(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 985 934**

(51) Int. Cl.:

H04R 3/00 (2006.01)
H04S 7/00 (2006.01)
G10L 19/008 (2013.01)
H04R 1/40 (2006.01)
H04S 3/02 (2006.01)
G10L 19/16 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **12.11.2019 PCT/US2019/060862**

(87) Fecha y número de publicación internacional: **22.05.2020 WO20102156**

(96) Fecha de presentación y número de la solicitud europea: **12.11.2019 E 19836166 (9)**

(97) Fecha y número de publicación de la concesión europea: **24.07.2024 EP 3881560**

(54) Título: **Representar audio espacial por medio de una señal de audio y metadatos asociados**

(30) Prioridad:

13.11.2018 US 201862760262 P
22.01.2019 US 201962795248 P
02.04.2019 US 201962828038 P
28.10.2019 US 201962926719 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
07.11.2024

(73) Titular/es:

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

(72) Inventor/es:

BRUHN, STEFAN

(74) Agente/Representante:

LINAGE GONZÁLEZ, Rafael

DESCRIPCIÓN

Representar audio espacial por medio de una señal de audio y metadatos asociados

Referencia cruzada a solicitudes relacionadas

Esta solicitud reivindica el beneficio de prioridad a la solicitud de patente provisional de los Estados Unidos n.º 62/760.262 presentada el 13 de noviembre de 2018; la solicitud de patente provisional de los Estados Unidos n.º 62/795.248, presentada el 22 de enero de 2019; la solicitud de patente provisional de los Estados Unidos n.º 62/828.038 presentada el 2 de abril de 2019; y la solicitud de patente provisional de los Estados Unidos n.º 62/926.719 presentada el 28 de octubre de 2019.

Campo técnico

La divulgación en el presente documento se refiere generalmente a codificación de una escena de audio que comprende objetos de audio. En particular, se refiere a métodos, sistemas, productos de programa de ordenador y formatos de datos para representar audio espacial, y un codificador, descodificador y renderizador asociados para codificar, descodificar y renderizar audio espacial.

Antecedentes

La introducción de acceso inalámbrico de alta velocidad de 4G/5G a redes de telecomunicaciones, combinada con la disponibilidad de plataformas de hardware cada vez más potentes, ha proporcionado una fundación para que comunicaciones avanzadas y servicios multimedia se desplieguen de manera más rápida y fácil que nunca hasta ahora.

El códec de servicios de voz potenciados (EVS) del proyecto de asociación de tercera generación (3GPP) ha ofrecido una mejora altamente significativa en la experiencia de usuario con la introducción de codificación de habla y audio de banda super ancha (SWB) y banda completa (FB), junto con una resiliencia mejorada a pérdida de paquetes. Sin embargo, el ancho de banda de audio extendido es solo una de las dimensiones requeridas para una experiencia verdaderamente inmersiva. Para la inmersión del usuario en un mundo virtual convincente de una manera eficiente en recursos se requiere idealmente soporte más allá del mono y multi-mono ofrecido actualmente por EVS.

Además, los códecs de audio especificados actualmente en 3GPP proporcionan calidad y compresión adecuadas para contenido estéreo pero carecen de los rasgos conversacionales (por ejemplo, latencia suficientemente baja) necesarios para voz conversacional y teleconferencia. Estos codificadores también carecen de funcionalidad de canal múltiple que es necesaria para servicios inmersivos, tales como recepción y visualización simultáneas a demanda de contenido multimedia (o streaming) en directo, teleconferencia inmersiva y de realidad virtual (VR).

Se ha propuesto una extensión al códec de EVS para servicios de voz y audio inmersivos (IVAS) para rellenar este espacio de tecnología y abordar la demanda creciente de servicios multimedia ricos. Además, aplicaciones de teleconferencia sobre 4G/5G se beneficiarán de un códec de IVAS usado como codificador conversacional mejorado que soporta codificación de flujo múltiple (por ejemplo, audio basado en canales, objetos y escena). Los casos de uso para este códec de próxima generación incluyen, pero no se limitan a, voz conversacional, teleconferencia de flujo múltiple, streaming de contenido en directo y en no directo generado por usuario y conversacional de VR.

Aunque el objetivo es desarrollar un único códec con rasgos y rendimiento atractivos (por ejemplo, excelente calidad de audio, bajo retardo, soporte de codificación de audio espacial, rango apropiado de tasas de bits, resiliencia a errores de alta calidad, complejidad de implementación práctica), actualmente no hay un acuerdo finalizado en el formato de entrada de audio del códec de IVAS. El formato de audio espacial asistido por metadatos (MASA) se ha propuesto como un posible formato de entrada de audio. Sin embargo, los parámetros de MASA convencionales hacen ciertas suposiciones idealistas, tales como captura de audio que se hace en un único punto. Sin embargo, en un escenario del mundo real, donde un teléfono móvil o tableta se usa como un dispositivo de captura de audio, tal suposición de captura de sonido en un único punto puede no mantenerse. En cambio, dependiendo del factor de forma del dispositivo particular, los diversos micrófonos del dispositivo pueden ubicarse a alguna distancia y las diferentes señales de micrófono capturadas pueden no estar completamente alineadas en el tiempo. Esto es particularmente cierto cuando también se considera cómo la fuente del audio puede moverse en el espacio.

Otra suposición subyacente del formato de MASA es que todos los canales de micrófono se proporcionan a igual nivel y que no hay diferencias en la respuesta de frecuencia y fase entre ellos. De nuevo, en un escenario del mundo real, los canales de micrófono pueden tener diferentes características de frecuencia y fase dependientes de dirección, que también pueden ser variables en el tiempo. Se podría suponer, por ejemplo, que el dispositivo

de captura de audio se mantiene temporalmente de tal manera que uno de los micrófonos está ocluido o que hay algún objeto en las proximidades del teléfono que causa reflexiones o difracciones de las ondas sonoras que llegan. De este modo, hay muchos factores adicionales para tener en cuenta cuando se determina qué formato de audio sería adecuado junto con un códec tal como el códec de IVAS.

El documento WO2017/182714 A1 describe codificar canales de señales de audio de múltiples micrófonos y combinarlos con metadatos espaciales en un flujo de bits y generar opcionalmente una mezcla descendente de canales. El documento US2015/0142427 A1 describe un codificador que comprende un mezclador descendente que mezcla de manera descendente un número de señales de audio a una señal de mezcla descendente. El documento US2016/0180826 describe un sistema de captación que incluye un detector de viento y un supresor de viento que emite una señal de indicación de nivel de viento indicativa de actividad de viento. El documento US2018/0098174 A1 describe incluir una señal de mezcla descendente de audio compatible hacia atrás de dos canales o de canal múltiple junto con extensiones opcionales (denominadas en el presente documento como "información colateral") en un flujo de bits de audio digital producido por un codificador de flujo de bits de audio.

Breve descripción de los dibujos

Realizaciones de ejemplo se describen ahora con referencia a los dibujos que se acompañan, en los que:

La figura 1 es un diagrama de flujo de un método para representar audio espacial de acuerdo con realizaciones ejemplares;

La figura 2 es una ilustración esquemática de un dispositivo de captura de audio y fuentes de sonido direccionales y difusas, respectivamente, de acuerdo con realizaciones ejemplares;

La figura 3A muestra una tabla (tabla 1A) de cómo un parámetro de valor de bit de canal indica cuántos canales se usan para el formato de MASA, de acuerdo con realizaciones ejemplares.

La figura 3B muestra una tabla (tabla 1B) de una estructura de metadatos que puede usarse para representar captura de FOA y FOA planaria con mezcla descendente en dos canales de MASA, de acuerdo con realizaciones ejemplares;

La figura 4 muestra una tabla (tabla 2) de valores de compensación de retardo para cada micrófono y por baldosa de TF, de acuerdo con realizaciones ejemplares;

La figura 5 muestra una tabla (tabla 3) de una estructura de metadatos que puede usarse para indicar qué conjunto de valores de compensación se aplica a qué baldosa de TF, de acuerdo con realizaciones ejemplares;

La figura 6 muestra una tabla (tabla 4) de una estructura de metadatos que puede usarse para representar ajuste de ganancia para cada micrófono, de acuerdo con realizaciones ejemplares;

La figura 7 muestra un sistema que incluye un dispositivo de captura de audio, un codificador, un descodificador y un renderizador, de acuerdo con realizaciones ejemplares.

La figura 8 muestra un dispositivo de captura de audio, de acuerdo con realizaciones ejemplares.

La figura 9 muestra un descodificador y renderizador, de acuerdo con realizaciones ejemplares.

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

Descripción detallada

Es un objeto de la invención superar las carencias de la técnica anterior. Este objeto de la invención se resuelve mediante las reivindicaciones independientes. Realizaciones específicas se definen en las reivindicaciones dependientes.

I. Visión general - Representación de audio espacial

De acuerdo con un primer aspecto, se proporciona un método, un sistema, un producto de programa de ordenador y un formato de datos para representar audio espacial.

De acuerdo con realizaciones ejemplares se proporciona un método para representar audio espacial, siendo el audio espacial una combinación de sonido direccional y sonido difuso, que comprende:

crear una señal de audio de mezcla descendente de canal individual o múltiple mediante la mezcla descendente de señales de audio de entrada procedentes de una pluralidad de micrófonos en una unidad de captura de audio que captura el audio espacial;

determinar primeros parámetros de metadatos asociados con la señal de audio de mezcla descendente, en el que los primeros parámetros de metadatos son indicativos de uno o más de: un valor de retardo de tiempo relativo, un valor de ganancia, y un valor de fase asociados con cada señal de audio de entrada; y

combinar la señal de audio de mezcla descendente creada y los primeros parámetros de metadatos en una representación del audio espacial.

