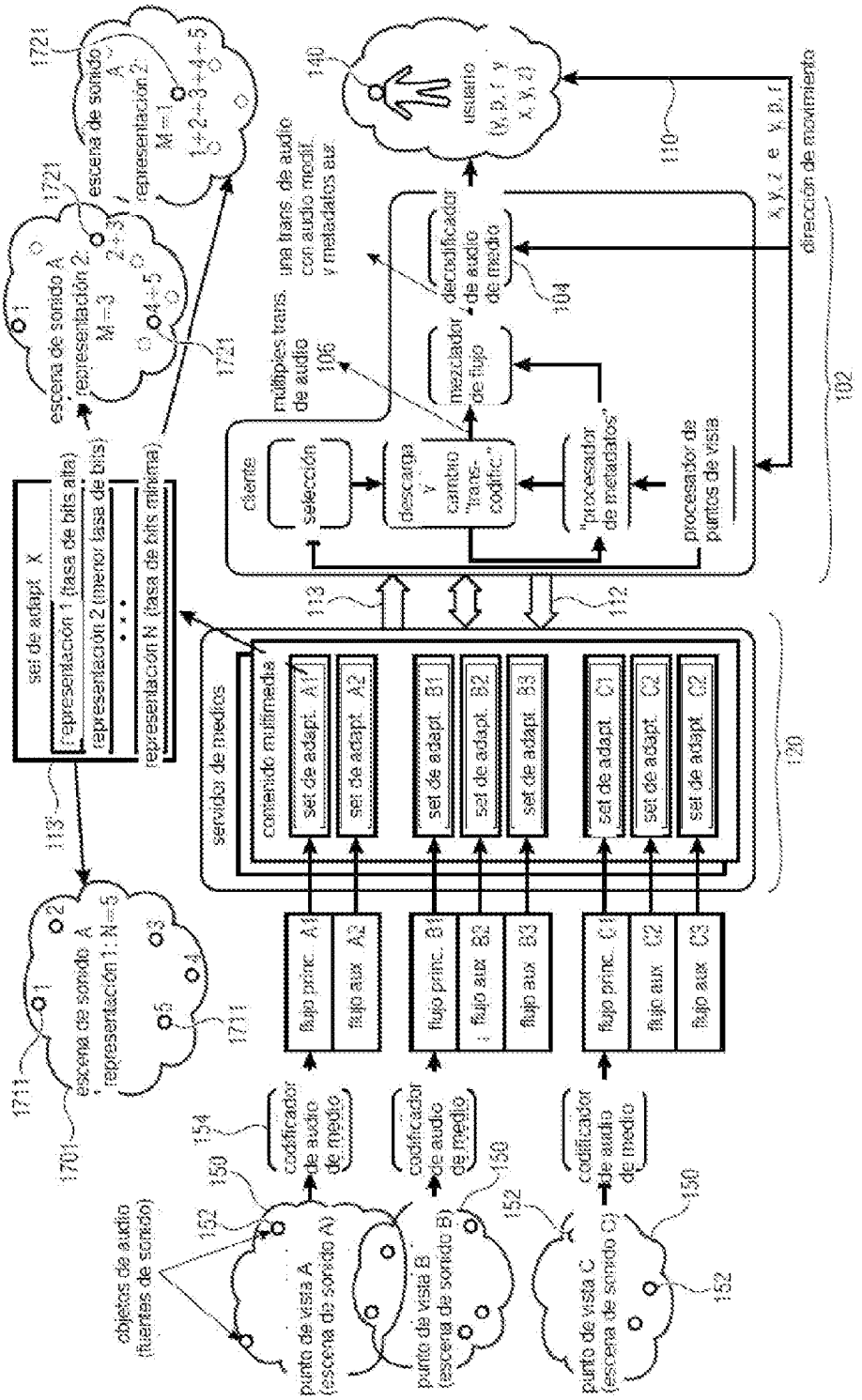
objetos de audio (fuentes de sonido)
punto de vista A (escena de sonido A)
punto de vista B (escena de sonido B)
punto de vista C (escena de sonido C)
codificador de audio de medio
flujo princ. A1
flujo aux A2
flujo princ. B1
flujo aux B2
flujo aux B3
flujo princ. C1
flujo aux C2
flujo aux C3
servidor de medios
contenido multimedia
set de adapt. A1
set de adapt. A2
set de adapt. B1
set de adapt. B2
set de adapt. B3
set de adapt. C1
set de adapt. C2
set de adapt. C3
cliente
selección
descarga y cambio "trans-codific."
"procesador de metadatos"
procesador de puntos de vista
múltiples trans. de audio de audio 106
una trans. de audio con audio modif. y metadatos aux
mezclador de flujo
decodificador de audio de medio
usuario
(x, y, z) (α, β, γ)
dirección de movimiento
X, Y, Z e α, β, γ
escena de sonido representación 1: M=5
escena de sonido representación 2: M=3
escena de sonido representación 2: M=1
set de adapt. X
representación 1 (tasa de bits alta)
representación 2 (menor tasa de bits)
representación N (tasa de bits mínima)

