

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 991 409**

51 Int. Cl.:
G10L 19/008 (2013.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

- 86 Fecha de presentación y número de la solicitud internacional: **29.10.2020** **PCT/US2020/057885**
- 87 Fecha y número de publicación internacional: **06.05.2021** **WO21087063**
- 96 Fecha de presentación y número de la solicitud europea: **29.10.2020** **E 20811838 (0)**
- 97 Fecha y número de publicación de la concesión europea: **02.10.2024** **EP 4052257**

54 Título: **Codificar y decodificar audio multicanal usando metadatos direccionales**

30 Prioridad:

30.10.2019 US 201962927790 P
01.10.2020 US 202063086465 P

45 Fecha de publicación y mención en BOPI de la
traducción de la patente:
03.12.2024

73 Titular/es:

**DOLBY LABORATORIES LICENSING
CORPORATION (100.0%)**
1275 Market Street
San Francisco, CA 94103, US

72 Inventor/es:

MCGRATH, DAVID S.

74 Agente/Representante:

LINAGE GONZÁLEZ, Rafael

ES 2 991 409 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Codificar y decodificar audio multicanal usando metadatos direccionales

5 Referencia cruzada a solicitudes relacionadas

Esta solicitud reivindica prioridad de la solicitud de patente provisional de Estados Unidos n.º 62/927.790, presentada el 30 de octubre de 2019, y la solicitud de patente provisional de Estados Unidos n.º 63/086.465, presentada el 1 de octubre de 2020.

10 Campo técnico

La presente divulgación se refiere generalmente al procesamiento de señales de audio. En particular, la presente divulgación se refiere a métodos de procesamiento de una representación comprimida de una señal de audio espacial para generar una representación reconstruida de la señal de audio espacial.

Antecedentes

La audición humana permite a los oyentes percibir su entorno en forma de una escena de audio espacial -por lo que el término "escena de audio espacial" se usa aquí para referirse al entorno acústico alrededor de un oyente o al entorno acústico percibido en la mente del oyente-.

Mientras que la experiencia humana está ligada a escenas de audio espacial, la técnica de la grabación y reproducción de audio implica la captura, manipulación, transmisión y reproducción de señales de audio o canales de audio. El término "flujo de audio" se usa para referirse a una colección de una o más señales de audio, particularmente donde el flujo de audio está destinado a representar una escena de audio espacial.

Se puede reproducir un flujo de audio para un oyente, a través de transductores electroacústicos o por otros medios, para proporcionar a uno o más oyentes una experiencia auditiva en forma de una escena de audio espacial. Un objetivo común de los profesionales de grabación de audio y los artistas de audio es crear flujos de audio que proporcionen al oyente la experiencia de una escena de audio espacial específica.

Un flujo de audio puede estar acompañado de datos asociados, denominados metadatos, que ayudan en el proceso de reproducción. Los metadatos acompañados pueden incluir información que varía en el tiempo y que puede usarse para efectuar modificaciones en el procesamiento que se aplica durante el proceso de reproducción. En lo sucesivo, el término "experiencia de audio capturada" puede usarse para referirse a un flujo de audio más cualquier metadato asociado.

En algunas aplicaciones, los metadatos constan únicamente de datos indicativos de la disposición prevista de los altavoces para la reproducción. A menudo, estos metadatos se omiten, asumiendo que la disposición de los altavoces durante la reproducción está estandarizada. En este caso, la experiencia de audio capturada consta únicamente de un flujo de audio. Un ejemplo de una de estas experiencias de audio capturadas es un flujo de audio de dos canales, grabada en un disco compacto, donde se supone que el sistema de reproducción previsto es en forma de dos altavoces dispuestos frente al oyente.

Alternativamente, una experiencia de audio capturada en forma de una señal de audio multicanal basada en escena puede estar destinada a ser presentada a un oyente mediante el procesamiento de las señales de audio, a través de una matriz de mezcla, para generar un conjunto de señales de altavoz, cada una de las cuales puede reproducirse posteriormente en un altavoz respectivo, en donde los altavoces pueden estar dispuestos arbitrariamente espacialmente alrededor del oyente. En este ejemplo, la matriz de mezcla se puede generar basándose en el conocimiento previo del formato basado en escena y la disposición de los altavoces de reproducción.

Un ejemplo de un formato basado en escena es la ambisónica de orden superior (HOA), y un método de ejemplo para computar matrices de mezcla adecuadas se proporciona en "Ambisónica", Franz Zotter y Matthias Frank, ISBN: 978-3-030-17206-0, capítulo 3.

Típicamente, estos formatos basados en escena incluyen una gran cantidad de canales u objetos de audio, lo que genera requisitos de ancho de banda o de almacenamiento comparativamente altos al transmitir o almacenar señales de audio espacial en estos formatos.

De este modo, existe una necesidad de representaciones compactas de señales de audio espacial que representen escenas de audio espacial. Esto se aplica tanto a las señales de audio espacial basadas en canal como a las basadas en objeto.

El documento US 2007/269063 A1 describe un marco de codificación de audio espacial en el dominio frecuencia

basado en escena de audio espacial percibida en lugar de en el contenido del canal. En una realización, los vectores de dirección espacial de tiempo-frecuencia se usan como señales para describir la escena de audio de entrada.

El documento WO 2019/086757 describe un método para el procesamiento de señales de audio espacial, que comprende: determinar, para dos o más señales de audio de reproducción, al menos un parámetro de audio espacial para proporcionar una reproducción de audio espacial; determinar entre las dos o más señales de audio de reproducción al menos un parámetro de relación de señal de audio, estando asociado el al menos un parámetro de relación de señal de audio con una determinación de información de relación de señal entre canales entre las dos o más señales de audio de reproducción y para al menos dos bandas de frecuencia, de modo que las dos o más señales de audio de reproducción estén configuradas para ser reproducidas basándose en el al menos un parámetro de audio espacial y el al menos un parámetro de relación de señal de audio.

El documento GB 2571949 A describe el procesamiento de señales de audio espacial donde se recibe al menos una señal de audio antes de determinar al menos un parámetro espacial asociado con la al menos una señal de audio y generar un parámetro de suavizado adaptativo basándose en el al menos un parámetro espacial. A esto le sigue la determinación de ganancias de panoramización para aplicar a una primera parte de la al menos una señal de audio, aplicando el parámetro de suavizado adaptativo a las ganancias de panoramización para generar ganancias de panoramización suavizadas asociadas, antes de aplicar las ganancias de panoramización suavizadas a la primera parte de la al menos una señal de audio para generar una señal de audio posicionada. Preferiblemente, se aplica la decorrelación a una segunda parte de la señal de audio para generar una señal ambiental. Preferiblemente, el parámetro de suavizado adaptativo se basa en un parámetro de suavidad de dirección. Preferiblemente, el parámetro de suavidad de dirección se promedia basándose en la energía de la señal de audio para generar el parámetro de suavizado adaptativo.

Sumario

La presente divulgación propone métodos de procesamiento de una representación comprimida de una señal de audio espacial para generar una representación reconstruida de la señal de audio espacial, correspondientes aparatos, programas y medios de almacenamiento legibles por ordenador.

Partes de la divulgación que no forman parte de la invención se refieren a un método de procesamiento de una señal de audio espacial para generar una representación comprimida de la señal de audio espacial. La señal de audio espacial puede ser una señal multicanal o una señal basada en objeto, por ejemplo. La representación comprimida puede ser una representación compacta o de tamaño reducido. El método puede incluir el análisis de la señal de audio espacial para determinar direcciones de llegada del uno o más elementos de audio en una escena de audio (escena de audio espacial) representada por la señal de audio espacial. Los elementos de audio pueden ser elementos de audio dominantes. Los elementos de audio (dominantes) pueden estar relacionados con objetos acústicos (dominantes), fuentes de sonido (dominantes) o componentes acústicos (dominantes) en la escena de audio, por ejemplo. El uno o más elementos de audio pueden incluir entre uno y diez elementos de audio, como cuatro elementos de audio, por ejemplo. Las direcciones de llegada pueden corresponder a ubicaciones en una esfera unitaria que indican las ubicaciones percibidas de los elementos de audio. El método puede incluir además, para al menos una subbanda de frecuencia (por ejemplo, para todas las subbandas de frecuencia) de la señal de audio espacial, determinar indicaciones respectivas de potencia de señal asociadas con las direcciones de llegada determinadas. El método puede incluir además la generación de metadatos que incluyen información de dirección e información de energía, incluyendo la información de dirección indicaciones de las direcciones de llegada determinadas de uno o más elementos de audio e incluyendo la información de energía indicaciones respectivas de potencia de señal asociada con las direcciones de llegada determinadas. El método puede incluir además la generación de una señal de audio basada en canal con un número predefinido de canales basándose en la señal de audio espacial. La señal de audio basada en canal puede denominarse señal de mezcla de audio o flujo de mezcla de audio. Se entiende que el número de canales de la señal de audio basada en canal puede ser menor que el número de canales o el número de objetos de la señal de audio espacial. El método puede incluir además la salida, como representación comprimida de la señal de audio espacial, de la señal de audio basada en canal y los metadatos. Los metadatos pueden estar relacionados con un flujo de metadatos.