Con la disposición anterior, una representación mejorada del audio espacial puede lograrse, teniendo en cuenta diferentes propiedades y/o posiciones espaciales de la pluralidad de micrófonos. Además, usar los metadatos en las etapas de procesamiento posteriores de codificación, descodificación o renderización puede contribuir a representar y reconstruir fielmente el audio capturado mientras se representa el audio en una forma codificada eficiente de tasa de bits.

De acuerdo con realizaciones ejemplares, combinar la señal de audio de mezcla descendente creada y los primeros parámetros de metadatos en una representación del audio espacial puede comprender además incluir segundos parámetros de metadatos en la representación del audio espacial, siendo los segundos parámetros de metadatos indicativos de una configuración de mezcla descendente para las señales de audio de entrada.

Esto es ventajoso porque permite reconstruir (por ejemplo, a través de una operación de mezcla ascendente) las señales de audio de entrada en un descodificador. Además, proporcionando los segundos metadatos, puede realizarse una mezcla descendente adicional mediante una unidad separada antes de codificar la representación del audio espacial en un flujo de bits.

De acuerdo con realizaciones ejemplares los primeros parámetros de metadatos pueden determinarse para una o más bandas de frecuencia de las señales de audio de entrada del micrófono.

Esto es ventajoso porque permite parámetros adaptados individualmente de retardo, ganancia y/o ajuste de fase, por ejemplo, considerando las diferentes respuestas de frecuencia para diferentes bandas de frecuencia de las señales del micrófono.

De acuerdo con realizaciones ejemplares, la mezcla descendente para crear una señal x de audio de mezcla descendente de canal individual o múltiple puede describirse mediante:

$$x = D \cdot m$$

en la que:

D es una matriz de mezcla descendente que contiene coeficientes de mezcla descendente que definen pesos para cada señal de audio de entrada procedente de la pluralidad de micrófonos, y

m es una matriz que representa las señales de audio de entrada procedentes de la pluralidad de micrófonos.

De acuerdo con realizaciones ejemplares, los coeficientes de mezcla descendente pueden elegirse para seleccionar la señal de audio de entrada de micrófono que tiene actualmente la mejor relación de señal sobre ruido con respecto al sonido direccional, y para descartar señales de audio de entrada de señal procedentes de cualesquiera otros micrófonos.

Esto es ventajoso porque permite lograr una buena representación de calidad del audio espacial con una complejidad de cálculo reducida en la unidad de captura de audio. En esta realización, solo se elige una señal de audio de entrada para representar el audio espacial en una trama de audio específica y/o baldosa de frecuencia-tiempo. En consecuencia, la complejidad de cálculo para la operación de mezcla descendente se reduce.

De acuerdo con realizaciones ejemplares la selección puede determinarse bajo una premisa por baldosa de tiempo-frecuencia (TF).

Esto es ventajoso porque permite una operación de mezcla descendente mejorada, por ejemplo, considerando las diferentes respuestas de frecuencia para diferentes bandas de frecuencia de las señales de micrófono.

De acuerdo con realizaciones ejemplares la selección puede hacerse para una trama de audio particular.

Ventajosamente, esto permite adaptaciones con respecto a señales de captura de micrófono que varían en el tiempo, y a su vez a una calidad de audio mejorada.

De acuerdo con realizaciones ejemplares, los coeficientes de mezcla descendente pueden elegirse para maximizar la relación de señal sobre ruido con respecto al sonido direccional, cuando se combinan las señales de audio de entrada procedentes de los diferentes micrófonos

Esto es ventajoso porque permite una calidad mejorada de la mezcla descendente debido a la atenuación de componentes de señal no deseados que no surgen de las fuentes direccionales.

De acuerdo con realizaciones ejemplares la maximización puede hacerse para una banda de frecuencia particular.

De acuerdo con realizaciones ejemplares la maximización puede hacerse para una trama de audio particular.

De acuerdo con realizaciones ejemplares determinar primeros parámetros de metadatos pueden incluir analizar una o más de: características de retardo, ganancia y fase de las señales de audio de entrada procedentes de la pluralidad de micrófonos.

De acuerdo con realizaciones ejemplares los primeros parámetros de metadatos pueden determinarse bajo una premisa por baldosa de tiempo-frecuencia (TF).

De acuerdo con realizaciones ejemplares al menos una porción de la mezcla descendente puede producirse en la unidad de captura de audio.

De acuerdo con realizaciones ejemplares al menos una porción de la mezcla descendente puede producirse en un codificador.

De acuerdo con realizaciones ejemplares, cuando se detecta más de una fuente de sonido direccional, se pueden determinar primeros metadatos para cada fuente.

De acuerdo con realizaciones ejemplares la representación del audio espacial puede incluir al menos uno de los siguientes parámetros: un índice de dirección, una relación de energía directa sobre total; una coherencia de dispersión; un ganancia, fase y tiempo de llegada para cada micrófono; una relación de energía difusa sobre total; una coherencia envolvente; una relación de energía restante sobre total; y una distancia.

De acuerdo con realizaciones ejemplares un parámetro de metadatos de los segundos o primeros parámetros de metadatos puede indicar si la señal de audio de mezcla descendente creada se genera a partir de: señales estéreo izquierda derecha, señales ambisónicas de primer orden (FOA) planarias, o señales de componente de FOA.

De acuerdo con realizaciones ejemplares la representación del audio espacial puede contener parámetros de metadatos organizados en un campo de definición y un campo de selector, en la que el campo de definición especifica al menos un conjunto de parámetros de compensación de retardo asociado con la pluralidad de micrófonos, y el campo de selector especifica la selección de un conjunto de parámetros de compensación de retardo.

De acuerdo con realizaciones ejemplares el campo de selector puede especificar qué conjunto de parámetros de compensación de retardo se aplica a cualquier baldosa de tiempo-frecuencia dada.

De acuerdo con realizaciones ejemplares el valor de retardo de tiempo relativo puede estar aproximadamente en el intervalo de [-2,0 ms, 2,0 ms]

De acuerdo con realizaciones ejemplares los parámetros de metadatos en la representación del audio espacial pueden incluir además un campo que especifica el ajuste de ganancia aplicado y un campo que especifica el ajuste de fase.

De acuerdo con realizaciones ejemplares el ajuste de ganancia puede estar aproximadamente en el intervalo de [+10 dB, -30 dB].

De acuerdo con realizaciones ejemplares al menos partes de los primeros y/o segundos elementos de metadatos se determinan en el dispositivo de captura de audio usando tablas de consulta almacenadas.

De acuerdo con realizaciones ejemplares al menos partes de los primeros y/o segundos elementos de metadatos se determinan en un dispositivo remoto conectado al dispositivo de captura de audio.

II. Visión general - Sistema

De acuerdo con un segundo aspecto, se proporciona un sistema para representar audio espacial.

De acuerdo con realizaciones ejemplares se proporciona un sistema para representar audio espacial, que comprende:

un componente de recepción configurado para recibir señales de audio de entrada procedentes de una pluralidad de micrófonos en una unidad de captura de audio que captura el audio espacial;

un componente de mezcla descendente configurado para crear una señal de audio de mezcla descendente de canal individual o múltiple mediante la mezcla descendente de las señales de audio recibidas;

un componente de determinación de metadatos configurado para determinar primeros parámetros de metadatos asociados con la señal de audio de mezcla descendente, en el que los primeros parámetros de metadatos son indicativos de uno o más de: un valor de retardo de tiempo relativo, un valor de ganancia y un valor de fase asociados con cada señal de audio de entrada; y

un componente de combinación configurado para combinar la señal de audio de mezcla descendente creada y los primeros parámetros de metadatos en una representación del audio espacial.

III. Visión general - Formato de datos

De acuerdo con un tercer aspecto, se proporciona formato de datos para representar audio espacial. El formato de datos puede usarse ventajosamente junto con componentes físicos relacionados con audio espacial, tales como dispositivos de captura de audio, codificadores, descodificadores, renderizadores, y así sucesivamente, y diversos tipos de productos de programa de ordenador y otro equipo que se usa para transmitir audio espacial entre dispositivos y/o ubicaciones.

De acuerdo con realizaciones de ejemplo, el formato de datos comprende:

una señal de audio de mezcla descendente resultante de una mezcla descendente de señales de audio de entrada procedentes de una pluralidad de micrófonos en una unidad de captura de audio que captura el audio espacial; y

primeros parámetros de metadatos indicativos de uno o más de: una configuración de mezcla descendente para las señales de audio de entrada, un valor de retardo de tiempo relativo, un valor de ganancia, y un valor de fase asociados con cada señal de audio de entrada.

De acuerdo con un ejemplo, el formato de datos se almacena en una memoria no transitoria.

IV. Visión general - Codificador

De acuerdo con un cuarto aspecto, se proporciona un codificador para codificar una representación de audio espacial.