Fig. 1.7

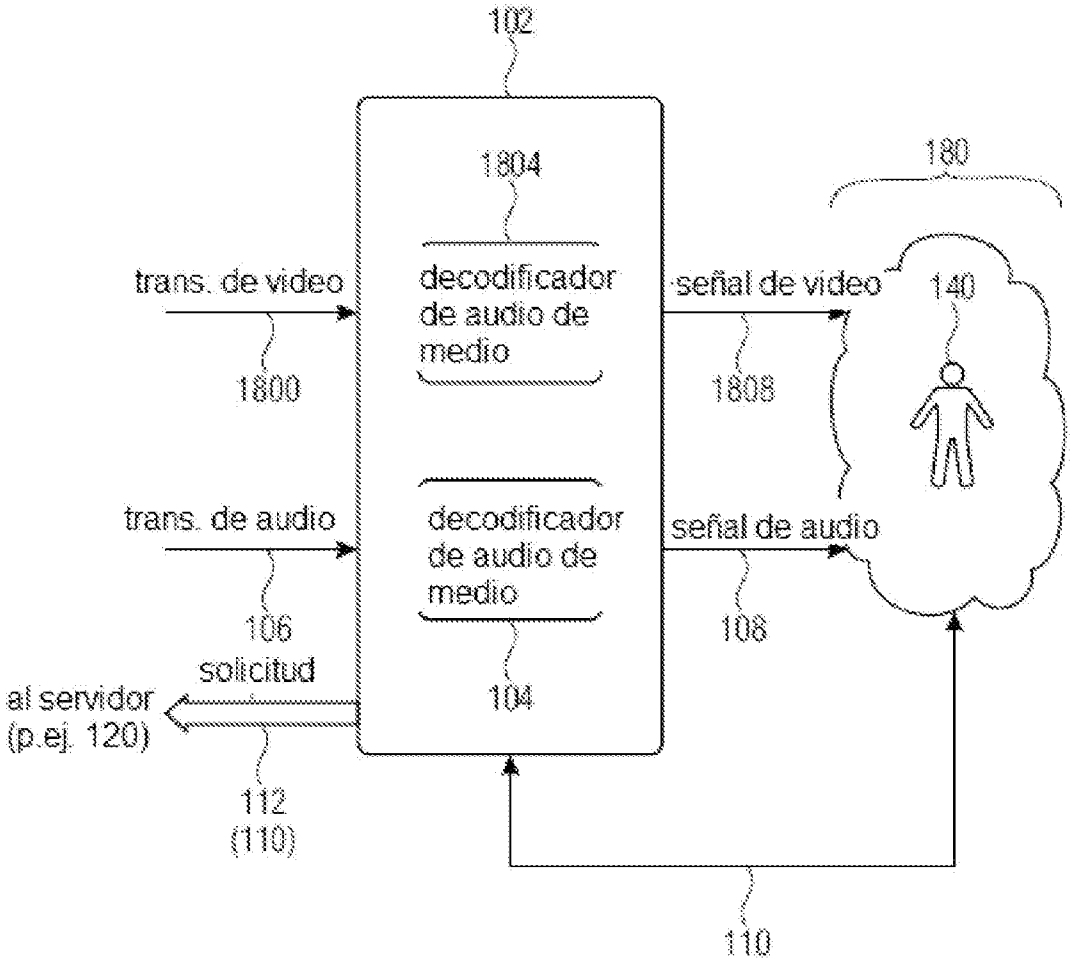
102
1804
decodificador de audio de medio
trans. de video
1800
señal de video
1808
180
140
trans. de audio
106
decodificador de audio de medio
señal de audio
108
solicitud
al servidor (p.ej. 120)
112 (110)
104
110

Fig. 1.8

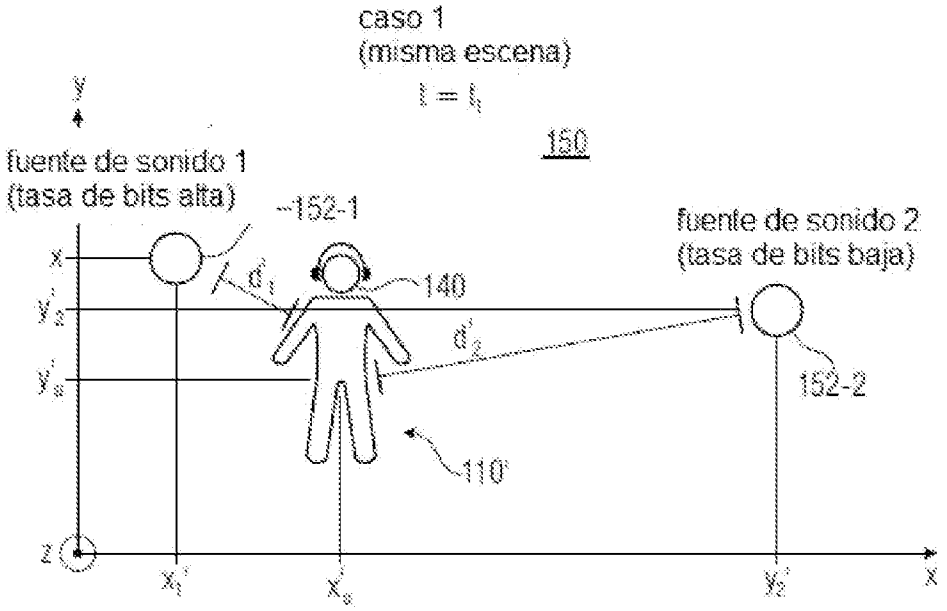
caso 1
(misma escena)
t = t1
150
y
fuente de sonido 1
(tasa de bits alta)
152-1
fuente de sonido 2
(tasa de bits baja)
x
140
152-2
110'
z
x

Fig. 2a

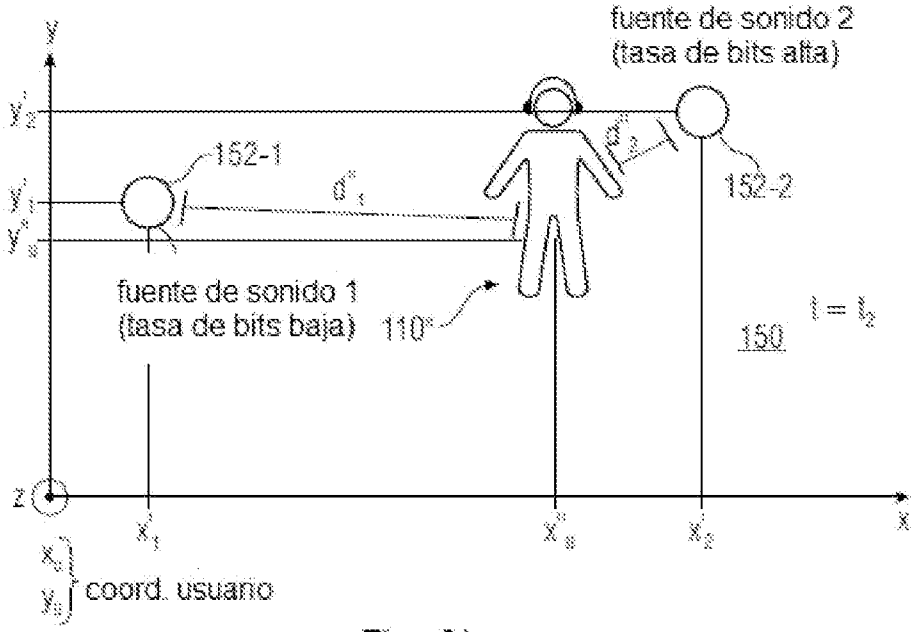
y
fuente de sonido 2
(tasa de bits alta)
152-1
152-2
fuente de sonido 1
(tasa de bits baja)
110"
150
t = t2
z
x
coord. usuario