Mediante ello, se puede generar una representación comprimida de una señal de audio espacial que incluya solo un número limitado de canales. Aún así, mediante el uso apropiado de la información de dirección y de la información de energía, un decodificador puede generar una versión reconstruida de la señal de audio espacial original que es una muy buena aproximación de la señal de audio espacial original en lo que respecta a la representación de la escena de audio espacial original.

El análisis de la señal de audio espacial puede basarse en una pluralidad de subbandas de frecuencia de la señal de audio espacial. Por ejemplo, el análisis puede basarse en el rango de frecuencia completo de la señal de audio espacial (es decir, la señal completa). Es decir, el análisis puede basarse en todas las subbandas de frecuencia.

El análisis de la señal de audio espacial puede implicar la aplicación del análisis de escena a la señal de audio espacial. Mediante ello, se pueden determinar de forma fiable y eficiente (las direcciones de) los elementos de audio dominantes en la escena de audio.

La señal de audio espacial puede ser una señal de audio multicanal. Alternativamente, la señal de audio espacial puede ser una señal de audio basada en objeto. En este caso, el método puede incluir además la conversión de la señal de audio basada en objeto en una señal de audio multicanal antes de aplicar el análisis de escena. Esto permite aplicar coherentemente herramientas de análisis de escena a la señal de audio.

Una indicación de la potencia de señal asociada con una dirección de llegada dada puede estar relacionada con una fracción de la potencia de señal en la subbanda de frecuencia para la dirección de llegada dada en relación con la potencia de señal total en la subbanda de frecuencia.

Las indicaciones de potencia de señal pueden determinarse para cada una de una pluralidad de subbandas de frecuencia. En este caso, pueden referirse, para una dirección de llegada dada y una subbanda de frecuencia dada, a una fracción de la potencia de señal en la subbanda de frecuencia dada para la dirección de llegada dada en relación con la potencia de señal total en la subbanda de frecuencia dada. En particular, las indicaciones de potencia de señal se pueden determinar por subbanda, mientras que la determinación de las direcciones (dominantes) de llegada se puede realizar sobre la señal completa (es decir, basándose en todas las subbandas de frecuencia).

El análisis de la señal de audio espacial, la determinación de las indicaciones respectivas de la potencia de señal y la generación de la señal de audio basada en canal se pueden realizar según una premisa de segmento de tiempo. Por consiguiente, la representación comprimida se puede generar y emitir para cada uno de una pluralidad de segmentos de tiempo, con una señal de audio de mezcla descendente y metadatos (bloque de metadatos) para cada segmento de tiempo. Alternativa o adicionalmente, el análisis de la señal de audio espacial, la determinación de las indicaciones respectivas de la potencia de señal y la generación de la señal de audio basada en canal se pueden realizar basándose en una representación de tiempo-frecuencia de la señal de audio espacial. Por ejemplo, los pasos mencionados anteriormente pueden realizarse basándose en una transformada de Fourier discreta (como una STFT, por ejemplo) de la señal de audio espacial. Es decir, para cada segmento de tiempo (bloque de tiempo), los pasos antes mencionados se pueden realizar basándose en los intervalos de tiempo-frecuencia (intervalos FFT) de la señal de audio espacial, es decir, en los coeficientes de Fourier de la señal de audio espacial.

La señal de audio espacial puede ser una señal de audio basada en objeto que incluye una pluralidad de objetos de audio y vectores de dirección asociados. Luego, el método puede incluir además la generación de la señal de audio multicanal mediante la panoramización de los objetos de audio a un conjunto predefinido de canales de audio. Allí, cada objeto de audio puede ser panoramizado hacia el conjunto predefinido de canales de audio de acuerdo con su vector de dirección. Además, la señal de audio basada en canal puede ser una señal de mezcla descendente generada mediante la aplicación de una operación de mezcla descendente a la señal de audio multicanal. La señal de audio multicanal puede ser una señal ambisónica de orden superior, por ejemplo.

La señal de audio espacial puede ser una señal de audio multicanal. Luego, la señal de audio basada en canal puede ser una señal de mezcla descendente generada mediante la aplicación de una operación de mezcla descendente a la señal de audio multicanal.

Un aspecto de la divulgación se refiere a un método de procesamiento de una representación comprimida de una señal de audio espacial para generar una representación reconstruida de la señal de audio espacial. La representación comprimida incluye una señal de audio basada en canal con un número predefinido de canales y metadatos. Los metadatos incluyen información de dirección e información de energía. La información de dirección incluye indicaciones de direcciones de llegada de uno o más elementos de audio en una escena de audio (escena de audio espacial). La información de energía incluye, para al menos una subbanda de frecuencia, indicaciones respectivas de potencia de señal asociadas a las direcciones de llegada. El método incluye generar señales de audio del uno o más elementos de audio basándose en la señal de audio basada en canal, la información de dirección y la información de energía. El método incluye además generar una señal de audio residual de la cual el uno o más elementos de audio están sustancialmente ausentes, basándose en la señal de audio basada en canal, la información de dirección y la información de energía. La señal residual se puede representar en el mismo formato de audio que la señal de audio basada en canal, por ejemplo puede tener el mismo número de canales.

En algunas realizaciones, una indicación de potencia de señal asociada con una dirección de llegada dada puede estar relacionada con una fracción de potencia de señal en la subbanda de frecuencia para la dirección de llegada dada en relación con la potencia de señal total en la subbanda de frecuencia.

En algunas realizaciones, la información de energía puede incluir indicaciones de potencia de señal para cada

una de una pluralidad de subbandas de frecuencia. Entonces, una indicación de potencia de señal puede referirse, para una dirección de llegada dada y una subbanda de frecuencia dada, a una fracción de potencia de señal en la subbanda de frecuencia dada para la dirección de llegada dada en relación con la potencia de señal total en la subbanda de frecuencia dada.

En algunas realizaciones, el método puede incluir además la panoramización de las señales de audio del uno o más elementos de audio a un conjunto de canales de un formato de audio de salida. El método puede incluir además la generación de una señal de audio multicanal reconstruida en el formato de audio de salida basándose en el uno o más elementos de audio panoramizados y la señal residual. El formato de audio de salida puede estar relacionado con una representación de salida, por ejemplo tal como HOA o cualquier otro formato multicanal adecuado. La generación de la señal de audio multicanal reconstruida puede incluir la mezcla ascendente de la señal residual con el conjunto de canales del formato de audio de salida. La generación de la señal de audio multicanal reconstruida puede incluir además la adición de uno o más elementos de audio panoramizados y la señal residual de mezcla ascendente.

De acuerdo con la invención, la generación de señales de audio del uno o más elementos de audio incluye la determinación de coeficientes de una matriz de mezcla inversa M para mapear la señal de audio basada en canal a una representación intermedia que incluye la señal de audio residual y las señales de audio del uno o más elementos de audio, basándose en la información de dirección y la información de energía. La representación intermedia también puede denominarse representación separada o separable, o representación híbrida.

De acuerdo con la invención, la determinación de los coeficientes de la matriz de mezcla inversa M incluye determinar, para cada uno del uno o más elementos de audio, un vector de panoramización $Pan_{down}(dir)$ para panoramizar el elemento de audio a los canales de la señal de audio basada en canal, basándose en la dirección de llegada dir del elemento de audio. Dicha determinación de los coeficientes de la matriz de mezcla inversa M incluye además la determinación de una matriz de mezcla E que se usaría para mapear la señal de audio residual y las señales de audio del uno o más elementos de audio a los canales de la señal de audio basada en canal, basándose en los vectores de panoramización determinados. Dicha determinación de los coeficientes de la matriz de mezcla inversa M incluye además determinar una matriz de covarianza S para la representación intermedia basándose en la información de energía. La determinación de la matriz de covarianza S puede basarse además en los vectores de panoramización determinados Pan_{down} . Dicha determinación de los coeficientes de la matriz de mezcla inversa M incluye además una determinación de los coeficientes de la matriz de mezcla inversa M basándose en la matriz de mezcla E y la matriz de covarianza S .