De acuerdo con realizaciones ejemplares, se proporciona un codificador configurado para:

recibir una representación de audio espacial, comprendiendo la representación:

una señal de audio de mezcla descendente de canal individual o múltiple creada mediante la mezcla descendente de señales de audio de entrada procedentes de una pluralidad de micrófonos en una unidad de captura de audio que captura el audio espacial, y

primeros parámetros de metadatos asociados con la señal de audio de mezcla descendente, en el que los primeros parámetros de metadatos son indicativos de uno o más de: un valor de retardo de tiempo relativo, un valor de ganancia, y un valor de fase asociados con cada señal de audio de entrada; y

codificar la señal de audio de mezcla descendente de canal individual o múltiple en un flujo de bits usando los primeros metadatos, o

codificar la señal de audio de mezcla descendente de canal individual o múltiple y los primeros metadatos en un flujo de bits.

V. Visión general - Descodificador

De acuerdo con un quinto aspecto, se proporciona un descodificador para descodificar una representación de audio espacial.

De acuerdo con realizaciones ejemplares se proporciona un descodificador configurado para:

recibir un flujo de bits indicativo de una representación codificada de audio espacial, comprendiendo la representación:

una señal de audio de mezcla descendente de canal individual o múltiple creada mediante la mezcla descendente de señales de audio de entrada procedentes de una pluralidad de micrófonos en una unidad de captura de audio que captura el audio espacial, y

primeros parámetros de metadatos asociados con la señal de audio de mezcla descendente, en el que los primeros parámetros de metadatos son indicativos de uno o más de: un valor de retardo de tiempo relativo, un valor de ganancia, y un valor de fase asociados con cada señal de audio de entrada; y

descodificar el flujo de bits en una aproximación del audio espacial, usando los primeros parámetros de metadatos.

VI. Visión general - Renderizador

De acuerdo con un sexto aspecto, se proporciona un renderizador para renderizar una representación de audio espacial.

De acuerdo con realizaciones ejemplares se proporciona un renderizador configurado para:

recibir una representación de audio espacial, comprendiendo la representación:

una señal de audio de mezcla descendente de canal individual o múltiple creada mediante la mezcla descendente de señales de audio de entrada procedentes de una pluralidad de micrófonos en una unidad de captura de audio que captura el audio espacial, y

primeros parámetros de metadatos asociados con la señal de audio de mezcla descendente, en el que los primeros parámetros de metadatos son indicativos de uno o más de: un valor de retardo de tiempo relativo, un valor de ganancia, y un valor de fase asociados con cada señal de audio de entrada; y

renderizar el audio espacial usando los primeros metadatos.

VII. Visión general - Generalmente

Los aspectos segundo a sexto pueden tener generalmente los mismos rasgos y ventajas que el primer aspecto.

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

Los pasos de cualquier método divulgado en el presente documento no tienen que realizarse en el orden exacto divulgado, a menos que se indique explícitamente.

VIII. Realizaciones de ejemplo

Como se ha descrito anteriormente, captura y representación de audio espacial presenta un conjunto específico de desafíos, de tal manera que el audio capturado puede reproducirse fielmente en el extremo de recepción. Las diversas realizaciones de la presente invención descritas en el presente documento abordan diversos aspectos de estos temas, incluyendo diversos parámetros de metadatos junto con la señal de audio de mezcla descendente cuando se transmite la señal de audio de mezcla descendente.

La invención se describirá a modo de ejemplo, y con referencia al formato de audio de MASA. Sin embargo, es importante darse cuenta de que los principios generales de la invención son aplicables a un amplio rango de formatos que pueden usarse para representar audio, y la descripción en el presente documento no se limita a MASA.

Además, se debe ser consciente de que los parámetros de metadatos que se describen a continuación no son una lista completa de parámetros de metadatos, sino que puede haber parámetros de metadatos adicionales (o un subconjunto más pequeño de parámetros de metadatos) que se pueden usar para llevar datos sobre la señal de audio de mezcla descendente a los diversos dispositivos usados en la codificación, descodificación y renderización del audio.

Además, mientras que los ejemplos en el presente documento se describirán en el contexto de un codificador de

IVAS, debe apreciarse que éste es simplemente un tipo de codificador en el que pueden aplicarse los principios generales de la invención, y que puede haber muchos otros tipos de codificadores, descodificadores, y renderizadores que pueden usarse junto con las diversas realizaciones descritas en el presente documento.

Por último, debe apreciarse que mientras que los términos "mezcla ascendente" y "mezcla descendente" se usan a lo largo de este documento, pueden no implicar necesariamente aumentar y reducir, respectivamente, el número de canales. Aunque este puede ser el caso a menudo, se debe ser consciente de que cualquier término puede referirse a reducir o aumentar el número de canales. De este modo, ambos términos caen dentro del concepto más general de "mezcla". De manera similar, el término "señal de audio de mezcla descendente" se usará a lo largo de la memoria descriptiva, pero se debe ser consciente de que ocasionalmente pueden usarse otros términos, tales como "canal de MASA", "canal de transporte" o "canal de mezcla descendente", todos los cuales tienen esencialmente el mismo significado que "señal de audio de mezcla descendente".

Volviendo ahora a la figura 1, un método 100 se describe para representar audio espacial, de acuerdo con una realización. Como puede verse en la figura 1, el método comienza capturando audio espacial usando un dispositivo de captura de audio, paso 102. La figura 2 muestra una vista esquemática de un entorno de sonido 200 en el que un dispositivo de captura de audio 202, tal como un teléfono móvil o tableta, por ejemplo, captura audio de una fuente ambiental difusa 204 y una fuente direccional 206, tal como un orador. En la realización ilustrada, el dispositivo de captura de audio 202 tiene tres micrófonos m1, m2 y m3, respectivamente.

El sonido direccional incide desde una dirección de llegada (DOA) representada por ángulos azimutal y de elevación. Se supone que el sonido ambiental difuso es omnidireccional, es decir, espacialmente invariable o espacialmente uniforme. También se considera en la siguiente explicación la posible aparición de una segunda fuente de sonido direccional, que no se muestra en la figura 2.

A continuación, las señales de los micrófonos se mezclan de manera descendente para crear una señal de audio de mezcla descendente de canal individual o múltiple, paso 104. Hay muchas razones para propagar solo una señal de audio de mezcla descendente mono. Por ejemplo, puede haber limitaciones de tasa de bits o la intención poner a disposición una señal de audio de mezcla descendente monocanal de alta calidad después de que se hayan hecho ciertos potenciamientos propietarios, tales como formación de haces y ecualización o supresión de ruido. En otras realizaciones, la mezcla descendente da como resultado una señal de audio de mezcla descendente de canal múltiple. Generalmente, el número de canales en la señal de audio de mezcla descendente es menor que el número de señales de audio de entrada, sin embargo, en algunos casos, el número de canales en la señal de audio de mezcla descendente puede ser igual al número de señales de audio de entrada y la mezcla descendente es, en cambio, para lograr una SNR aumentada, o reducir la cantidad de datos en la señal de audio de mezcla descendente resultante en comparación con las señales de audio de entrada. Esto se elabora adicionalmente a continuación.

Propagar los parámetros relevantes usados durante la mezcla descendente al códec de IVAS como parte de los metadatos de MASA puede dar la posibilidad de recuperar la señal estéreo y/o una señal de audio de mezcla descendente espacial con la mejor fidelidad posible.

En este escenario, se obtiene un único canal de MASA mediante la siguiente operación de mezcla descendente:

$$x = D \cdot m,$$

con

$$D = (K_{1,1}\ K_{1,2}\ K_{1,3})$$

y

$$m = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix}.$$

Las señales m y x pueden, durante las diversas etapas de procesamiento, no estar necesariamente representadas como señales de tiempo de banda completa sino posiblemente también como señales de componente de diversas subbandas en el dominio de tiempo o frecuencia (baldosas de TF). En ese caso, eventualmente se recombinarían y se transformarían potencialmente al dominio de tiempo antes de propagarse al códec de IVAS.

Los sistemas de codificación/descodificación de audio dividen típicamente el espacio de tiempo-frecuencia en baldosas de tiempo/frecuencia, por ejemplo aplicando bancos de filtros adecuados a las señales de audio de entrada. Por una baldosa de tiempo/frecuencia se entiende generalmente una porción del espacio de tiempo-

frecuencia correspondiente a un intervalo de tiempo y una banda de frecuencia. El intervalo de tiempo puede corresponder típicamente a la duración de una trama de tiempo usada en el sistema de codificación/descodificación de audio. La banda de frecuencia es una parte de todo el rango de frecuencia de la señal/objeto de audio que está siendo codificado o descodificado. La banda de frecuencia puede corresponder típicamente a una o varias bandas de frecuencia vecinas definidas por un banco de filtros usado en el sistema de codificación/descodificación. En el caso de que la banda de frecuencia corresponda a varias bandas de frecuencia vecinas definidas por el banco de filtros, esto permite tener bandas de frecuencia no uniformes en el proceso de descodificación de la señal de audio de mezcla descendente, por ejemplo, bandas de frecuencia más anchas para frecuencias más altas de la señal de audio de mezcla descendente.

En una implementación que usa un único canal de MASA, hay al menos dos opciones en cuanto a cómo se puede definir la matriz D de mezcla descendente. Una elección es escoger esa señal de micrófono que tiene la mejor relación de señal sobre ruido (SNR) con respecto al sonido direccional. En la configuración mostrada en la figura 2 es probable que el micrófono m1 capture la mejor señal a medida que se dirige hacia la fuente de sonido direccional. Las señales procedentes de los otros micrófonos podrían entonces descartarse. En ese caso, la matriz de mezcla descendente podría ser como sigue:

D = (1 0 0).