Fig. 2b

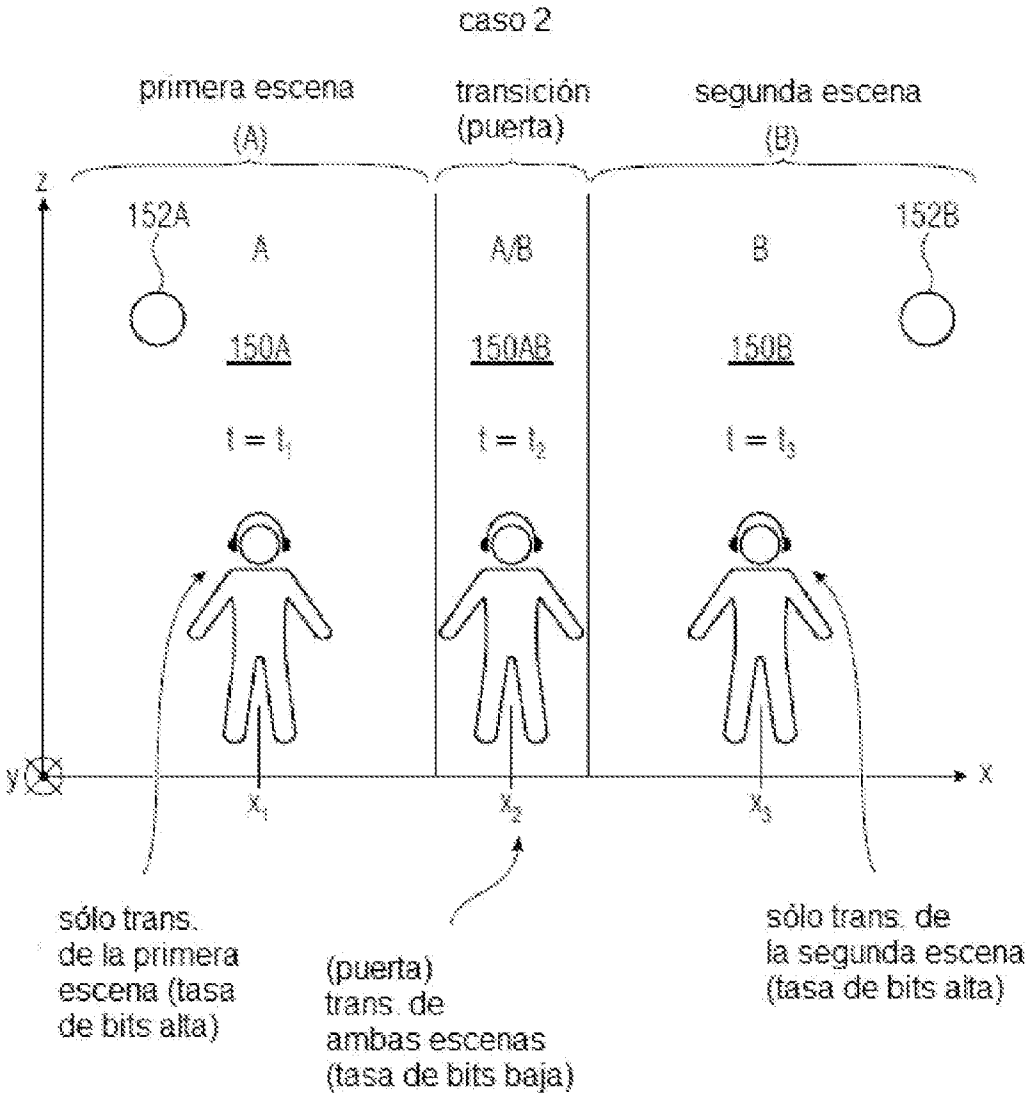
caso 2
primera escena
(A)
transición
(puerta)
segunda escena
(B)
z
152A
A
A/B
B
152B
150A
150AB
150B
t = t1
t = t2
t = t3
y
x
x1
x2
x3
sólo trans.
de la primera
escena (tasa
de bits alta)
(puerta)
trans. de
ambas escenas
(tasa de bits baja)
sólo trans. de
la segunda escena
(tasa de bits alta)