En algunas realizaciones, la matriz de mezcla E puede determinarse de acuerdo con $E = (I_N | Pan_{down}(dir_1) | \dots | Pan_{down}(dir_P))$. Aquí, I_N puede ser una matriz de identidad $N \times N$, donde N indica el número de canales de la señal basada en canal, $Pan_{down}(dir_p)$ puede ser el vector de panoramización para el p -ésimo elemento de audio con la dirección asociada de llegada dir_p que panoramizaría (por ejemplo, mapearía) el p -ésimo elemento de audio a los N canales de la señal basada en canal, donde $p = 1, \dots, P$ indica uno respectivo de entre el uno o más elementos de audio y P indica el número total del uno o más elementos de audio. Por consiguiente, la matriz E puede ser una matriz $N \times P$. La matriz E puede determinarse para cada uno de una pluralidad de segmentos de tiempo k . En ese caso, la matriz E y las direcciones de llegada dir_p tendrían un índice k que indica el segmento de tiempo, por ejemplo, $E_k = (I_N | Pan_{down}(dir_{k,1}) | \dots | Pan_{down}(dir_{k,P}))$. Aunque el método propuesto puede funcionar en términos de banda, la matriz E puede ser la misma para todas las subbandas de frecuencia.

En algunas realizaciones, la matriz de covarianza S puede determinarse como una matriz diagonal de acuerdo con $\{S\}_{n,n} = rms(Pan_{down})_n(1 - \sum_{p=1}^P e_p)$ para $1 \leq n \leq N$, y $\{S\}_{N+p,N+p} = e_p$ para $1 \leq p \leq P$. Aquí, e_p puede ser la potencia de señal asociada con la dirección de llegada del p -ésimo elemento de audio. La matriz S puede determinarse para cada uno de una pluralidad de segmentos de tiempo k , y/o para cada una de una pluralidad de subbandas de frecuencia b . En ese caso, la matriz S y las potencias de señal e_p tendrían un índice k que indicaría el segmento de tiempo y/o un índice b que indicaría la subbanda de frecuencia, por ejemplo $\{S_{k,b}\}_{n,n} = rms(Pan_{down})_n(1 - \sum_{p=1}^P e_{k,p,b})$ para $1 \leq n \leq N$, y $\{S_{k,b}\}_{N+p,N+p} = e_{k,p,b}$ para $1 \leq p \leq P$.

En algunas realizaciones, la determinación de los coeficientes de la matriz de mezcla inversa M basada en la matriz de mezcla E y la matriz de covarianza S puede implicar la determinación de una pseudoinversa basada en la matriz de mezcla E y la matriz de covarianza S .

En algunas realizaciones, la matriz de mezcla inversa M puede determinarse de acuerdo con $M = S \times E^* \times (E \times S \times E^*)^{-1}$. Aquí, " $\times$ " indica el producto de matriz y " * " indica la transpuesta conjugada de una matriz. La matriz de mezcla inversa M puede determinarse para cada uno de una pluralidad de segmentos de tiempo k , y/o para cada una de una pluralidad de subbandas de frecuencia b . En ese caso, las matrices M y S tendrían un índice k que indica el segmento de tiempo y/o un índice b que indica la subbanda de frecuencia, y la matriz E tendría un índice k que indica el segmento de tiempo, por ejemplo

$$M_{k,b} = S_{k,b} \times E_k^* \times (E_k \times S_{k,b} \times E_k^*)^{-1}.$$

En algunas realizaciones, la señal de audio basada en canal puede ser una señal ambisónica de primer orden.

- 5 Otro aspecto se refiere a un aparato que incluye un procesador y una memoria acoplada al procesador, en donde el procesador está adaptado para llevar a cabo todos los pasos de los métodos de acuerdo con cualquiera de los aspectos y realizaciones mencionados anteriormente.

- 10 Otro aspecto de la divulgación se refiere a un programa que incluye instrucciones que, cuando son ejecutadas por un procesador, hacen que el procesador lleve a cabo todos los pasos de los métodos mencionados anteriormente.

- Otro aspecto adicional de la divulgación se refiere a un medio de almacenamiento legible por ordenador que almacena el programa mencionado anteriormente.

- 15 También se apreciará que las características del aparato y los pasos del método pueden intercambiarse de muchas maneras. En particular, los detalles del método o métodos divulgados se pueden realizar mediante el aparato correspondiente, y viceversa, como apreciará el experto en la técnica. Además, cualquiera de las declaraciones anteriores realizadas con respecto al método o métodos se entiende que se aplican igualmente al
20 aparato correspondiente, y viceversa.

Breve descripción de los dibujos

- 25 Se ilustran realizaciones de ejemplo de la divulgación a modo de ejemplo en los dibujos que se acompañan, en los que números de referencia iguales indican elementos iguales o similares y en los que:

- la figura 1 ilustra esquemáticamente un ejemplo de una disposición de un codificador que genera una representación comprimida de una escena de audio espacial y un decodificador correspondiente para generar una escena de audio reconstituida a partir de la representación comprimida, que no forma parte de la invención,

- 30 la figura 2 ilustra esquemáticamente otro ejemplo de una disposición de un codificador que genera una representación comprimida de una escena de audio espacial y un decodificador correspondiente para generar una escena de audio reconstituida a partir de la representación comprimida, que no forma parte de la invención,

- 35 la figura 3 ilustra esquemáticamente un ejemplo de generación de una representación comprimida de una escena de audio espacial, que no forma parte de la invención,

- la figura 4 ilustra esquemáticamente un ejemplo de decodificación de una representación comprimida de una escena de audio espacial para formar una escena de audio reconstituida, de acuerdo con realizaciones de la divulgación,

- 40 la figura 5 y la figura 6 son diagramas de flujo que ilustran ejemplos de métodos de procesamiento de una escena de audio espacial para generar una representación comprimida de la escena de audio espacial, que no forman parte de la invención,

- 45 las figuras 7 a 11 ilustran esquemáticamente ejemplos de detalles de generación de una representación comprimida de una escena de audio espacial, que no forman parte de la invención,

- 50 la figura 12 ilustra esquemáticamente un ejemplo de detalles de decodificación de una representación comprimida de una escena de audio espacial para formar una escena de audio reconstituida, de acuerdo con realizaciones de la divulgación,

- la figura 13 es un diagrama de flujo que ilustra un ejemplo de un método de decodificación de una representación comprimida de una escena de audio espacial para formar una escena de audio reconstituida, de acuerdo con realizaciones de la divulgación,

- 55 la figura 14 es un diagrama de flujo que ilustra detalles del método de la figura 13,

- 60 la figura 15 es un diagrama de flujo que ilustra otro ejemplo de un método de decodificación de una representación comprimida de una escena de audio espacial para formar una escena de audio reconstituida, que no forma parte de la invención,

la figura 16 ilustra esquemáticamente un aparato para generar una representación comprimida de una escena de audio espacial y/o decodificar la representación comprimida de una escena de audio espacial para formar una

escena de audio reconstituida, de acuerdo con realizaciones de la divulgación.

Descripción detallada

- 5 En general, la presente divulgación se refiere a permitir el almacenamiento y/o la transmisión, usando una cantidad reducida de datos, de una escena de audio espacial.

A continuación se describirán conceptos de procesamiento de audio que pueden usarse en el contexto de la presente divulgación.

10

Funciones de panoramización

- Una señal de audio multicanal (o flujo de audio) se puede formar mediante la panoramización de elementos sonoros individuales (o elementos de audio, objetos de audio) de acuerdo con una ley de mezcla lineal. Por ejemplo, si un conjunto de R objetos de audio están representados por R señales, $\{o_r(t): 1 \leq r \leq R\}$, entonces una mezcla panoramizada multicanal, $\{z_n(t): 1 \leq n \leq N\}$ puede formarse mediante

15

$$\begin{pmatrix} z_1(t) \\ z_2(t) \\ \vdots \\ z_N(t) \end{pmatrix} = \sum_{r=1}^R \text{Pan}(\theta_r) o_r(t) \quad (1)$$

- 20 La función de panoramización, $\text{Pan}(\theta_r)$, representa un vector de columna que contiene N factores de escala (ganancias de panoramización) indicativos de las ganancias que se usan para mezclar la señal del objeto, $o_r(t)$, para formar la salida multicanal, y donde θ_r es indicativo de la ubicación del objeto respectivo.