Mientras que la fuente de sonido se mueve con respecto al dispositivo de captura de audio, se podría seleccionar otro micrófono más adecuado de modo que cualquier señal m_2 o m_3 se usa como canal de MASA resultante.

Cuando se conmutan las señales de micrófono, es importante asegurarse de que la señal x de canal de MASA no sufre ninguna discontinuidad potencial. Podrían producirse discontinuidades debido a diferentes tiempos de llegada de la fuente de sonido direccional a los diferentes micrófonos, o debido a diferentes características de ganancia o fase de la trayectoria acústica desde la fuente a los micrófonos. Consiguientemente, las características individuales de retardo, ganancia y fase de las diferentes entradas de micrófono deben analizarse y compensarse. Las señales de micrófono reales pueden por lo tanto sufrir cierta operación de algo de ajuste de retardo y filtrado antes de la mezcla descendente de MASA.

En otra realización, los coeficientes de la matriz de mezcla descendente se establecen de tal manera que la SNR del canal de MASA con respecto a la fuente direccional se maximiza. Esto puede lograrse, por ejemplo, añadiendo las diferentes señales de micrófono con pesos $K_{1,1}$ $K_{1,2}$ $K_{1,3}$ ajustados adecuadamente. Para realizar este trabajo de manera eficaz, las características individuales de retardo, ganancia y fase de las diferentes entradas de micrófono deben analizarse y compensarse de nuevo, lo que también podría entenderse como formación de haces acústicos hacia la fuente direccional.

Los ajustes de ganancia/fase pueden entenderse como una operación de filtrado selectivo en frecuencia. Como tal, los ajustes correspondientes también pueden optimizarse para conseguir reducción de ruido acústico o potenciamiento de las señales de sonido direccionales, por ejemplo siguiendo un enfoque de Wiener.

Como una variación adicional, puede haber un ejemplo con tres canales de MASA. En ese caso, la matriz D de mezcla descendente puede definirse por la matriz de 3 por 3 siguiente:

$$D = \begin{pmatrix} K_{1,1} & K_{1,2} & K_{1,3} \\ K_{2,1} & K_{2,2} & K_{2,3} \\ K_{3,1} & K_{3,2} & K_{3,3} \end{pmatrix}$$

En consecuencia, hay ahora tres señales x_1, x_2, x_3 (en lugar de una en el primer ejemplo) que pueden codificarse con el códec de IVAS.

El primer canal de MASA puede generarse como se describe en el primer ejemplo. El segundo canal de MASA puede usarse para portar un segundo sonido direccional, si hay. Los coeficientes de matriz de mezcla descendente pueden seleccionarse entonces de acuerdo con principios similares a los del primer canal de MASA, sin embargo, de tal manera que la SNR del segundo sonido direccional se maximiza. Los coeficientes de matriz de mezcla descendente $K_{3,1}$ $K_{3,2}$ $K_{3,3}$ para el tercer canal de MASA pueden adaptarse para extraer el componente de sonido difuso mientras que se minimizan los sonidos direccionales.

Típicamente, la captura estéreo de fuentes direccionales dominantes en presencia de algún sonido ambiental puede realizarse, como se muestra en la figura 2 y se describió anteriormente. Esto puede producirse frecuentemente en ciertos casos de uso, por ejemplo en telefonía. De acuerdo con las diversas realizaciones descritas en el presente documento, parámetros de metadatos también se determinan junto con la mezcla descendente, paso 104, que posteriormente se agregará y propagará junto con la única señal de audio de mezcla descendente mono.

En una realización, tres parámetros de metadatos principales están asociados con cada señal de audio capturada: un valor de retardo de tiempo relativo, un valor de ganancia y un valor de fase. De acuerdo con un enfoque general, el canal de MASA se obtiene de acuerdo con las siguientes operaciones:

- Ajuste de retardo de cada señal de micrófono m_i (i = 1, 2) en una cantidad $\tau_i = \Delta\tau_i + \tau_{ref}$.

- Ajuste de ganancia y fase de cada componente/baldosa de tiempo frecuencia (TF) de cada señal de micrófono ajustada en retardo mediante un parámetro de ajuste de ganancia y fase, a y φ respectivamente.

El término de ajuste de retardo τ_i en la expresión anterior puede interpretarse como un tiempo de llegada de una onda sonora plana desde la dirección de la fuente direccional y, como tal, también se expresa convenientemente como tiempo de llegada con respecto al tiempo de llegada de la onda sonora en un punto de referencia τ_{ref}, tal como el centro geométrico del dispositivo de captura de audio 202, aunque podría usarse cualquier punto de referencia. Por ejemplo, cuando se usan dos micrófonos, el ajuste de retardo puede formularse como la diferencia entre τ_1 y τ_2, lo que equivale a mover el punto de referencia a la posición del segundo micrófono. En una realización, el parámetro de tiempo de llegada permite modelar tiempos de llegada relativos en un intervalo de [-2,0 ms, 2,0 ms], que corresponde a un desplazamiento máximo de un micrófono con respecto al origen de aproximadamente 68 cm.

En cuanto a los ajustes de ganancia y fase, en una realización se parametrizan para cada baldosa de TF, de tal manera que pueden modelarse cambios de ganancia en el rango [+10 dB, -30 dB], mientras que pueden representarse cambios de fase en el rango [-Pi, +Pi].

En el caso fundamental con una única fuente direccional dominante, tal como la fuente 206 mostrada en la figura 2, el ajuste del retardo es típicamente constante a través del espectro de frecuencia completo. A medida que la posición de la fuente direccional 206 puede cambiar, los dos parámetros de ajuste de retardo (uno para cada micrófono) variarían con el tiempo. De este modo, los parámetros de ajuste del retardo son dependientes de la señal.

En un caso más complejo, donde puede haber múltiples fuentes 206 de sonido direccional, una fuente desde una primera dirección podría ser dominante en una cierta banda de frecuencia, mientras que una fuente diferente desde otra dirección puede ser dominante en otra banda de frecuencia. En tal escenario, por el contrario, el ajuste de retardo se lleva a cabo ventajosamente para cada banda de frecuencia.

En una realización, esto puede hacerse compensando el retardo de las señales de micrófono en una baldosa de tiempo-frecuencia (TF) dada con respecto a la dirección de sonido que se halla dominante. Si no se detecta ninguna dirección de sonido dominante en la baldosa de TF, no se lleva a cabo ninguna compensación de retardo.

En una realización diferente, las señales de micrófono en una baldosa de TF dado pueden compensarse en retardo con el objetivo de maximizar una relación de señal sobre ruido (SNR) con respecto al sonido direccional, mientras se captura con todos los micrófonos.

En una realización, un límite adecuado de diferentes fuentes para las que se puede hacer una compensación de retardo es tres. Esto ofrece la posibilidad de hacer compensación de retardo en una baldosa de TF bien con respecto a una de las tres fuentes dominantes, o bien con ninguna. El conjunto correspondiente de valores de compensación de retardo (un conjunto que se aplica a todas las señales de micrófono) puede señalizarse de este modo mediante solo dos bits por baldosa de TF. Esto cubre la mayoría de los escenarios de captura relevantes en la práctica y tiene la ventaja de que la cantidad de metadatos o su tasa de bits se mantiene baja.

Otro escenario posible es donde señales ambisónicas de primer orden (FOA) en lugar de señales estéreo se capturan y se mezclan de manera descendente en por ejemplo un único canal de MASA. El concepto de FOA es bien conocido por los expertos en la técnica, pero puede describirse brevemente como un método para grabar, mezclar y reproducir audio tridimensional de 360 grados. El enfoque básico ambisónico es tratar una escena de audio como una esfera completa de 360 grados de sonido que proviene de diferentes direcciones alrededor de un punto central donde se coloca el micrófono mientras se graba, o donde se ubica el "punto dulce" del oyente mientras se reproduce.

La captura de FOA y FOA planaria con mezcla descendente a un único canal de MASA son extensiones relativamente directas del caso de captura estéreo descrito anteriormente. El caso de FOA planaria se caracteriza por un triple micrófono, tal como el mostrado en la figura 2, que hace la captura antes de la mezcla descendente. En el último caso de FOA, la captura se hace con cuatro micrófonos, cuya disposición o selectividades direccionales se extienden en las tres dimensiones espaciales.

Los parámetros de ajuste de compensación de retardo, amplitud y fase pueden usarse para recuperar las tres o, respectivamente, cuatro señales de captura originales y para permitir un renderizado espacial más fiel usando los metadatos de MASA de lo que sería posible simplemente en base a la señal de mezcla descendente mono. Alternativamente, los parámetros de ajuste de compensación de retardo, amplitud y fase pueden usarse para generar una representación de FOA (planaria) más precisa que se acerca a la que se habría capturado con una cuadrícula de micrófono normal.

En otro escenario más, FOA o FOA planaria pueden capturarse y mezclarse de manera descendente en dos o más canales de MASA. Este caso es una ampliación del caso anterior con la diferencia de que las tres o cuatro señales de micrófono capturadas se mezclan de manera descendente a dos en lugar de solo a un único canal de MASA. Se aplican los mismos principios, donde el propósito de proporcionar parámetros de ajuste de compensación de retardo, amplitud y fase es permitir la mejor reconstrucción posible de las señales originales antes de la mezcla descendente.

Como el lector experto se da cuenta, con el fin de acomodar todos estos escenarios de uso, la representación del audio espacial necesitará incluir metadatos sobre no solo el retardo, ganancia y fase, sino también parámetros que son indicativos de la configuración de mezcla descendente para la señal de audio de mezcla descendente.