Fig. 3

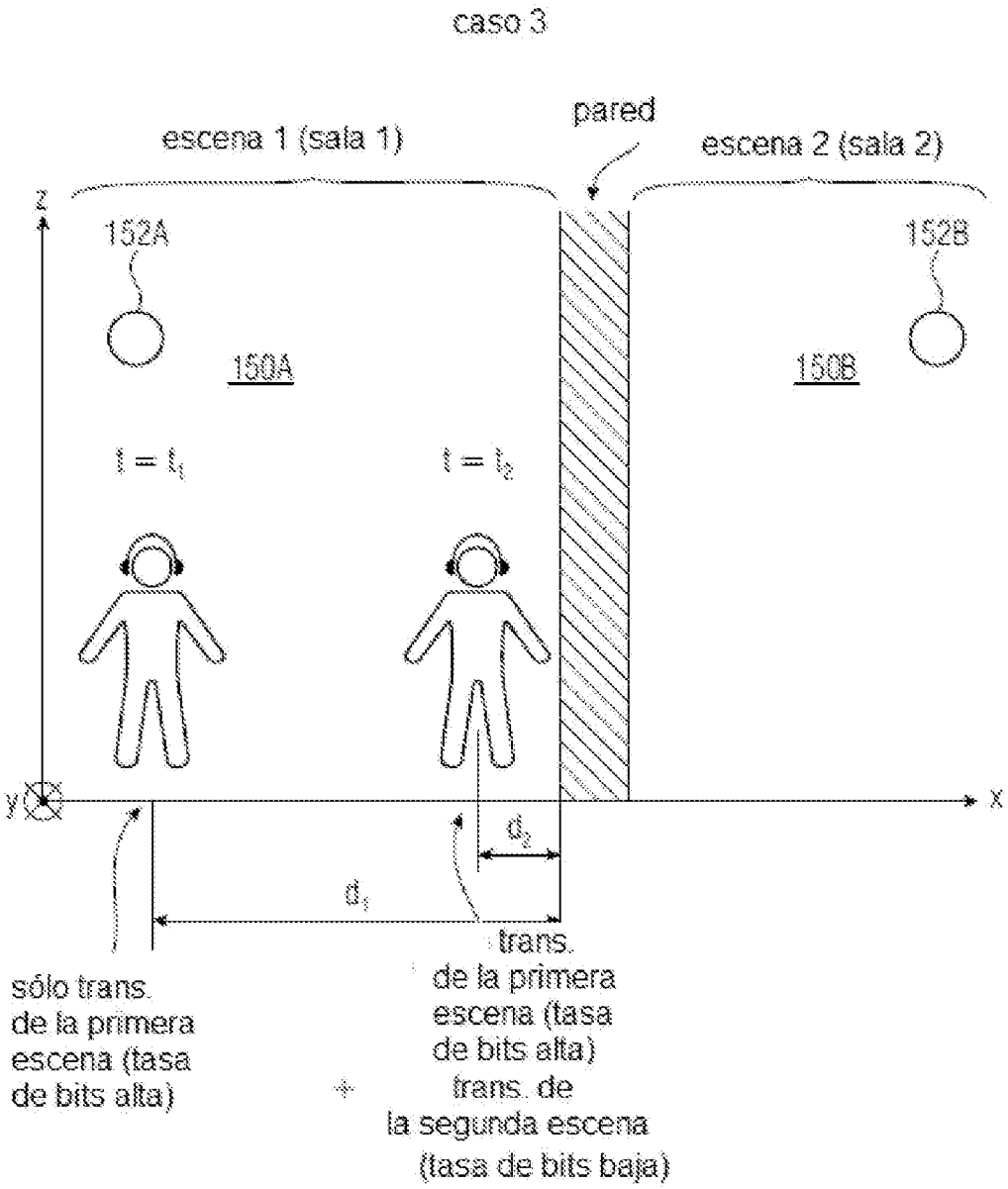
caso 3
pared
escena 1 (sala 1)
escena 2 (sala 2)
z
152A
152B
150A
150B
$l = l_1$
$l = l_2$
y
x
d_2
d_1
sólo trans. de la primera escena (tasa de bits alta)
trans. de la primera escena (tasa de bits alta)
trans. de la segunda escena (tasa de bits baja)

Fig. 4

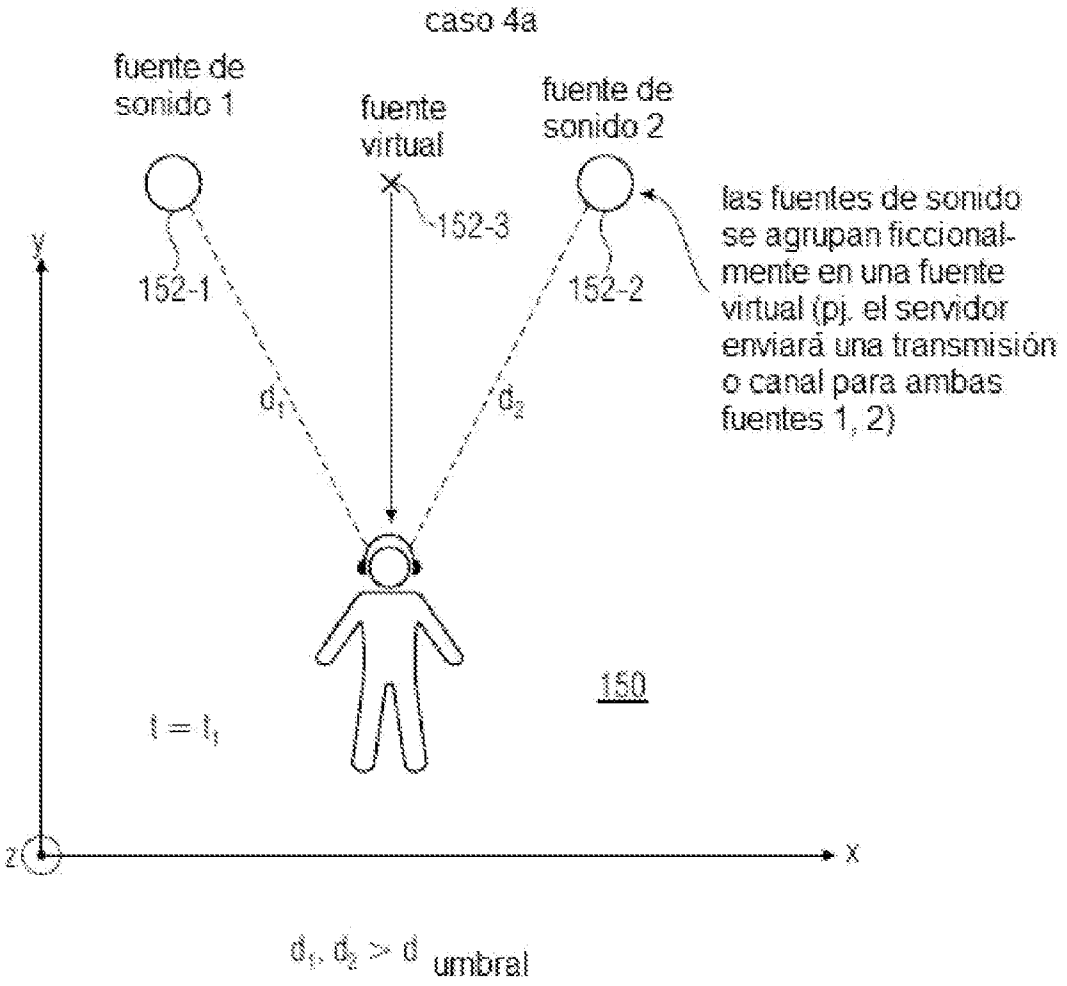
caso 4a
fuente de sonido 1
fuente virtual
fuente de sonido 2
152-3
152-1
152-2
las fuentes de sonido se agrupan ficcionalmente en una fuente virtual (pj. el servidor enviará una transmisión o canal para ambas fuentes 1, 2)
d_1
d_2
150
$l = l_1$
y
x
z
$d_1, d_2 > d_{umbral}$