- Una posible función de panoramización es un panoramizador ambisónico de primer orden (FOA). Un ejemplo de una función de panoramización FOA se da mediante

25

$$\text{Pan}_{\text{FOA}}(x, y, z) = \begin{pmatrix} 1 \\ y \\ z \\ x \end{pmatrix} \quad (2)$$

- Una función de panoramización alternativa es un panoramizador ambisónico de tercer orden (3OA). Un ejemplo de una función de panoramización de 3OA se da mediante

30

$$\text{Pan}_{\text{3OA}}(x, y, z) = \begin{pmatrix} 1 \\ y \\ z \\ x \\ \sqrt{3}xy \\ \sqrt{3}yz \\ \frac{1}{2}(2z^2 - x^2 - y^2) \\ \sqrt{3}xz \\ \frac{\sqrt{3}}{2}(x^2 - y^2) \\ \frac{\sqrt{10}}{4}y(3x^2 - y^2) \\ \sqrt{15}xyz \\ \frac{\sqrt{6}}{4}y(4z^2 - x^2 - y^2) \\ \frac{1}{2}z(2x^2 - 3x^2 - 3y^2) \\ \frac{\sqrt{6}}{4}x(4z^2 - x^2 - y^2) \\ \frac{\sqrt{15}}{2}z(x^2 - y^2) \\ \frac{\sqrt{10}}{4}x(x^2 - 3y^2) \end{pmatrix} \quad (3)$$

Se entiende que la presente divulgación no se limita a las funciones de panoramización FOA o HOA, y que se puede considerar el uso de otras funciones de panoramización, como apreciará el experto en la técnica.

Transformada de Fourier de corto plazo

Un flujo de audio, que consta de una o más señales de audio, puede convertirse en forma de transformada de Fourier de corto plazo (STFT), por ejemplo. Para este fin, se puede aplicar una transformada de Fourier discreta a segmentos de tiempo (opcionalmente en ventanas) de las señales de audio (por ejemplo, canales, señales de objetos de audio) del flujo de audio. Este proceso, aplicado a una señal de audio $x(t)$, puede expresarse de la siguiente manera

$$X_{c,k}(f) = STFT\{x_c(t)\} \quad (4)$$

Se entiende que la STFT es un ejemplo de una transformada de tiempo-frecuencia y que la presente divulgación no se limitará a las STFT.

En la ecuación (4), la variable $X_{c,k}(f)$ indica la transformada de Fourier de corto plazo del canal c ($1 \leq c \leq \text{NumChans}$), para el segmento de tiempo de audio k ($k \in \mathbb{Z}$), en los intervalos de frecuencia f ($1 \leq f \leq F$), donde F indica el número de intervalos de frecuencia producidos por la transformada de Fourier discreta. Se apreciará que la terminología usada aquí es a modo de ejemplo, y que los detalles de implementación específicos de varios métodos STFT (incluidas varias funciones de ventana) pueden conocerse en la técnica. El segmento de tiempo de audio k puede definirse, por ejemplo, como un rango de muestras de audio centradas alrededor de $t = k \times \text{paso} + \text{constante}$, de modo que los segmentos de tiempo estén espaciados uniformemente en el tiempo, con un espaciado igual al paso.

Los valores numéricos de la STFT (como $X_{c,k}(1)$, $X_{c,k}(2)$, ..., $X_{c,k}(F)$) pueden denominarse intervalos FFT.

Además, la forma STFT se puede convertir en un flujo de audio. El flujo de audio resultante puede ser una aproximación a la entrada original y puede estar dado por

$$x'_c(t) = STFT^{-1}\{X_{c,k}(f)\} \approx x_c(t) \quad (5)$$

Análisis de frecuencias en bandas

Los datos característicos pueden formarse a partir de un flujo de audio donde los datos característicos están asociados con un número de bandas de frecuencia (subbandas de frecuencia), donde una banda (subbanda) está definida por una región del rango de frecuencia.

A modo de ejemplo, la potencia de señal en el canal c de un flujo, en la banda de frecuencia b (donde el número de bandas es B y $1 \leq b \leq B$), donde la banda b abarca los intervalos FFT $f_{\min} \leq f \leq f_{\max}$, se puede computar de acuerdo con

$$\text{potencia}_{c,b,k} = \sum_{f=f_{\min}}^{f_{\max}} |X_{c,k}(f)|^2 \quad (6)$$

De acuerdo con un ejemplo más general, la banda de frecuencia b puede definirse mediante un vector de ponderación, $FR_b(f)$, que asigna pesos a cada intervalo de frecuencia, de modo que un cálculo alternativo de la potencia en una banda puede darse mediante

$$\text{potencia}_{c,b,k} = \sum_{f=1}^F FR_b(f) |X_{c,k}(f)|^2 \quad (7)$$

En una generalización adicional de la ecuación (7), la STFT de un flujo que se compone de C señales de audio se puede procesar para producir la covarianza en varias bandas, donde la covarianza, $R_{b,k}$ es una matriz $C \times C$, y

donde el elemento $\{R_{b,k}\}_{i,j}$ se computa de acuerdo con

$$\{R_{b,k}\}_{i,j} = \sum_{f=1}^F FR_b(f) X_{i,k}(f) \overline{X_{j,k}(f)} \quad (8)$$

5 donde $\overline{X_{j,k}(f)}$ representa el conjugado complejo de $X_{j,k}(f)$.

En otro ejemplo, se pueden emplear filtros de paso de banda para formar señales filtradas representativas del flujo de audio original en bandas de frecuencia de acuerdo con las respuestas del filtro de paso de banda. Por ejemplo, una señal de audio $x_c(t)$ puede filtrarse para producir $x'_{c,b}(t)$, que representa una señal con energía derivada predominantemente de la banda b de $x_c(t)$, y por lo tanto, un método alternativo para computar la covarianza de un flujo en la banda b para el bloque de tiempo k (que corresponde a las muestras de tiempo $t_{\min} \leq t \leq t_{\max}$) puede expresarse por

$$\{R_{b,k}\}_{i,j} = \sum_{t=t_{\min}}^{t_{\max}} x'_{i,b}(t) x'_{j,b}(t) \quad (9)$$

15 Mezcla de frecuencias en bandas

Un flujo de audio compuesto de N canales puede procesarse para producir un flujo de audio compuesto de M canales de acuerdo con una matriz de mezcla lineal $M \times N$, Q , de modo que

$$y_m(t) = \sum_{n=1}^N Q_{m,n} x_n(t) \quad (10)$$

que puede escribirse en forma de matriz como

$$\hat{y}(t) = Q \times \hat{x}(t) \quad (11)$$

donde $\hat{x}(t)$ se refiere al vector de columna formado a partir de los N elementos: $x_1(t), x_2(t), \dots, x_N(t)$.

Además, se puede implementar un proceso de mezcla alternativo en el dominio STFT, en donde la matriz, Q , puede tomar valores diferentes en cada bloque de tiempo, k , y en cada banda de frecuencia, b . En este caso, el procesamiento puede considerarse aproximadamente dado por

$$Y_{m,k}(f) = \sum_{b=1}^B \sum_{n=1}^N FR_b(f) Q_{b,k,m,n} X_{n,k}(f) \quad (12)$$

o, en forma de matriz

$$\hat{Y}_k(f) = \sum_{b=1}^B FR_b(f) (Q_{b,k} \times \hat{X}_k(f)) \quad (13)$$

Se apreciará que se pueden emplear métodos alternativos para producir un comportamiento equivalente al procesamiento descrito en la ecuación (13).

Implementaciones de ejemplo

A continuación, se describirán con más detalle implementaciones de ejemplo de métodos y aparatos de acuerdo con realizaciones de la divulgación.

5 En términos generales, los métodos que no forman parte de la invención representan una escena de audio espacial en forma de un flujo de mezcla de audio y un flujo de metadatos de dirección, donde el flujo de metadatos de dirección incluye datos indicativos de la ubicación de elementos sónicos direccionales en la escena de audio espacial y datos indicativos de la potencia de cada elemento sónico direccional, en un número de subbandas, en relación con la potencia total de la escena de audio espacial en esa subbanda. Los métodos de acuerdo con las realizaciones de la divulgación se relacionan con la determinación del flujo de metadatos de dirección a partir de una escena de audio espacial de entrada, y con la creación de una escena de audio reconstituida (por ejemplo, reconstruida) a partir de un flujo de metadatos de dirección y un flujo de mezcla de audio asociado.

15 Los ejemplos de métodos que no forman parte de la invención son eficientes (por ejemplo, en términos de datos reducidos para almacenamiento o transmisión) para representar una escena de sonido espacial. La escena de audio espacial puede representarse mediante una señal de audio espacial. Dichos métodos pueden implementarse definiendo un formato de almacenamiento o transmisión (por ejemplo, el flujo de audio espacial compacto) que consta de un flujo de mezcla de audio y un flujo de metadatos (por ejemplo, flujo de metadatos de dirección).