Volviendo ahora a la figura 1, los parámetros de metadatos determinados se combinan con la señal de audio de mezcla descendente en una representación del audio espacial, paso 108, que finaliza el proceso 100. Lo siguiente es una descripción de cómo estos parámetros de metadatos pueden representarse de acuerdo con una realización de la invención.

Para soportar los casos de uso descritos anteriormente con mezcla descendente a un único o múltiples canales de MASA, se usan dos elementos de metadatos. Un elemento de metadatos son metadatos de configuración independientes de la señal que son indicativos de la mezcla descendente. Este elemento de metadatos se describe a continuación junto con las figuras 3A-3B. El otro elemento de metadatos está asociado con la mezcla descendente. Este elemento de metadatos se describe a continuación junto con las figuras 4-6 y puede determinarse como se describió anteriormente junto con la figura 1. Este elemento se requiere cuando se señaliza la mezcla descendente.

La tabla 1A, mostrada en la figura 3A, es una estructura de metadatos que puede usarse para indicar el número de canales de MASA, desde un único canal de MASA (mono), sobre dos canales de MASA (estéreo) hasta un máximo de cuatro canales de MASA, representados por valores de bit de canal 00, 01, 10 y 11, respectivamente.

La tabla 1B, mostrada en la figura 3B, contiene los valores de bit de canal de la tabla 1A (en este caso particular, solo se muestran los valores de canal “00” y “01” con propósitos ilustrativos), y muestra cómo la configuración de captura de micrófono puede representarse. Por ejemplo, como puede verse en la tabla 1B para un único canal de MASA (mono) se puede señalar si las configuraciones de captura son mono, estéreo, FOA planaria o FOA. Como puede verse además en la tabla 1B, la configuración de captura de micrófono se codifica como un campo de 2 bits (en la columna denominada valor de bit). La tabla 1B también incluye una descripción adicional de los metadatos. Otra configuración independiente de señal puede representar por ejemplo que el audio se originó desde una cuadrícula de micrófonos de un teléfono inteligente o un dispositivo similar.

En el caso en el que los metadatos de mezcla descendente son dependientes de la señal, se necesitan algunos detalles adicionales, como se describirá a continuación. Como se indica en la tabla 1B para el caso específico cuando la señal de transporte es una señal monocanal obtenida a través de mezcla descendente de señales de micrófono múltiple, estos detalles se proporcionan en un campo de metadatos dependientes de señal. La información proporcionada en ese campo de metadatos describe el ajuste de retardo aplicado (con el posible propósito de formación de haces acústicos hacia fuentes direccionales) y filtrado de las señales de micrófono (con el posible propósito de ecualización/supresión de ruido) antes de la mezcla descendente. Esto ofrece información adicional que puede beneficiar la codificación, descodificación y/o renderización.

En una realización, los metadatos de mezcla descendente comprenden cuatro campos, un campo de definición y de selector para señalizar la compensación de retardo aplicada, seguido de dos campos que señalizan los ajustes aplicados de ganancia y fase, respectivamente.

El número de señales n de micrófono de mezcla descendente se señaliza mediante el campo "valor de bit" de la tabla 1B, es decir, n = 2 para mezcla descendente estéreo ("valor de bit = 01"), n = 3 para mezcla descendente de FOA planaria ("valor de bit = 10") y n = 4 para mezcla descendente de FOA ("valor de bit = 11").

Por baldosa de TF pueden definirse y señalizarse hasta tres conjuntos diferentes de valores de compensación de retardo para las hasta n señales de micrófono. Cada conjunto es respectivo de la dirección de una fuente direccional. La definición de los conjuntos de valores de compensación de retardo y la señalización de qué conjunto se aplica a qué baldosa de TF se hace con dos campos separados (de definición y de selector).

En una realización, el campo de definición es una matriz de n x 3 con elementos $B_{i,j}$ de 8 bits que codifica la compensación $\Delta\tau_{i,j}$ de retardo aplicada. Estos parámetros son respectivos del conjunto al que pertenecen, es decir, respectivos de la dirección de una fuente direccional (j = 1... 3). Los elementos $B_{i,j}$ son además respectivos del micrófono de captura (o la señal de captura asociada) (i = 1 ... n, n ≤ 4). Esto se ilustra esquemáticamente en la tabla 2, mostrada en la figura 4.

La figura 4 junto con la figura 3 de este modo muestra una realización donde la representación del audio espacial contiene parámetros de metadatos que están organizados en un campo de definición y un campo de selector. El campo de definición especifica al menos un conjunto de parámetros de compensación de retardo asociado con la pluralidad de micrófonos, y el campo de selector especifica la selección de un conjunto de parámetros de compensación de retardo. Ventajosamente, la representación del valor de retardo de tiempo relativo entre los micrófonos es compacta y de este modo requiere menos velocidad de bits cuando se transmite a un codificador posterior o similar.

El parámetro de compensación de retardo representa un tiempo de llegada relativo de una supuesta onda sonora plana desde la dirección de una fuente en comparación con la llegada de la onda a un punto central geométrico (arbitrario) del dispositivo de captura de audio 202. La codificación de ese parámetro con la palabra clave B de número entero de 8 bits se hace de acuerdo con la siguiente ecuación:

$$\Delta\tau = \frac{B - 128}{128} \cdot 2\,ms, \quad \text{Ecuación n.º (1)}$$

Esto cuantifica el parámetro de retardo relativo linealmente en un intervalo de [-2,0 ms, 2,0 ms], que corresponde a un desplazamiento máximo de un micrófono con relación al origen de aproximadamente 68 cm. Esto es, por supuesto, meramente un ejemplo y otras características y resoluciones de cuantificación también pueden considerarse.

La señalización de qué conjunto de valores de compensación de retardo se aplica a qué baldosa de TF se hace usando un campo de selector que representa las 4*24 baldosas de TF en una trama de 20 ms, que supone 4 subtramas en una trama de 20 ms y 24 bandas de frecuencia. Cada elemento de campo contiene un conjunto de codificación de entrada de 2 bits 1 ... 3 de valores de compensación de retardo con los códigos respectivos '01', '10' y '11'. Se usa una entrada '00' si no se aplica compensación de retardo para la baldosa de TF. Esto se ilustra esquemáticamente en la tabla 3, mostrada en la figura 5.

El ajuste de ganancia se señaliza en 2-4 campos de metadatos, uno para cada micrófono. Cada campo es una matriz de códigos B_a de ajuste de ganancia de 8 bits, respectiva para las 4*24 baldosas de TF en una trama de 20 ms. La codificación de los parámetros de ajuste de ganancia con la palabra clave B_a de número entero se hace de acuerdo con la siguiente ecuación:

$$a = \frac{B_a}{256} \cdot 40 - 30[dB], \quad \text{Ecuación n.º (2)}$$

Los 2-4 campos de metadatos para cada micrófono se organizan como se muestra en la tabla 4, mostrada en la figura 6.

El ajuste de fase se señaliza de manera análoga a ajustes de ganancia en 2-4 campos de metadatos, uno para cada micrófono. Cada campo es una matriz de códigos B_φ de ajuste de fase de 8 bits respectiva para las 4*24 baldosas de TF en una trama de 20 ms. La codificación de los parámetros de ajuste de fase con la palabra clave B_φ de número entero se hace de acuerdo con la siguiente ecuación:

$$\varphi = \frac{B_\varphi}{256} \cdot 2\pi, \quad \text{Ecuación n.º (3)}$$

Los 2-4 campos de metadatos para cada micrófono se organizan como se muestra en la tabla 4 con la única diferencia de que los elementos de campo son las palabras clave B_φ de ajuste de fase.

Esta representación de señales de MASA, que incluyen metadatos asociados puede ser usada entonces por codificadores, descodificadores, renderizadores y otros tipos de equipos de audio que se usan para transmitir, recibir y restaurar fielmente el entorno de sonido espacial grabado. Las técnicas para hacer esto son bien conocidas por los expertos en la técnica, y pueden adaptarse fácilmente para ajustarse a la representación de audio espacial descrita en el presente documento. Por lo tanto, no se considera necesaria una discusión adicional sobre estos dispositivos específicos en este contexto.

Como se entiende por los expertos en la técnica, los elementos de metadatos descritos anteriormente pueden residir o determinarse de diferentes maneras. Por ejemplo, los metadatos pueden determinarse localmente en un dispositivo (tal como un dispositivo de captura de audio, un dispositivo codificador, etc.), pueden derivarse de otro modo de otros datos (por ejemplo, de una nube o de otro modo un servicio remoto), o pueden almacenarse en una tabla de valores predeterminados. Por ejemplo, en base al ajuste de retardo entre micrófonos, el valor de compensación de retardo (figura 4) para un micrófono puede determinarse mediante una tabla de consulta almacenada en el dispositivo de captura de audio, o recibirse desde un dispositivo remoto en base a un cálculo de ajuste de retardo hecho en el dispositivo de captura de audio, o recibirse desde tal dispositivo remoto en base a un cálculo de ajuste de retardo realizado en ese dispositivo remoto (es decir en base a las señales de entrada).