Fig. 5a

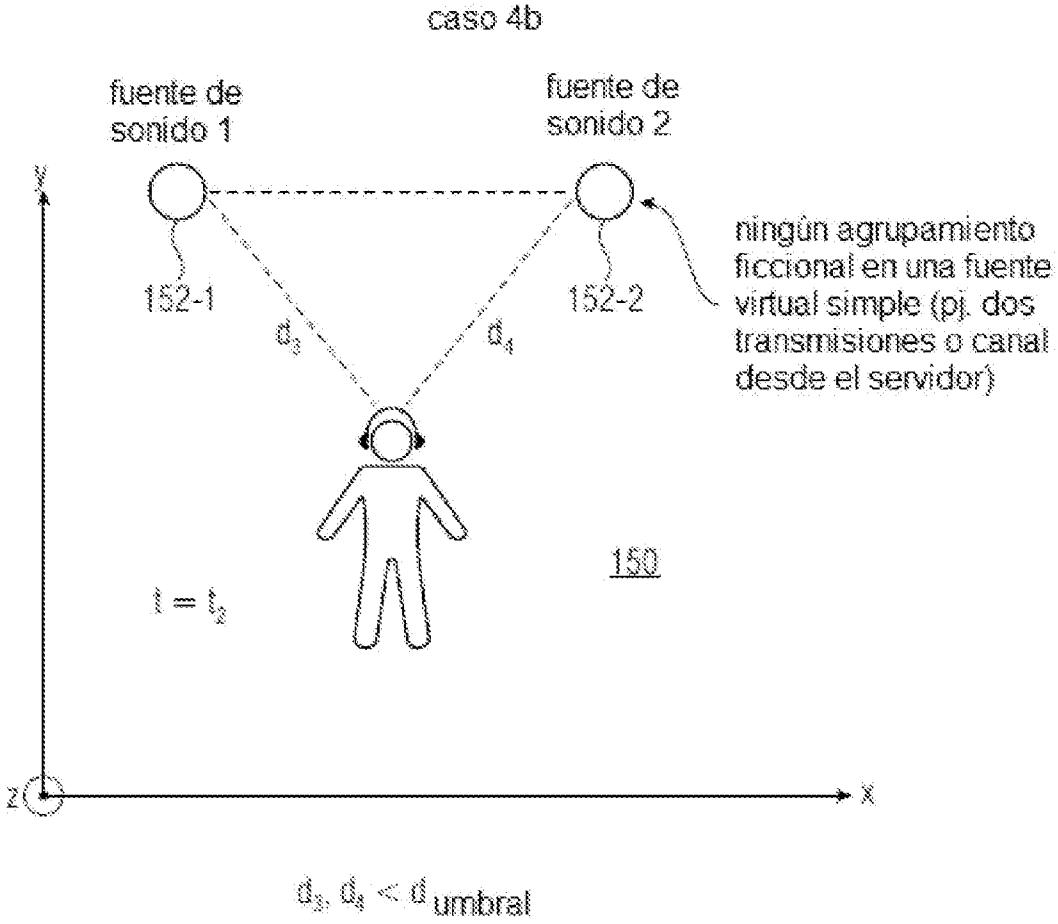
caso 4b
fuente de sonido 1
fuente de sonido 2
152-1
152-2
d_3
d_4
ningún agrupamiento ficcional en una fuente virtual simple (pj. dos transmisiones o canal desde el servidor)
150
$t = t_2$
Y
Z
X
$d_3, d_4 < d_{umbral}$

Fig. 5b

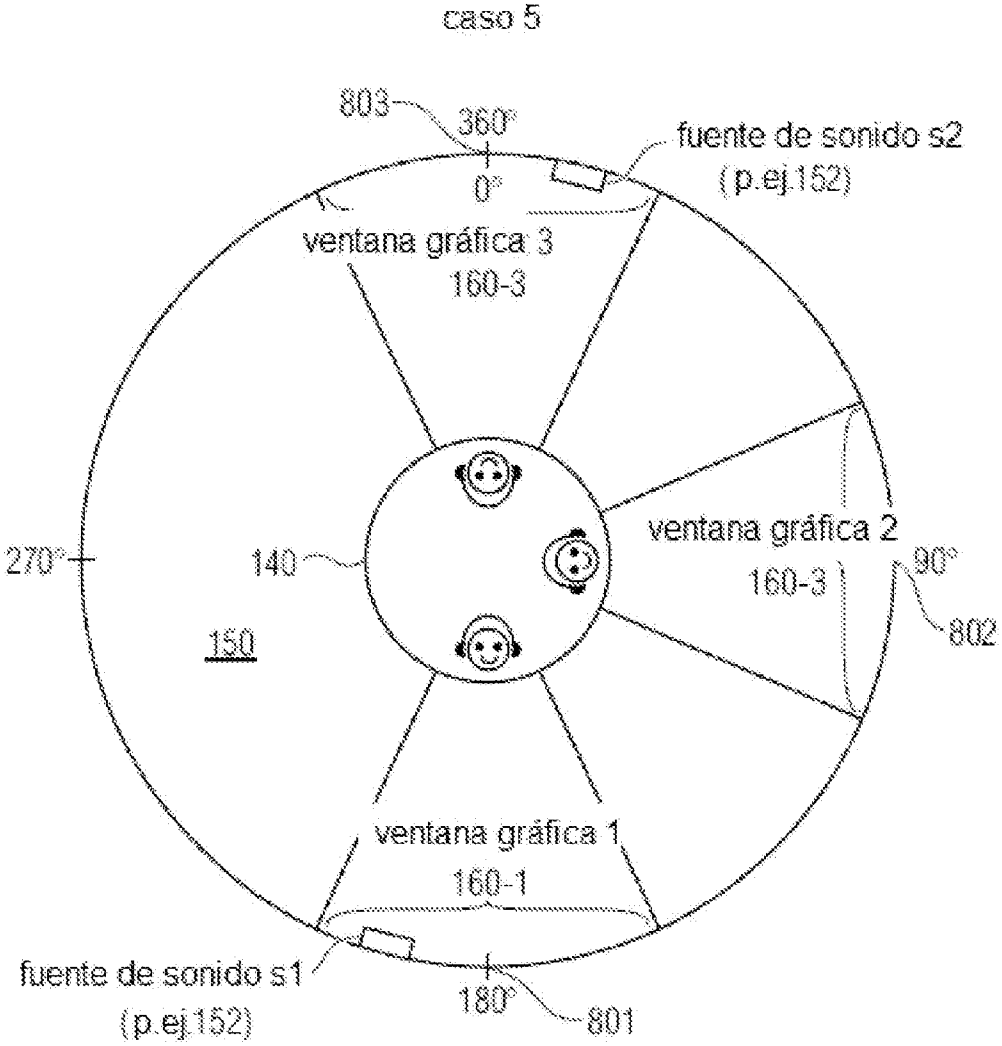
caso 5
803
360°
0°
fuente de sonido s2
(p.ej.152)
ventana gráfica 3
160-3
270°
140
150
ventana gráfica 2
160-3
90°
802
ventana gráfica 1
160-1
fuente de sonido s1
(p.ej.152)
180°
801