20 El flujo de mezcla de audio comprende una serie de señales de audio que transmiten una representación reducida de la escena de sonido espacial. Como tal, el flujo de mezcla de audio puede relacionarse con una señal de audio basada en canal con un número predefinido de canales. Se entiende que el número de canales de la señal de audio basada en canal es menor que el número de canales o el número de objetos de audio de la señal de audio espacial. Por ejemplo, la señal de audio basada en canal puede ser una señal de audio ambisónica de primer orden. En otras palabras, el flujo de audio espacial compacto puede incluir un flujo de mezcla de audio en forma de una representación ambisónica de primer orden del campo sonoro.

30 El flujo de metadatos (de dirección) comprende metadatos que definen propiedades espaciales de la escena de sonido espacial. Los metadatos de dirección pueden constar de una secuencia de bloques de metadatos de dirección, donde cada bloque de metadatos de dirección contiene metadatos que indican propiedades de la escena de sonido espacial en un segmento de tiempo correspondiente en el flujo de mezcla de audio.

35 En general, los metadatos incluyen información de dirección e información de energía. La información de dirección incluye indicaciones de direcciones de llegada de uno o más elementos de audio (dominantes) en la escena de audio. La información de energía comprende, para cada dirección de llegada, una indicación de la potencia de señal asociada a las direcciones de llegada determinadas. En algunas implementaciones, las indicaciones de potencia de señal pueden proporcionarse para una, algunas o cada una de una pluralidad de bandas (subbandas de frecuencia). Además, los metadatos pueden proporcionarse para cada uno de una pluralidad de segmentos de tiempo consecutivos, tal como en forma de bloques de metadatos, por ejemplo.

40 En un ejemplo, los metadatos (metadatos de dirección) incluyen metadatos que indican propiedades de la escena de sonido espacial en varias bandas de frecuencia, donde los metadatos definen:

- 45 - una o más direcciones (por ejemplo, direcciones de llegada) indicativas de la ubicación de objetos de audio (elementos de audio) en la escena de sonido espacial, y
- una fracción de energía (o potencia de señal), en cada banda de frecuencia, que se atribuye al objeto de audio respectivo (por ejemplo, atribuida a la dirección respectiva).

A continuación se proporcionarán detalles sobre la determinación de la información de dirección y la información de energía.

55 La figura 1 muestra esquemáticamente un ejemplo de una disposición que no forma parte de la invención. Específicamente, la figura muestra una disposición 100 en la que una escena 10 de audio espacial se introduce en un codificador 200 de escena que genera un flujo 30 de mezcla de audio y un flujo 20 de metadatos de dirección. La escena 10 de audio espacial puede estar representada por una señal de audio espacial o un flujo de audio espacial que se introduce al codificador 200 de escena. El flujo 30 de mezcla de audio y el flujo 20 de metadatos de dirección juntos forman un ejemplo de una escena de audio espacial compacta, es decir, una representación comprimida de la escena 10 de audio espacial (o de la señal de audio espacial).

60 La representación comprimida, es decir, el flujo 30 de audio de mezcla y el flujo 20 de metadatos de dirección, se introduce en el decodificador 300 de escena que produce una escena 50 de audio reconstruida. Los elementos de audio que existen dentro de la escena 10 de audio espacial se representarán dentro del flujo 30 de mezcla de audio de acuerdo con una función de panoramización de mezcla.

La figura 2 muestra esquemáticamente otro ejemplo de una disposición que no forma parte de la invención. Específicamente, la figura muestra una disposición alternativa 110 en donde la escena de audio espacial compacta, compuesta por un flujo 30 de mezcla de audio y un flujo 20 de metadatos de dirección, se codifica además proporcionando el flujo 30 de mezcla de audio a un codificador 35 de audio para producir un flujo 37 de audio codificado de tasa de bits reducida, y proporcionando el flujo 20 de metadatos de dirección a un codificador 25 de metadatos para producir un flujo 27 de metadatos codificados. El flujo 37 de audio codificado a tasa de bits reducida y el flujo 27 de metadatos codificados juntos forman una escena de audio espacial codificada (codificada a tasa de bits reducida).

La escena de audio espacial codificada se puede recuperar aplicando primero el flujo 37 de audio codificado de tasa de bits reducida y el flujo 27 de metadatos codificados a los respectivos decodificadores 36 y 26 para producir un flujo 38 de mezcla de audio recuperado y un flujo 28 de metadatos de dirección recuperado. Los flujos recuperados 38, 28 pueden ser idénticos o aproximadamente iguales a los respectivos flujos 30, 20. El flujo 38 de mezcla de audio recuperado y el flujo 28 de metadatos de dirección recuperados pueden ser decodificados por el decodificador 300 para producir una escena 50 de audio reconstruida.

La figura 3 ilustra esquemáticamente un ejemplo de una disposición para generar un flujo de audio codificado de tasa de bits reducida y un flujo de metadatos codificados a partir de una escena de audio espacial de entrada. Específicamente, la figura muestra una disposición 150 del codificador 200 de escena que proporciona un flujo 20 de metadatos de dirección y un flujo 30 de mezcla de audio a los codificadores respectivos 25, 35 para producir una escena 40 de audio espacial codificada que incluye un flujo 37 de audio codificado de tasa de bits reducida y el flujo 27 de metadatos codificados. El flujo 40 de audio espacial codificado está dispuesto preferiblemente para ser adecuado para el almacenamiento y/o transmisión con un requerimiento de datos reducido, en relación con los datos requeridos para el almacenamiento/transmisión de la escena de audio espacial original.

La figura 4 ilustra esquemáticamente un ejemplo de una disposición para generar una escena de audio espacial reconstruida a partir del flujo de audio codificado de tasa de bits reducida y el flujo de metadatos codificados. Específicamente, la figura muestra una disposición 160 en donde un flujo 40 de audio espacial codificado, compuesto por un flujo 37 de audio codificado de tasa de bits reducida y un flujo 27 de metadatos codificados, se proporciona como entrada a los decodificadores 36, 26 para producir un flujo 38 de mezcla de audio y un flujo 28 de metadatos de dirección, respectivamente. Luego, los flujos 38, 28 son procesados por el decodificador 300 de escena para producir una escena 50 de audio reconstruida.

A continuación se describirán los detalles de la generación de la escena de audio espacial compacta, es decir, la representación comprimida de la escena de audio espacial (o de la señal de audio espacial/flujo de audio espacial).

La figura 5 es un diagrama de flujo de un ejemplo de un método 500 de procesamiento de una señal de audio espacial para generar una representación comprimida de la señal de audio espacial. El método 500 comprende los pasos S510 a S550.

En el paso S510, se analiza la señal de audio espacial para determinar las direcciones de llegada de uno o más elementos de audio (por ejemplo, elementos de audio dominantes) en una escena de audio (escena de audio espacial) representada por la señal de audio espacial. Los elementos de audio (dominantes) pueden estar relacionados con objetos acústicos (dominantes), fuentes de sonido (dominantes) o componentes acústicos (dominantes) en la escena de audio, por ejemplo. El análisis de la señal de audio espacial puede implicar o estar relacionado con la aplicación del análisis de escena a la señal de audio espacial. Se entiende que el experto en la técnica conoce una variedad de herramientas de análisis de escena adecuadas. Las direcciones de llegada determinadas en este paso pueden corresponder a ubicaciones en una esfera unitaria que indican las ubicaciones (percibidas) de los elementos de audio.

En línea con la descripción anterior del análisis de bandas de frecuencia, el análisis de la señal de audio espacial en el paso S510 puede basarse en una pluralidad de subbandas de frecuencia de la señal de audio espacial. Por ejemplo, el análisis puede basarse en el rango de frecuencia completo de la señal de audio espacial (es decir, la señal completa). Es decir, el análisis puede basarse en todas las subbandas de frecuencia.

En el paso S520, se determinan indicaciones respectivas de potencia de señal asociadas con las direcciones de llegada determinadas para al menos una subbanda de frecuencia de la señal de audio espacial.

En el paso S530, se generan metadatos que comprenden información de dirección e información de energía. La información de dirección comprende indicaciones de las direcciones de llegada determinadas del uno o más elementos de audio. La información de energía comprende indicaciones respectivas de la potencia de señal asociada a las direcciones de llegada determinadas. Los metadatos generados en este paso pueden estar relacionados con un flujo de metadatos.

En el paso S540, se genera una señal de audio basada en canal con un número predefinido de canales basándose en la señal de audio espacial.

- 5 Finalmente, en el paso S550, la señal de audio basada en canal y los metadatos se emiten como la representación comprimida de la señal de audio espacial.

Se entiende que los pasos anteriores pueden realizarse en cualquier orden o en paralelo entre sí, siempre que el orden de los pasos garantice que esté disponible la entrada necesaria para cada paso.