La figura 7 muestra un sistema 700 de acuerdo con una realización ejemplar, en el que los rasgos de la invención descritos anteriormente pueden implementarse. El sistema 700 incluye un dispositivo de captura de audio 202, un codificador 704, un descodificador 706 y un renderizador 708. Los diferentes componentes del sistema 700 pueden comunicarse entre sí a través de una conexión cableada o inalámbrica, o cualquier combinación de las mismas, y los datos se envían típicamente entre las unidades en forma de un flujo de bits. El dispositivo de captura de audio 202 se ha descrito anteriormente y en la figura 2, y está configurado para capturar audio espacial que es una combinación de sonido direccional y sonido difuso. El dispositivo de captura de audio 202 crea una señal de audio de mezcla descendente de canal individual o múltiple mediante mezcla descendente de señales de audio de entrada procedentes de una pluralidad de micrófonos en una unidad de captura de audio que captura el audio espacial. A continuación, el dispositivo de captura de audio 202 determina primeros parámetros de metadatos asociados con la señal de audio de mezcla descendente. Esto se ejemplificará adicionalmente a continuación junto con la figura 8. Los primeros parámetros de metadatos son indicativos de un valor de retardo de tiempo relativo, un valor de ganancia, y/o un valor de fase asociados con cada señal de audio de entrada. El dispositivo de captura de audio 202 combina finalmente la señal de audio de mezcla descendente y los primeros parámetros de metadatos en una representación del audio espacial. Debe apreciarse que mientras que, en la realización actual, toda la captura y combinación de audio se hace en el dispositivo 202 de captura de audio, también puede haber realizaciones alternativas, en las que ciertas porciones de las operaciones de creación, determinación y combinación se producen en el codificador 704.

El codificador 704 recibe la representación de audio espacial desde el dispositivo de captura de audio 202. Es decir, el codificador 704 recibe un formato de datos que comprende una señal de audio de mezcla descendente de canal individual o múltiple que resultade una mezcla descendente de señales de audio de entrada procedentes de una pluralidad de micrófonos en una unidad de captura de audio que captura el audio espacial, y primeros parámetros de metadatos indicativos de una configuración de mezcla descendente para las señales de audio de entrada, un valor de retardo de tiempo relativo, un valor de ganancia, y/o un valor de fase asociados con cada señal de audio de entrada. Debe apreciarse que el formato de datos puede almacenarse en una memoria no transitoria antes/después de ser recibido por el codificador. El codificador 704 codifica entonces la señal de audio de mezcla descendente de canal individual o múltiple en un flujo de bits usando los primeros metadatos. En algunas realizaciones, el codificador 704 puede ser un codificador de IVAS, como se ha descrito anteriormente, pero como aprecia el experto en la técnica, otros tipos de codificadores 704 pueden tener capacidades similares y su uso también es posible.

El flujo de bits codificado, que es indicativo de la representación codificada del audio espacial, es recibido entonces por el descodificador 706. El descodificador 706 descodifica el flujo de bits en una aproximación del audio espacial, usando los parámetros de metadatos que se incluyen en el flujo de bits del codificador 704. Finalmente, el renderizador 708 recibe la representación descodificada del audio espacial y renderiza el audio espacial usando los metadatos, para crear una reproducción fiel del audio espacial en el extremo de recepción, por ejemplo por medio de uno o más altavoces.

La figura 8 muestra un dispositivo de captura de audio 202 de acuerdo con algunas realizaciones. El dispositivo de captura de audio 202 en algunas realizaciones puede comprender una memoria 802 con tablas de consulta almacenadas para determinar los primeros y segundos metadatos. El dispositivo de captura de audio 202 puede en algunas realizaciones estar conectado a un dispositivo remoto 804 (que puede estar ubicado en la nube o ser un dispositivo físico conectado al dispositivo de captura de audio 202) que comprende o puede comprender una memoria 806 con tablas de consulta almacenadas para determinar los primeros y segundos metadatos. El dispositivo de captura de audio puede en algunas realizaciones hacer cálculos/procesamiento necesarios (por ejemplo usando un procesador 803) para por ejemplo determinar el valor de retardo de tiempo relativo, un valor de ganancia, y un valor de fase asociados con cada señal de audio de entrada y transmitir tales parámetros al dispositivo remoto para recibir los primeros y segundos metadatos desde este dispositivo. En otras realizaciones, el dispositivo de captura de audio 202 está transmitiendo las señales de entrada al dispositivo remoto 804 que hace los cálculos/procesamiento necesarios (por ejemplo usando un procesador 805) y determina los primeros y segundos metadatos para transmisión de vuelta al dispositivo de captura de audio 202. En aún otra realización, el dispositivo remoto 804 que hace los cálculos/procesamiento necesarios, transmite parámetros de vuelta al dispositivo de captura de audio 202 que determina los primeros y segundos metadatos localmente en base a los parámetros recibidos (por ejemplo mediante el uso de la memoria 806 con tablas de consulta almacenadas).

La figura 9 muestra un descodificador 706 y renderizador 708 (cada uno que comprende un procesador 910, 912 para realizar diversos procesamientos, por ejemplo descodificación, renderización, etc.) de acuerdo con las realizaciones. El descodificador y renderizador pueden ser en dispositivos separados o en un mismo dispositivo. El procesador o procesadores 910, 912 pueden compartirse entre el descodificador y el renderizador o ser procesadores separados. De manera similar a lo que se describe junto con la figura 8, la interpretación de los primeros y segundos metadatos puede hacerse usando una tabla de consulta almacenada en una memoria 902 en el descodificador 706, una memoria 904 en el renderizador 708, o una memoria 906 en un dispositivo remoto 905 (que comprende un procesador 908) conectado al descodificador o al renderizador.

Equivalentes, extensiones, alternativas y miscelánea

Realizaciones adicionales de la presente divulgación resultarán evidentes para un experto en la técnica después de estudiar la descripción anterior. Aunque la presente descripción y dibujos divulgan realizaciones y ejemplos, la divulgación no se restringe a estos ejemplos específicos. Numerosas modificaciones y variaciones pueden hacerse sin salirse del alcance de la presente divulgación, que se define por las reivindicaciones que se acompañan. Cualquier signo de referencia que aparezca en las reivindicaciones no debe entenderse como limitante de su alcance.

Además, variaciones a las realizaciones divulgadas pueden entenderse y efectuarse por el experto en poner en práctica la divulgación, a partir de un estudio de los dibujos, la divulgación, y las reivindicaciones adjuntas. En las reivindicaciones, la palabra "comprender" no excluye otros elementos o pasos, y el artículo indefinido "un" o "una" no excluye una pluralidad. El mero hecho de que ciertas medidas se citan en reivindicaciones dependientes diferentes entre sí no indica que una combinación de estas medidas no pueda usarse ventajosamente.

Los sistemas y métodos descritos anteriormente en el presente documento pueden implementarse como software, firmware, hardware o una combinación de los mismos. En una implementación de hardware, la división de tareas entre unidades funcionales a las que se hace referencia en la descripción anterior no corresponde necesariamente a la división en unidades físicas; por el contrario, un componente físico puede tener múltiples funcionalidades, y una tarea puede llevarse a cabo por varios componentes físicos en cooperación. Ciertos componentes o todos los componentes pueden implementarse como software ejecutado por un procesador de señal digital o microprocesador, o pueden implementarse como hardware o como un circuito integrado de aplicación específica. Tal software puede distribuirse en medios legibles por ordenador, que pueden comprender medios de almacenamiento de ordenador (o medios no transitorios) y medios de comunicación (o medios transitorios). Como es bien conocido por un experto en la técnica, el término medios de almacenamiento de ordenador incluye medios tanto volátiles como no volátiles, extraíbles y no extraíbles implementados en cualquier método o tecnología para el almacenamiento de información tal como instrucciones legibles por ordenador, estructuras de datos, módulos de programa u otros datos. Los medios de almacenamiento de ordenador incluyen, pero no se limitan a, RAM, ROM, EEPROM, memoria flash u otra tecnología de memoria, CD-ROM, discos versátiles digitales (DVD) u otro almacenamiento en disco óptico, casetes magnéticos, cinta magnética, almacenamiento en disco magnético, u otros dispositivos de almacenamiento magnético, o cualquier otro medio que pueda usarse para almacenar la información deseada y al que pueda accederse mediante un ordenador. Además, es bien conocido por el experto que los medios de comunicación típicamente incorporan instrucciones legibles por ordenador, estructuras de datos, módulos de programa u otros datos en una señal de datos modulada tal como una onda portadora u otro mecanismo de transporte e incluye cualquier medio de entrega de información.

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

REIVINDICACIONES

1. Un método para representar audio espacial, siendo el audio espacial una combinación de sonido direccional y sonido difuso, comprendiendo el método:

crear (104) una señal de audio de mezcla descendente de canal individual o múltiple mediante la mezcla descendente de señales de audio de entrada procedentes de una pluralidad de micrófonos (m1, m2, m3) en una unidad de captura de audio que captura el audio espacial;

determinar (106) primeros parámetros de metadatos asociados con la señal de audio de mezcla descendente, en el que los primeros parámetros de metadatos son indicativos de uno o más de: un valor de retardo de tiempo relativo, un valor de ganancia, y un valor de fase asociados con cada señal de audio de entrada; y

combinar (108) la señal de audio de mezcla descendente creada y los primeros parámetros de metadatos en una representación del audio espacial;

caracterizado porque la mezcla descendente para crear una señal x de audio de mezcla descendente de canal individual o múltiple se describe mediante:

$$x = D \cdot m$$

en la que:

D es una matriz de mezcla descendente que contiene coeficientes de mezcla descendente que definen pesos para cada señal de audio de entrada procedente de la pluralidad de micrófonos, y

m es una matriz que representa las señales de audio de entrada procedentes de la pluralidad de micrófonos;

en el que los coeficientes de mezcla descendente se eligen para seleccionar la señal de audio de entrada del micrófono que tiene actualmente la mejor relación de señal sobre ruido con respecto al sonido direccional, y para descartar señales de audio de entrada de señal procedentes de cualesquiera otros micrófonos.