Fig. 6

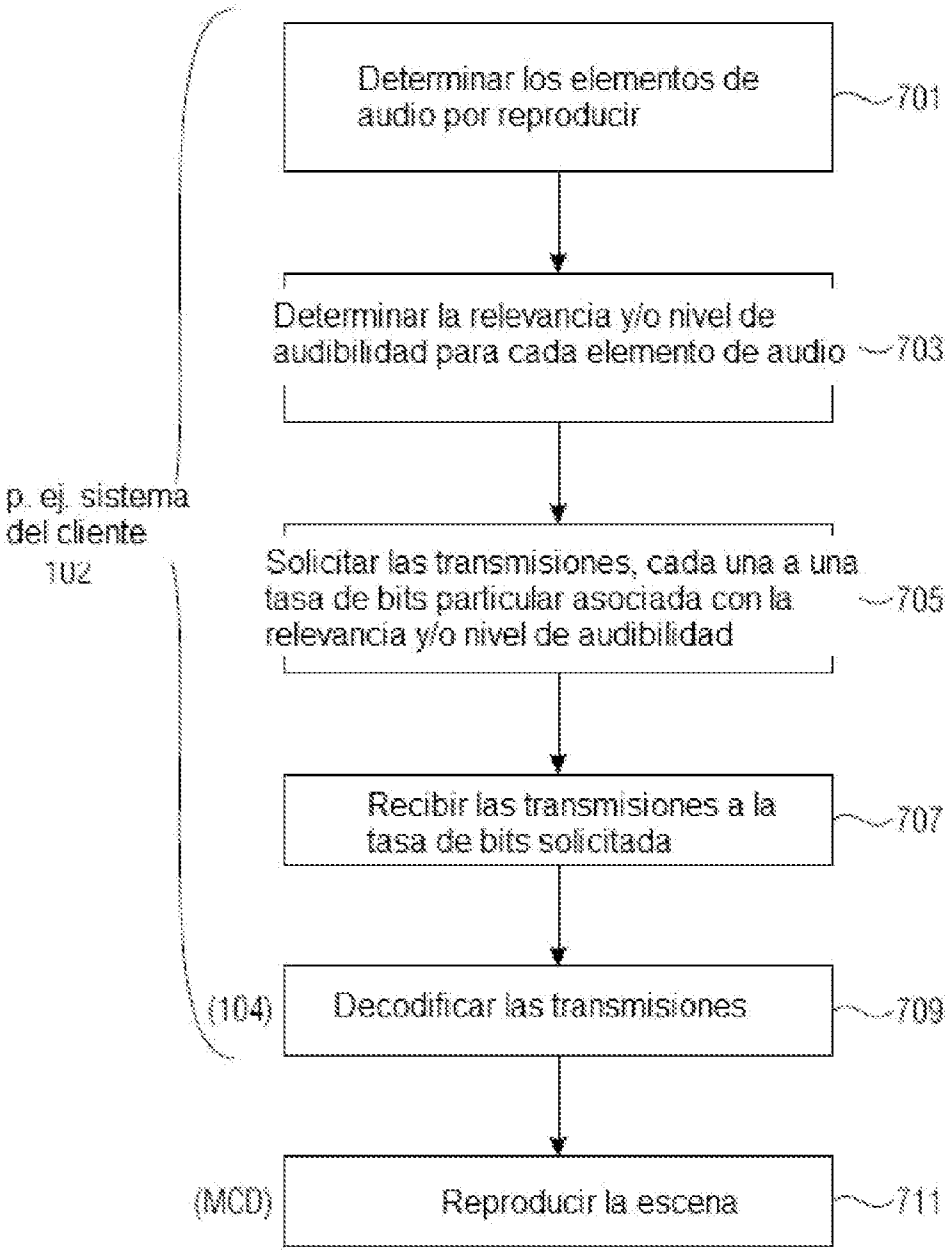
Determinar los elementos de audio por reproducir
701
Determinar la relevancia y/o nivel de audibilidad para cada elemento de audio
703
Solicitar las transmisiones, cada una a una tasa de bits particular asociada con la relevancia y/o nivel de audibilidad
705
Recibir las transmisiones a la tasa de bits solicitada
707
Decodificar las transmisiones
709
Reproducir la escena
711
p. ej. sistema del cliente
102
(104)
(MCD)