- 10 Típicamente, se puede considerar que una escena espacial (o señal de audio espacial) está compuesta por una suma de señales acústicas que inciden sobre un oyente desde un conjunto de direcciones, en relación con la posición de escucha. Por lo tanto, la escena de audio espacial puede modelarse como una colección de R objetos acústicos, donde el objeto r ($1 \leq r \leq R$) está asociado con una señal de audio $o_r(t)$ que incide en la posición de escucha desde una dirección de llegada definida por el vector de dirección θ_r . El vector de dirección también puede ser un vector de dirección variable en el tiempo $\theta_r(t)$.

- 20 Por lo tanto, de acuerdo con algunas implementaciones, la señal de audio espacial (flujo de audio espacial) puede definirse como una señal de audio espacial basada en objeto (escena de audio espacial basada en objeto), en forma de un conjunto de señales de audio y vectores de dirección asociados

$$\text{Escena de audio espacial (basada en objeto)} = \{(o_r(t), \theta_r(t)) : 1 \leq r \leq R\} \quad (14)$$

- 25 Además, de acuerdo con algunas implementaciones, la señal de audio espacial (flujo de audio espacial) puede definirse en términos de señales de transformada de Fourier de corto plazo, $O_{r,k}(f)$, de acuerdo con la ecuación (4), y los vectores de dirección pueden especificarse de acuerdo con el índice de bloque, k, de modo que:

$$\text{Escena de audio espacial (basada en objeto)} = \{(O_{r,k}(f), \theta_r(k)) : 1 \leq r \leq R\} \quad (15)$$

- 30 Alternativamente, la señal de audio espacial (flujo de audio espacial) puede representarse en términos de una señal de audio espacial basada en canal (escena de audio espacial basada en canal). Un flujo basado en canal consta de una colección de señales de audio, en donde cada objeto acústico de la escena de audio espacial se mezcla en los canales de acuerdo con una función de panoramización ($\text{Pan}(\theta)$), de acuerdo con la ecuación (1). A modo de ejemplo, una escena de audio espacial basada en canal Q, $\{C_{q,k}(f) : 1 \leq q \leq Q\}$, puede formarse a partir de una escena de audio espacial basada en objeto de acuerdo con

$$\text{Escena de audio espacial (basada en canal)} = \{C_{q,k}(f) : 1 \leq q \leq Q\}$$

$$\text{donde} \quad \begin{pmatrix} C_{1,k}(f) \\ C_{2,k}(f) \\ \vdots \\ C_{Q,k}(f) \end{pmatrix} = \sum_{r=1}^R \text{Pan}(\theta_r(k)) O_{r,k}(f) \quad (16)$$

- 40 Se apreciará que muchas características de una escena de audio espacial basada en canal están determinadas por la elección de la función de panoramización, y en particular, la longitud (Q) del vector de columna devuelto por la función de panoramización determinará la cantidad de canales de audio contenidos en la escena de audio espacial basada en canal. En términos generales, se puede lograr una representación de mayor calidad de una escena de audio espacial mediante una escena de audio espacial basada en canal que contenga un mayor número de canales.

- 45 Como ejemplo, en el paso S540 del método 500, la señal de audio espacial (escena de audio espacial) se puede procesar para crear una señal de audio basada en canal (flujo basado en canal) de acuerdo con la ecuación (16). La función de panoramización se puede elegir para crear una representación de resolución relativamente baja de la escena de audio espacial. Por ejemplo, la función de panoramización puede elegirse como la función ambisónica de primer orden (FOA), tal como la definida en la ecuación (2). Como tal, la representación comprimida puede ser una representación compacta o de tamaño reducido.

La figura 6 es un diagrama de flujo que proporciona otra formulación de un método 600 para generar una representación compacta de una escena de audio espacial. El método 600 está provisto de un flujo de entrada, en forma de una escena de audio espacial o un flujo basado en escena, y produce una escena de audio espacial compacta como representación compacta. Para su fin, el método 600 comprende los pasos S610 a S660. Allí, el paso S610 puede verse como correspondiente al paso S510, el paso 620 puede verse como correspondiente al paso S520, el paso S630 puede verse como correspondiente al paso S540, el paso S650 puede verse como correspondiente al paso S530 y el paso S660 puede verse como correspondiente al paso S550.

En el paso S610, se analiza el flujo de entrada para determinar las direcciones de llegada dominantes.

En el paso S620, para cada banda (subbanda de frecuencia), se determina una fracción de energía asignada a cada dirección, en relación con una energía total en el flujo en esa banda.

En el paso S630, se forma un flujo de mezcla descendente que contiene una serie de canales de audio que representan la escena de audio espacial.

En el paso S640, el flujo de mezcla descendente se codifica para formar una representación comprimida del flujo.

En el paso S650, la información de dirección y la información de fracción de energía se codifican para formar metadatos codificados.

Finalmente, en el paso S660, el flujo de mezcla descendente codificada se combina con los metadatos codificados para formar una escena de audio espacial compacta.

Se entiende que los pasos anteriores pueden realizarse en cualquier orden o en paralelo entre sí, siempre que el orden de los pasos garantice que esté disponible la entrada necesaria para cada paso.

Las figuras 7 a 11 ilustran esquemáticamente ejemplos de detalles de generación de una representación comprimida de una escena de audio espacial, que no forman parte de la invención.

Se entiende que los detalles de, por ejemplo, analizar la señal de audio espacial para determinar direcciones de llegada, determinar indicaciones de potencia de señal asociadas con las direcciones de llegada determinadas, generar metadatos que comprenden información de dirección e información de energía, y/o generar la señal de audio basada en canal con un número predefinido de canales como se describe a continuación pueden ser independientes de la disposición específica del sistema y pueden aplicarse, por ejemplo, a cualquiera de las disposiciones mostradas en las figuras 7 a 11, o cualquier disposición alternativa adecuada.

La figura 7 ilustra esquemáticamente un primer ejemplo de detalles de generación de la representación comprimida de la escena de audio espacial. Específicamente, la figura 7 muestra un codificador 200 de escena en el que una escena 10 de audio espacial es procesada por una función 203 de mezcla descendente para producir un flujo 30 de mezcla de audio de N canales, de acuerdo con, por ejemplo, los pasos S540 y S630. En algunos ejemplos, la función 203 de mezcla descendente puede incluir el proceso de panoramización de acuerdo con la ecuación (1) o la ecuación (16), en donde se elige una función de panoramización de mezcla

descendente: $Pan(\theta) \approx \underset{down}{Pan}(\theta)$. Por ejemplo, se puede elegir un panoramizador ambisónico de primer orden como la función de panoramización de mezcla descendente: $\underset{down}{Pan}(\theta) \approx \underset{FOA}{Pan}(\theta)$ y, por lo tanto, $N = 4$.

Para cada segmento de tiempo de audio, el análisis 202 de escena toma como entrada la escena de audio espacial y determina las direcciones de llegada de hasta P componentes acústicos dominantes dentro de la escena de audio espacial, de acuerdo con, por ejemplo, los pasos S510 y S610. Los valores típicos para P están entre 1 y 10, y un valor preferido para P es $P \approx 4$. Por consiguiente, el uno o más elementos de audio determinados en el paso S510 pueden comprender entre uno y diez elementos de audio, tales como cuatro elementos de audio, por ejemplo.

El análisis 202 de escena produce un flujo 20 de metadatos compuesto de información 21 de dirección e información 22 de fracción de banda de energía (información de energía). Opcionalmente, el análisis 202 de escena también puede proporcionar coeficientes 207 a la función 203 de mezcla descendente para permitir que se modifique la mezcla descendente. Sin limitación alguna, el análisis de la señal de audio espacial (por ejemplo, en el paso S510), la determinación de las indicaciones respectivas de la potencia de señal (por ejemplo, en el paso S520) y la generación de la señal de audio basada en canal (por ejemplo, en el paso S540) se pueden realizar según una premisa de segmento de tiempo, de acuerdo con, por ejemplo, la descripción anterior de las STFT. Esto implica que la representación comprimida se generará y se emitirá para cada uno de una pluralidad de segmentos de tiempo, con una señal de audio de mezcla descendente y metadatos (bloque de metadatos) para cada segmento de tiempo.