2. El método de la reivindicación 1, en el que combinar la señal de audio de mezcla descendente creada y los primeros parámetros de metadatos en una representación del audio espacial comprende además:

incluir segundos parámetros de metadatos en la representación del audio espacial, siendo los segundos parámetros de metadatos indicativos de una configuración de mezcla descendente para las señales de audio de entrada.

3. El método de la reivindicación 1 o 2, en el que los primeros parámetros de metadatos se determinan para una o más bandas de frecuencia de las señales de audio de entrada de micrófono.

4. El método de la reivindicación 1, en el que la selección se hace para la premisa por baldosa de tiempo-frecuencia (TF) o la selección se hace para todas las bandas de frecuencia de una trama de audio particular.

5. El método de la reivindicación 1, en el que los coeficientes de mezcla descendente se eligen para maximizar la relación de señal sobre ruido con respecto al sonido direccional, cuando se combinan las señales de audio de entrada procedentes de los diferentes micrófonos.

6. El método de la reivindicación 5, en el que la maximización se hace para una banda de frecuencia particular o la maximización se hace para una trama de audio particular.

7. El método de cualquiera de las reivindicaciones 1 a 6, en el que determinar primeros parámetros de metadatos incluye analizar una o más de: características de retardo, ganancia y fase de las señales de audio de entrada procedentes de la pluralidad de micrófonos.

8. Un sistema para representar audio espacial, siendo el audio espacial una combinación de sonido direccional y sonido difuso, que comprende:

un componente de recepción configurado para recibir señales de audio de entrada procedentes de una pluralidad de micrófonos (m1, m2, m3) en una unidad de captura de audio que captura el audio espacial;

un componente de mezcla descendente configurado para crear una señal de audio de mezcla descendente de canal individual o múltiple mediante mezcla descendente de las señales de audio recibidas;

un componente de determinación de metadatos configurado para determinar primeros parámetros de metadatos

asociados con la señal de audio de mezcla descendente, en el que los primeros parámetros de metadatos son indicativos de uno o más de: un valor de retardo de tiempo relativo, un valor de ganancia, y un valor de fase asociados con cada señal de audio de entrada; y

un componente de combinación configurado para combinar la señal de audio de mezcla descendente creada y los primeros parámetros de metadatos en una representación del audio espacial;

caracterizado porque la mezcla descendente para crear una señal x de audio de mezcla descendente de canal individual o múltiple se describe mediante:

$$x = D \cdot m$$

en la que:

D es una matriz de mezcla descendente que contiene coeficientes de mezcla descendente que definen pesos para cada señal de audio de entrada procedente de la pluralidad de micrófonos, y

m es una matriz que representa las señales de audio de entrada procedentes de la pluralidad de micrófonos;

en el que los coeficientes de mezcla descendente se eligen para seleccionar la señal de audio de entrada del micrófono que tiene actualmente la mejor relación de señal sobre ruido con respecto al sonido direccional, y para descartar señales de audio de entrada de señal procedentes de cualesquiera otros micrófonos.

9. Un producto de programa de ordenador que comprende un medio legible por ordenador con instrucciones para realizar el método de una cualquiera de las reivindicaciones 1 a 7.

10. Un codificador (704) configurado para:

recibir una representación de audio espacial, siendo el audio espacial una combinación de sonido direccional y sonido difuso, comprendiendo la representación:

una señal de audio de mezcla descendente de canal individual o múltiple creada mediante mezcla descendente de señales de audio de entrada procedentes de una pluralidad de micrófonos (m1, m2, m3) en una unidad de captura de audio que captura el audio espacial, y

primeros parámetros de metadatos asociados con la señal de audio de mezcla descendente, en el que los primeros parámetros de metadatos son indicativos de uno o más de: un valor de retardo de tiempo relativo, un valor de ganancia, y un valor de fase asociados con cada señal de audio de entrada; y

realizar uno de:

codificar la señal de audio de mezcla descendente de canal individual o múltiple en un flujo de bits usando los primeros metadatos, y

codificar la señal de audio de mezcla descendente de canal individual o múltiple y los primeros metadatos en un flujo de bits;

caracterizado porque la mezcla descendente para crear una señal x de audio de mezcla descendente de canal individual o múltiple se describe mediante:

$$x = D \cdot m$$

en la que:

D es una matriz de mezcla descendente que contiene coeficientes de mezcla descendente que definen pesos para cada señal de audio de entrada procedente de la pluralidad de micrófonos, y

m es una matriz que representa las señales de audio de entrada procedentes de la pluralidad de micrófonos;

en el que los coeficientes de mezcla descendente se eligen para seleccionar la señal de audio de entrada del micrófono que tiene actualmente la mejor relación de señal sobre ruido con respecto al sonido direccional, y para descartar señales de audio de entrada de señal procedentes de cualesquiera otros micrófonos.

11. Un descodificador (706) configurado para:

recibir un flujo de bits indicativo de una representación codificada de audio espacial, siendo el audio espacial una

combinación de sonido direccional y sonido difuso, comprendiendo la representación:

una señal de audio de mezcla descendente de canal individual o múltiple creada mediante mezcla descendente de señales de audio de entrada procedentes de una pluralidad de micrófonos (m1, m2, m3) en una unidad de captura de audio (202) que captura el audio espacial, y

primeros parámetros de metadatos asociados con la señal de audio de mezcla descendente, en el que los primeros parámetros de metadatos son indicativos de uno o más de: un valor de retardo de tiempo relativo, un valor de ganancia, y un valor de fase asociados con cada señal de audio de entrada; y

descodificar el flujo de bits en una aproximación del audio espacial, mediante el uso de los primeros parámetros de metadatos;

caracterizado porque la mezcla descendente para crear una señal x de audio de mezcla descendente de canal individual o múltiple se describe mediante:

$$x = D \cdot m$$

en la que:

D es una matriz de mezcla descendente que contiene coeficientes de mezcla descendente que definen pesos para cada señal de audio de entrada procedente de la pluralidad de micrófonos, y

m es una matriz que representa las señales de audio de entrada procedentes de la pluralidad de micrófonos;

en el que los coeficientes de mezcla descendente se eligen para seleccionar la señal de audio de entrada del micrófono que tiene actualmente la mejor relación de señal sobre ruido con respecto al sonido direccional, y para descartar señales de audio de entrada de señal procedentes de cualesquiera otros micrófonos.

12. Un renderizador (708) configurado para:

recibir una representación de audio espacial, siendo el audio espacial una combinación de sonido direccional y sonido difuso, comprendiendo la representación:

una señal de audio de mezcla descendente de canal individual o múltiple creada mediante mezcla descendente de señales de audio de entrada procedentes de una pluralidad de micrófonos (m1, m2, m3) en una unidad de captura de audio que captura el audio espacial, y

primeros parámetros de metadatos asociados con la señal de audio de mezcla descendente, en el que los primeros parámetros de metadatos son indicativos de uno o más de: un valor de retardo de tiempo relativo, un valor de ganancia, y un valor de fase asociados con cada señal de audio de entrada; y

renderizar el audio espacial usando los primeros metadatos;

caracterizado porque la mezcla descendente para crear una señal x de audio de mezcla descendente de canal individual o múltiple se describe mediante:

$$x = D \cdot m$$

en la que:

D es una matriz de mezcla descendente que contiene coeficientes de mezcla descendente que definen pesos para cada señal de audio de entrada procedente de la pluralidad de micrófonos, y

m es una matriz que representa las señales de audio de entrada procedentes de la pluralidad de micrófonos;

en el que los coeficientes de mezcla descendente se eligen para seleccionar la señal de audio de entrada del micrófono que tiene actualmente la mejor relación de señal sobre ruido con respecto al sonido direccional, y para descartar señales de audio de entrada de señal procedentes de cualesquiera otros micrófonos.