Fig. 7a

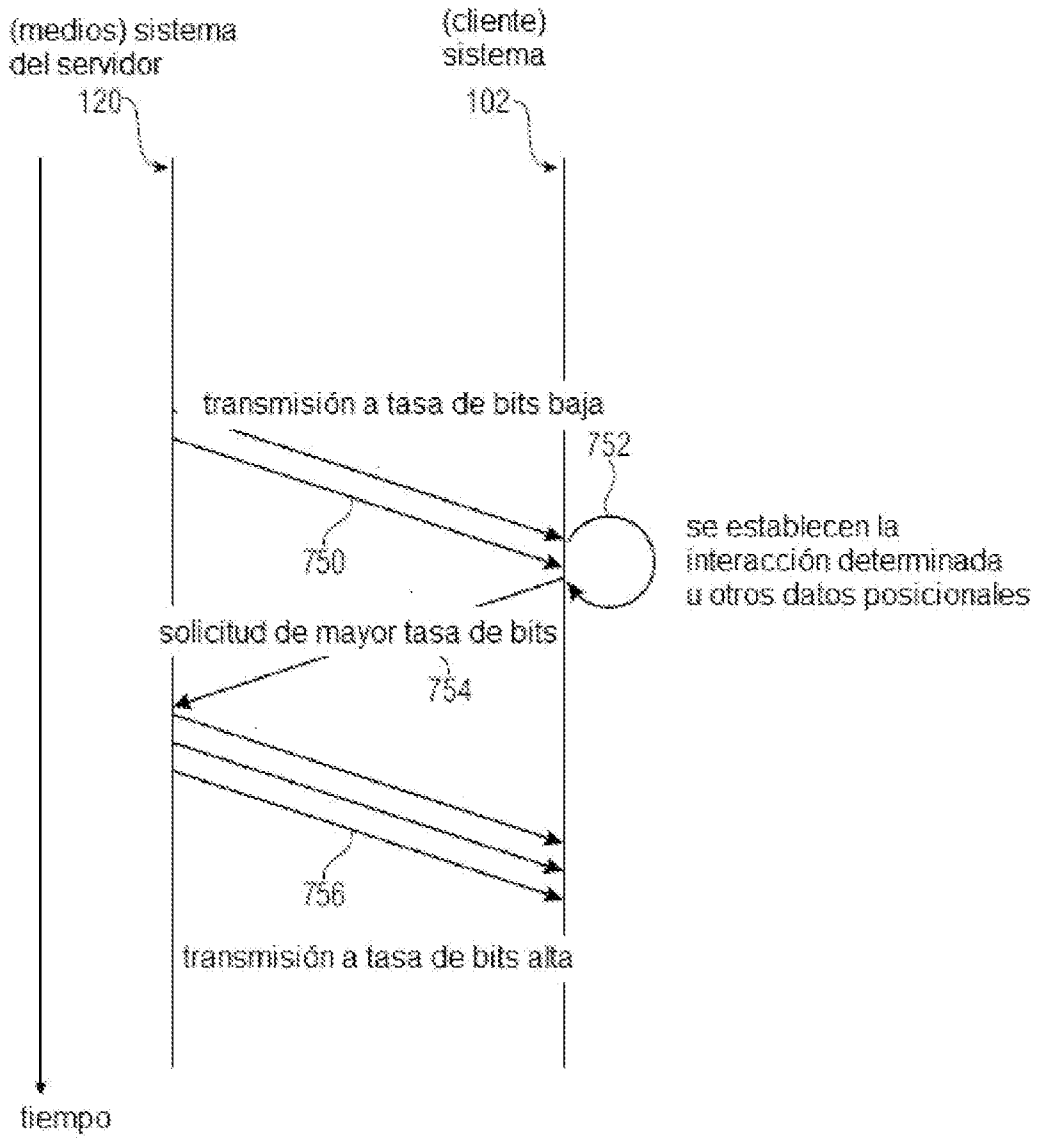
(medios) sistema del servidor
120
(cliente) sistema
102
transmisión a tasa de bits baja
752
750
se establecen la interacción determinada u otros datos posicionales
solicitud de mayor tasa de bits
754
756
transmisión a tasa de bits alta
tiempo

Fig. 7b

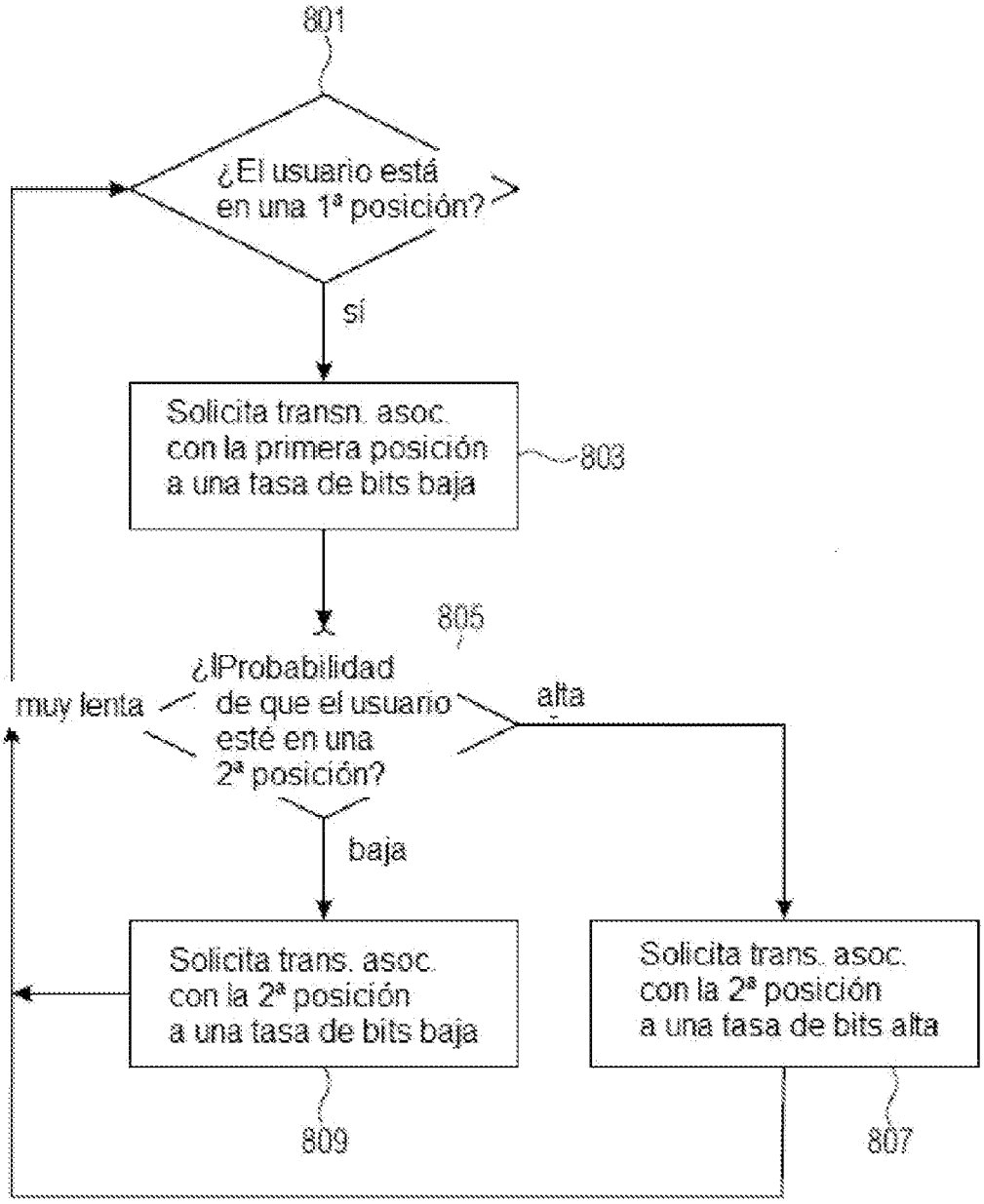
801
¿El usuario está en una 1ª posición?
sí
Solicita transn. asoc. con la primera posición a una tasa de bits baja
803
805
¿Probabilidad de que el usuario esté en una 2ª posición?
muy lenta
alta
baja
Solicita trans. asoc. con la 2ª posición a una tasa de bits baja
Solicita trans. asoc. con la 2ª posición a una tasa de bits alta
809
807