Para cada segmento de tiempo, k , la información 21 de dirección (por ejemplo, incorporada por las direcciones de llegada del uno o más elementos de audio) puede tomar la forma de P vectores de dirección, $\{dir_{k,p}: 1 \leq p \leq P\}$. El vector de dirección p indica la dirección asociada con el índice de objeto dominante p , y puede representarse en términos de vectores unitarios, donde:

$$\begin{aligned} dir_{k,p} &= (x_{k,p}, y_{k,p}, z_{k,p}) \\ x_{k,p}^2 + y_{k,p}^2 + z_{k,p}^2 &\approx 1 \end{aligned} \quad (17)$$

o en términos de coordenadas esféricas, donde:

$$\begin{aligned} dir_{k,p} &= (az_{k,p}, el_{k,p}) \\ -180 \leq az_{k,p} \leq 180 \text{ y } -90 \leq el_{k,p} \leq 90 \end{aligned} \quad (18)$$

En algunos ejemplos, las indicaciones respectivas de potencia de señal determinadas en el paso S520 toman la forma de una fracción de potencia de señal. Es decir, una indicación de potencia de señal asociada con una dirección de llegada dada en la subbanda de frecuencia se relaciona con una fracción de potencia de señal en la subbanda de frecuencia para la dirección de llegada dada en relación con la potencia de señal total en la subbanda de frecuencia.

Además, en algunos ejemplos, las indicaciones de potencia de señal se determinan para cada una de una pluralidad de subbandas de frecuencia (es decir, por subbanda). Luego, se relacionan, para una dirección de llegada dada y una subbanda de frecuencia dada, con una fracción de la potencia de señal en la subbanda de frecuencia dada para la dirección de llegada dada en relación con la potencia de señal total en la subbanda de frecuencia dada. En particular, si bien las indicaciones de potencia de señal se pueden determinar por subbanda, la determinación de las direcciones de llegada (dominantes) todavía se puede realizar sobre la señal completa (es decir, basándose en todas las subbandas de frecuencia).

Aún además, en algunos ejemplos, el análisis de la señal de audio espacial (por ejemplo, en el paso S510), la determinación de las indicaciones respectivas de la potencia de señal (por ejemplo, en el paso S520) y la generación de la señal de audio basada en canal (por ejemplo, en el paso S540) se realizan basándose en una representación de tiempo-frecuencia de la señal de audio espacial. Por ejemplo, los pasos antes mencionados y otros pasos según sea adecuado se pueden realizar basándose en una transformada de Fourier discreta (como una STFT, por ejemplo) de la señal de audio espacial. Por ejemplo, para cada segmento de tiempo (bloque de tiempo), los pasos antes mencionados se pueden realizar basándose en los intervalos de tiempo-frecuencia (intervalos FFT) de la señal de audio espacial, es decir, en los coeficientes de Fourier de la señal de audio espacial.

Dado lo anterior, para cada segmento de tiempo, k , y para cada índice de objeto dominante p ($1 \leq p \leq P$), la información 22 de fracción de banda de energía puede incluir un valor de fracción $e_{k,p,b}$ para cada banda b de un conjunto de bandas ($1 \leq b \leq B$). El valor de fracción $e_{k,p,b}$ se determina para el segmento de tiempo k de acuerdo con:

$$e_{k,p,b} = \frac{\text{Energía en dirección } dir_{k,p} \text{ para banda } b}{\text{Energía total en escena para banda } b} \quad (19)$$

El valor de fracción $e_{k,p,b}$ puede representar la fracción de energía en una región espacial alrededor de la dirección $dir_{k,p}$, de modo que la energía de múltiples objetos acústicos en la escena de audio espacial original se puede combinar para representar un único componente acústico dominante asignado a la dirección $dir_{k,p}$. En algunos ejemplos, la energía de todos los objetos acústicos de la escena se puede ponderar, usando una función de ponderación de diferencia angular $w(\theta)$ que representa una ponderación mayor para una dirección, θ , que está cerca de $dir_{k,p}$, y una ponderación menor para una dirección, θ , que está lejos de $dir_{k,p}$. Las diferencias direccionales pueden considerarse cercanas para diferencias angulares menores a , por ejemplo, 10° , y lejanas para diferencias angulares mayores a , por ejemplo, 45° . En ejemplos alternativos, la función de ponderación puede elegirse basándose en opciones alternativas de las diferencias angulares cercanas/lejanas.

En general, la señal de audio espacial de entrada para la que se genera la representación comprimida puede ser una señal de audio multicanal o una señal de audio basada en objeto, por ejemplo. En el último caso, el método para generar la representación comprimida de la señal de audio espacial comprendería además un paso de

conversión de la señal de audio basada en objeto en una señal de audio multicanal antes de aplicar el análisis de escena (por ejemplo, antes del paso S510).

En el ejemplo de la figura 7, la señal de audio espacial de entrada puede ser una señal de audio multicanal. Luego, la señal de audio basada en canal generada en el paso S540 sería una señal de mezcla descendente generada al aplicar una operación de mezcla descendente a la señal de audio multicanal.

La figura 8 ilustra esquemáticamente otro ejemplo de detalles de la generación de la representación comprimida de la escena de audio espacial. En este caso, la señal de audio espacial de entrada puede ser una señal de audio basada en objeto que comprende una pluralidad de objetos de audio y vectores de dirección asociados. En este caso, el método de generación de la representación comprimida de la señal de audio espacial comprende generar una señal de audio multicanal, como una representación intermedia o escena intermedia, mediante la panoramización de los objetos de audio hacia un conjunto predefinido de canales de audio, en donde cada objeto de audio es panoramizado hacia el conjunto predefinido de canales de audio de acuerdo con su vector de dirección. De este modo, la figura 8 muestra un ejemplo alternativo de un codificador 200 de escena en donde la escena 10 de audio espacial se introduce en un convertidor 201 que produce la escena intermedia 11 (por ejemplo, incorporada por la señal multicanal). La escena intermedia 11 se puede crear de acuerdo con la ecuación (1), donde la función de panoramización se selecciona de modo que el producto escalar de los vectores de ganancia de panoramización $Pan(\theta_1)$ y $Pan(\theta_2)$ represente aproximadamente una función de ponderación de diferencia angular, como se describió anteriormente.

En algunos ejemplos, la función de panoramización usada en el convertidor 201 es una función de panoramización ambisónico de tercer orden, $Pan_{3OA}(\theta)$, como se muestra en la ecuación (3). Por consiguiente, la señal de audio multicanal puede ser, por ejemplo, una señal ambisónica de orden superior.

La escena intermedia 11 se introduce luego en el análisis 202 de escena. El análisis 202 de escena puede determinar las direcciones, $dir_{k,p}$, de los objetos acústicos dominantes en la escena de audio espacial a partir del análisis de la escena intermedia 11. La determinación de las direcciones dominantes se puede realizar estimando la energía en un conjunto de direcciones, siendo la energía estimada más grande la dirección dominante.

La información 22 de fracción de banda de energía para el segmento de tiempo k puede incluir un valor de fracción $e_{k,p,b}$ para cada banda b que se deriva de la energía en la banda b de la escena intermedia 11 en cada dirección $dir_{k,p}$, en relación con la energía total en la banda b de la escena intermedia 11 en el segmento de tiempo k .

El flujo 30 de mezcla de audio (por ejemplo, señal de audio basada en canal) de la escena de audio espacial compacta (por ejemplo, representación compacta) en este caso es una señal de mezcla descendente generada mediante la aplicación de la función 203 de mezcla descendente (operación de mezcla descendente) a la escena de audio espacial.

La figura 10 muestra una disposición alternativa de un codificador de escena que incluye un convertidor 201 para convertir la escena 10 de audio espacial en un formato intermedio 11 basado en escena. El formato intermedio 11 se introduce en el análisis 202 de escena y en la función 203 de mezcla descendente. En algunos ejemplos, la función 203 de mezcla descendente puede incluir un mezclador de matriz con coeficientes adaptados para convertir el formato intermedio 11 en el flujo 30 de mezcla de audio. Es decir, el flujo 30 de mezcla de audio (por ejemplo, señal de audio basada en canal) de la escena de audio espacial compacta (por ejemplo, representación compacta) en este caso puede ser una señal de mezcla descendente generada mediante la aplicación de la función 203 de mezcla descendente (operación de mezcla descendente) a la escena intermedia (por ejemplo, señal de audio multicanal).

En un ejemplo alternativo, mostrado en la figura 11, el codificador espacial 200 puede tomar una entrada en forma de una entrada 11 basada en escena, en donde los objetos acústicos se representan de acuerdo con una regla de panoramización, $Pan(\theta)$. En algunos ejemplos, la función de panoramización puede ser una función de panoramización ambisónica de orden superior. En un ejemplo, la función de panoramización es una función de panoramización ambisónica de tercer orden.

En otro ejemplo alternativo, ilustrado en la figura 9, una escena 10 de audio espacial es convertida por el convertidor 201 en el codificador espacial 200 para producir una escena intermedia 11 que se introduce en la función 203 de mezcla descendente. El análisis 202 de escena se proporciona con información de la escena 10 de audio espacial.