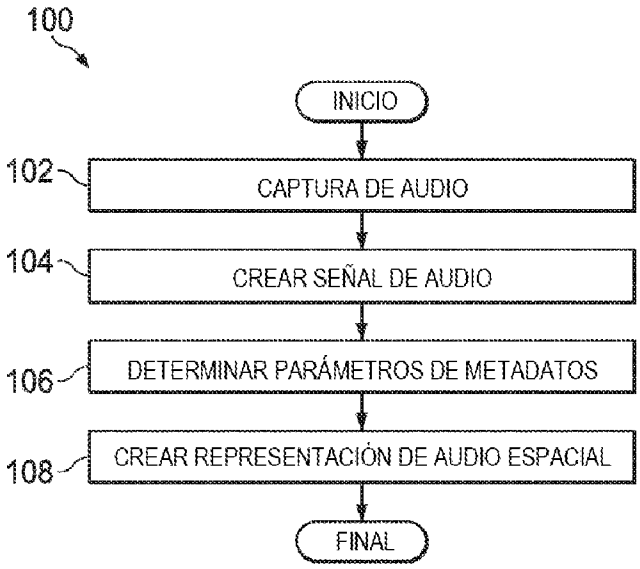
100
INICIO
102
CAPTURA DE AUDIO
104
CREAR SEÑAL DE AUDIO
106
DETERMINAR PARÁMETROS DE METADATOS
108
CREAR REPRESENTACIÓN DE AUDIO ESPACIAL
FINAL

FIG. 1

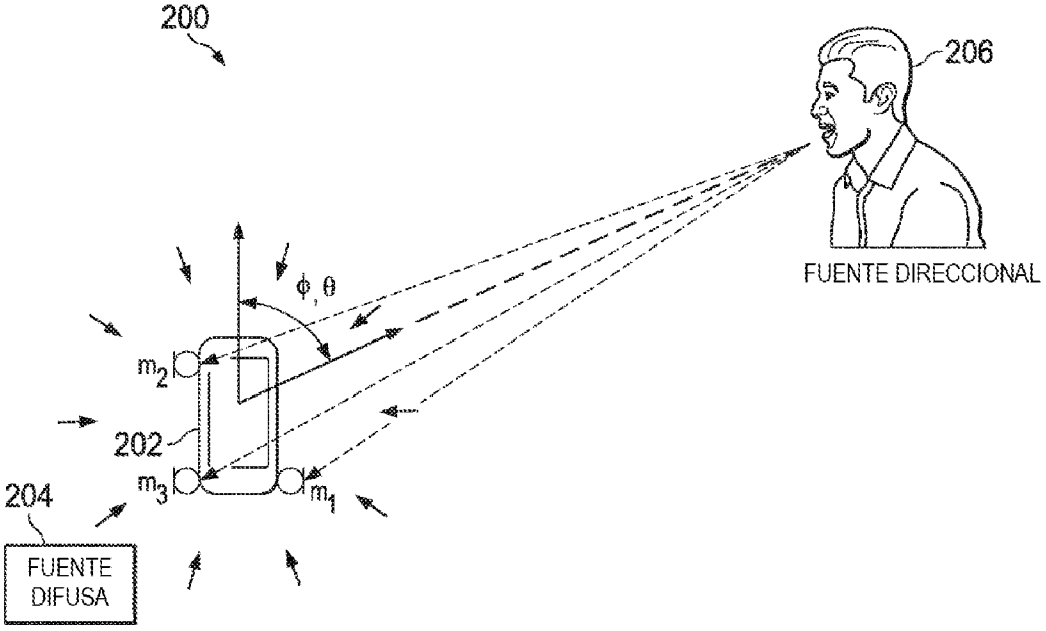
200
206
FUENTE DIRECCIONAL
ϕ, θ
m2
202
m3
m1
204
FUENTE DIFUSA

FIG. 2

TABLA 1A

VALOR DE BIT DE CANAL	VALOR DESCODIFICADO	DESCRIPCIÓN ADICIONAL
00	1 CANAL	
01	2 CANALES	
10	3 CANALES	
11	4 CANALES	

FIG. 3A

TABLA 1B

VALOR DE BIT DE CANAL	VALOR DE BIT	VALOR DESCODIFICADO	DESCRIPCIÓN ADICIONAL
00	00	MONO	
	01	MEZCLA DESCENDENTE ESTÉREO	LA MEZCLA DESCENDENTE SE GENERA A PARTIR DE UNA SEÑAL ESTÉREO I/D. MEZCLA DESCENDENTE RELACIONADA CON METADATOS DEPENDIENTES DE SEÑAL SE SEÑALIZA EN EL CAMPO DE METADATOS DE MEZCLA DESCENDENTE.
	10	MEZCLA DESCENDENTE DE FOA PLANARIA	LA MEZCLA DESCENDENTE SE GENERA A PARTIR DE SEÑALES DE COMPONENTE DE FOA PLANARIA. MEZCLA DESCENDENTE RELACIONADA CON METADATOS DEPENDIENTES DE SEÑAL SE SEÑALIZA EN EL CAMPO DE METADATOS DE MEZCLA DESCENDENTE.
	11	MEZCLA DESCENDENTE DE FOA	LA MEZCLA DESCENDENTE SE GENERA A PARTIR DE SEÑALES DE UN COMPONENTE DE FOA. MEZCLA DESCENDENTE RELACIONADA CON METADATOS DEPENDIENTES DE SEÑAL SE SEÑALIZA EN EL CAMPO DE METADATOS DE MEZCLA DESCENDENTE.
01	00	ESTÉREO I/D	
	01	BINAURAL	
	10	2 MONO (MEZCLADOS)	LAS DOS SEÑALES SE MEZCLAN EN SÍNTESIS. PROVISIÓN PARA ENTRADA DE USUARIO EN RENDERIZADOR.
	11	2 MONO (ALTERNATIVOS)	LAS DOS SEÑALES SON ALTERNATIVAS, Y SOLO SE USA UNA SEÑAL EN SÍNTESIS. SE USA POR DEFECTO LA PRIMERA DE LAS DOS SEÑALES MONO. PROVISIÓN PARA ENTRADA DE USUARIO EN RENDERIZADOR.

FIG. 3B

TABLA 2

		CONJUNTO DE VALORES DE COMPENSACIÓN DE RETARDO		
		1	2	3
MICRÓFONO	1	...	...	...
	2	...	...	...
	3	...	B_{ij}	...
	4	...	...	...

FIG. 4

TABLA 3

		SUBBANDA			
		1	2	○ ○ ○	24
SUBTRAMA	1	...	...	○ ○ ○	...
	2	...	SELECTOR DE 2 BITS	○ ○ ○	...
	3	...	...	○ ○ ○	...
	4	...	...	○ ○ ○	...

FIG. 5

TABLA 4

MICRÓFONO 1		SUBBANDA			
		1	2	○ ○ ○	24
SUBTRAMA	1	...	...	○ ○ ○	...
	2	...	B_a	○ ○ ○	...
	3	...	...	○ ○ ○	...
	4	...	...	○ ○ ○	...
MICRÓFONO 2		SUBBANDA			
		1	2	○ ○ ○	24
SUBTRAMA	1	...	...	○ ○ ○	...
	2	...	B_a	○ ○ ○	...
	3	...	...	○ ○ ○	...
	4	...	...	○ ○ ○	...
MICRÓFONO 3		SUBBANDA			
		1	2	○ ○ ○	24
SUBTRAMA	1	...	...	○ ○ ○	...
	2	...	B_a	○ ○ ○	...
	3	...	...	○ ○ ○	...
	4	...	...	○ ○ ○	...
MICRÓFONO 4		SUBBANDA			
		1	2	○ ○ ○	24
SUBTRAMA	1	...	...	○ ○ ○	...
	2	...	B_a	○ ○ ○	...
	3	...	...	○ ○ ○	...
	4	...	...	○ ○ ○	...

FIG. 6

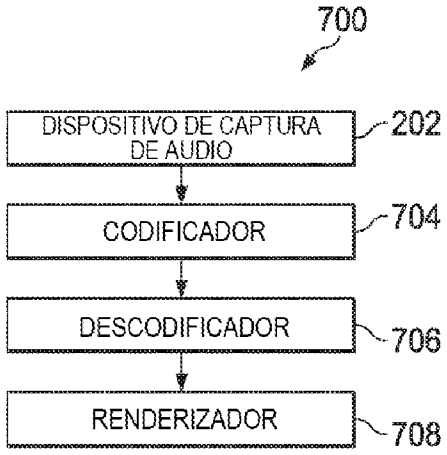

FIG. 7

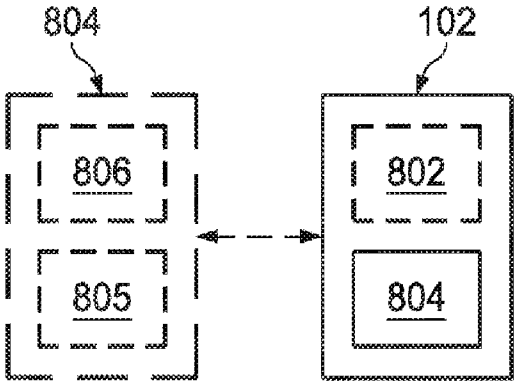

FIG. 8

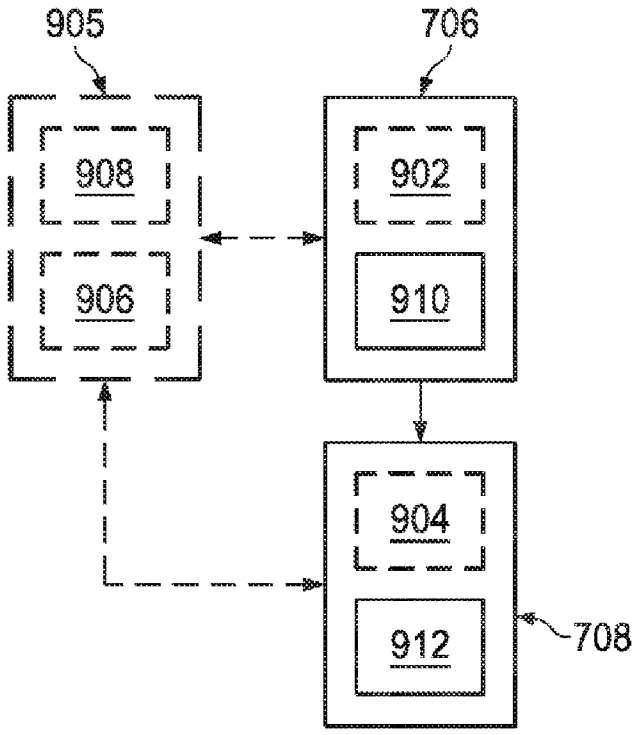

FIG. 9