Fig. 8a

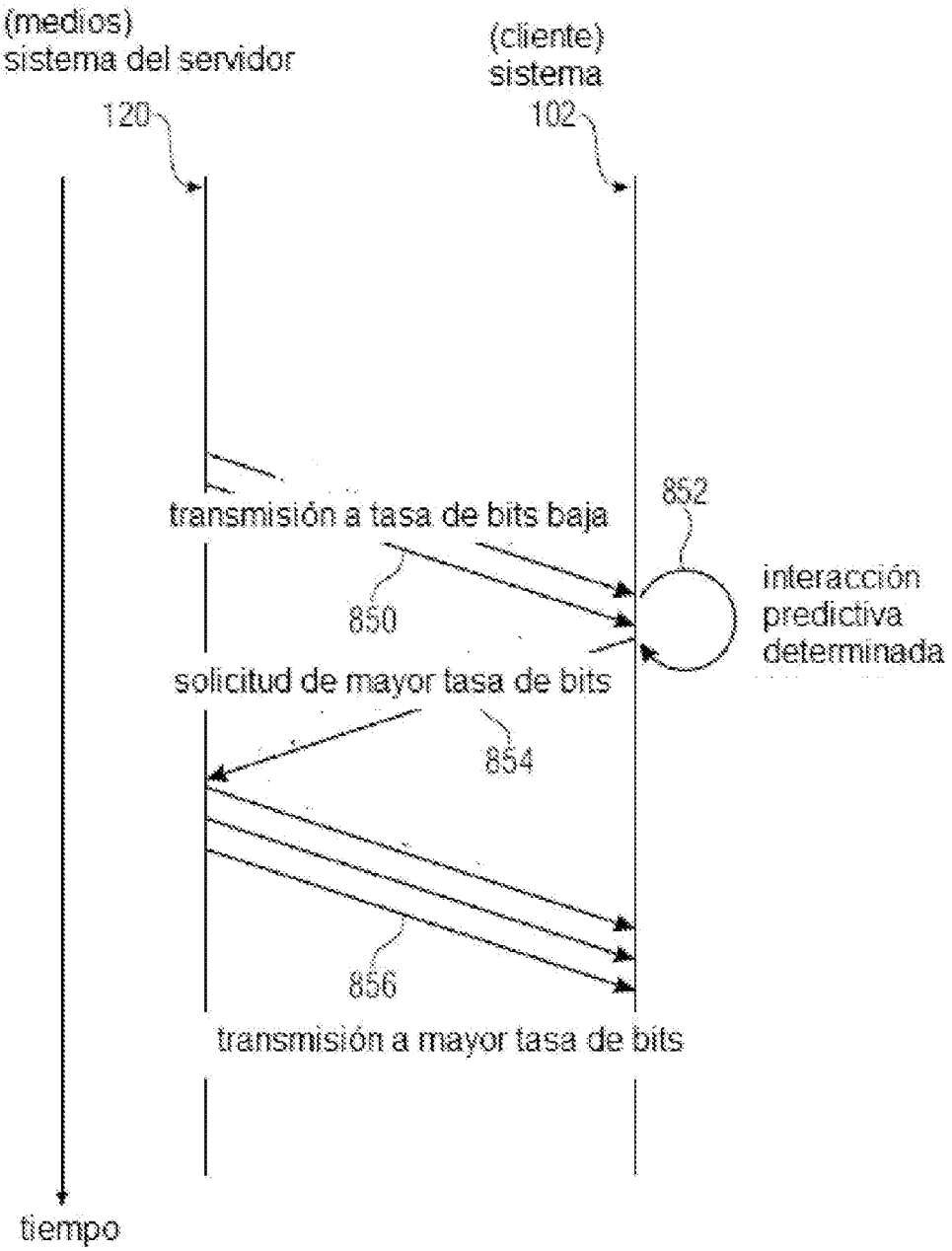
(medios)
sistema del servidor
120
(cliente)
sistema
102
transmisión a tasa de bits baja
852
interacción
predictiva
determinada
850
solicitud de mayor tasa de bits
854
856
transmisión a mayor tasa de bits
tiempo

Fig. 8b