La figura 12 ilustra esquemáticamente un ejemplo de detalles de decodificación de una representación comprimida de una escena de audio espacial para formar una escena de audio reconstituida, de acuerdo con realizaciones de la divulgación. Específicamente, la figura muestra un decodificador 300 de escena que incluye

un desmezclador 302 que toma un flujo 30 de mezcla de audio y produce un flujo 70 de audio espacial separado. El flujo 70 de audio espacial separado está compuesto por P señales 90 de objeto dominantes y un flujo residual 80. El decodificador residual 81 toma la entrada del flujo residual 80 y crea un flujo residual decodificado 82. El panoramizador 91 de objetos toma la entrada de las señales 90 de objetos dominantes y crea un flujo de objetos panoramizados 92. El flujo residual decodificado 82 y el flujo 92 de objeto panoramizado se suman 75 para producir la escena 50 de audio reconstituida.

Además, la figura 12 muestra la entrada de información 21 de dirección e información 22 de fracción de banda de energía a una calculadora 301 de matriz de desmezcla que determina una matriz 60 de desmezcla (matriz de mezcla inversa) que será usada por el desmezclador 302.

A continuación se describirán los detalles del procesamiento de la escena de audio espacial compacta (por ejemplo, la representación comprimida de la señal de audio espacial) para generar la representación reconstruida de la señal de audio espacial.

La figura 13 es un diagrama de flujo de un ejemplo de un método 1300 de procesamiento de una representación comprimida de una señal de audio espacial para generar una representación reconstruida de la señal de audio espacial. Se entiende que la representación comprimida comprende una señal de audio basada en canal (por ejemplo, incorporada por el flujo 30 de mezcla de audio) con un número predefinido de canales y metadatos, comprendiendo los metadatos información de dirección (por ejemplo, incorporada por la información 21 de dirección) e información de energía (por ejemplo, incorporada por la información 22 de fracción de banda de energía), comprendiendo la información de dirección indicaciones de direcciones de llegada de uno o más elementos de audio en una escena de audio y comprendiendo la información de energía, para al menos una subbanda de frecuencia, indicaciones respectivas de potencia de señal asociada con las direcciones de llegada. La señal de audio basada en canal puede ser una señal ambisónica de primer orden, por ejemplo. El método 1300 comprende los pasos S1310 y S1320 y, opcionalmente, los pasos S1330 y S1340. Se entiende que estos pasos pueden ser realizados por el decodificador 300 de escena de la figura 12, por ejemplo.

En el paso S1310, se generan señales de audio del uno o más elementos de audio basándose en la señal de audio basada en canal, la información de dirección y la información de energía.

En el paso S1320, se genera una señal de audio residual de la cual el uno o más elementos de audio están sustancialmente ausentes, basándose en la señal de audio basada en canal, la información de dirección y la información de energía. Aquí, la señal residual se puede representar en el mismo formato de audio que la señal de audio basada en canal, por ejemplo, puede tener el mismo número de canales que la señal de audio basada en canal.

En el paso opcional S1330, las señales de audio del uno o más elementos de audio son panoramizados a un conjunto de canales de un formato de audio de salida. Aquí, el formato de audio de salida puede estar relacionado con una representación de salida, por ejemplo, como HOA o cualquier otro formato multicanal adecuado.

En el paso opcional S1340, se genera una señal de audio multicanal reconstruida en el formato de audio de salida basándose en el uno o más elementos de audio panoramizados y la señal residual. La generación de la señal de audio multicanal reconstruida puede incluir la mezcla ascendente de la señal residual con el conjunto de canales del formato de audio de salida. La generación de la señal de audio multicanal reconstruida puede incluir además la adición de uno o más elementos de audio panoramizados y la señal residual de mezcla ascendente.

Se entiende que los pasos anteriores pueden realizarse en cualquier orden o en paralelo entre sí, siempre que el orden de los pasos garantice que esté disponible la entrada necesaria para cada paso.

En línea con la descripción anterior de los métodos de procesamiento de la escena de audio espacial para generar la representación comprimida de la escena de audio espacial, una indicación de potencia de señal asociada con una dirección de llegada dada puede relacionarse con una fracción de potencia de señal en la subbanda de frecuencia para la dirección de llegada dada en relación con la potencia de señal total en la subbanda de frecuencia.

Además, en algunas realizaciones, la información de energía puede incluir indicaciones de potencia de señal para cada una de una pluralidad de subbandas de frecuencia. Luego, una indicación de potencia de señal puede relacionarse, para una dirección de llegada dada y una subbanda de frecuencia dada, con una fracción de potencia de señal en la subbanda de frecuencia dada para la dirección de llegada dada en relación con la potencia de señal total en la subbanda de frecuencia dada.

La generación de señales de audio del uno o más elementos de audio en el paso S1310 puede comprender la determinación de coeficientes de una matriz de mezcla inversa M para mapear la señal de audio basada en canal a una representación intermedia que comprende la señal de audio residual y las señales de audio del uno

o más elementos de audio, basándose en la información de dirección y la información de energía. La representación intermedia también puede denominarse representación separada o separable, o representación híbrida.

5 A continuación se describirán los detalles de dicha determinación de los coeficientes de la matriz de mezcla inversa M con referencia al diagrama de flujo de la figura 14. El método 1400 ilustrado por este diagrama de flujo comprende los pasos S1410 a S1440.

10 En el paso S1410, para cada uno del uno o más elementos de audio, se determina un vector de panoramización $Pan_{down}(dir)$ para panoramizar el elemento de audio a los canales de la señal de audio basada en canal, basándose en la dirección de llegada dir del elemento de audio.

15 En el paso S1420, se determina una matriz de mezcla E que se usaría para mapear la señal de audio residual y las señales de audio del uno o más elementos de audio a los canales de la señal de audio basada en canal, basándose en los vectores de panoramización determinados.

20 En el paso S1430, se determina una matriz de covarianza S para la representación intermedia basándose en la información de energía. La determinación de la matriz de covarianza S puede basarse además en los vectores de panoramización determinados Pan_{down} .

Finalmente, en el paso S1440, los coeficientes de la matriz de mezcla inversa M se determinan basándose en la matriz de mezcla E y la matriz de covarianza S.

25 Se entiende que los pasos anteriores pueden realizarse en cualquier orden o en paralelo entre sí, siempre que el orden de los pasos garantice que esté disponible la entrada necesaria para cada paso.

Volviendo a la figura 12, la calculadora 301 de matriz de desmezcla computa la matriz 60 de desmezcla (matriz de mezcla inversa), $M_{k,b}$ de acuerdo con un proceso que incluye los siguientes pasos:

30 1. Las entradas a la calculadora de matriz de desmezcla, para el segmento de tiempo k, son la información de dirección, $dir_{k,p}$ ($1 \leq p \leq P$), y la información de fracción de banda de energía, $e_{k,p,b}$ ($1 \leq p \leq P$ y $1 \leq b \leq B$). P representa el número de componentes acústicos dominantes y B indica el número de bandas de frecuencia.

35 2. Para cada banda, b, la matriz de desmezcla $M_{k,b}$ se computa de acuerdo con:

$$M = S \times E^* \times (E \times S \times E^*)^{-1} \quad (20)$$

40 donde "x" indica el producto de matrices y "*" indica la transpuesta conjugada de una matriz. El cálculo de acuerdo con la ecuación (20) puede corresponder al paso S1440, por ejemplo.

45 La matriz de desmezcla M puede determinarse para cada uno de una pluralidad de segmentos de tiempo k, y/o para cada una de una pluralidad de subbandas de frecuencia b. En ese caso, las matrices M y S tendrían un índice k que indica el segmento de tiempo y/o un índice b que indica la subbanda de frecuencia, y la matriz E tendría un índice k que indica el segmento de tiempo, por ejemplo

$$M_{k,b} = S_{k,b} \times E_k^* \times (E_k \times S_{k,b} \times E_k^*)^{-1} \quad (20a)$$

50 En general, la determinación de los coeficientes de la matriz de mezcla inversa M basada en la matriz de mezcla E y la matriz de covarianza S puede implicar la determinación de una pseudoinversa basada en la matriz de mezcla E y la matriz de covarianza S. Un ejemplo de dicha pseudoinversa se da en las ecuaciones (20) y (20a).

En la ecuación (20), la matriz E_k (matriz de mezcla) se forma apilando juntas una matriz (I_N) de identidad $N \times N$ y las P columnas formadas por la función de panoramización aplicada a las direcciones de cada uno de los P componentes acústicos dominantes:

$$E = (I_N [Pan_{down}(dir_1)] \dots [Pan_{down}(dir_P)]) \quad (21)$$

60 En la ecuación (21), I_N es una matriz de identidad $N \times N$, donde N indica el número de canales de la señal basada en canal. $Pan_{down}(dir_p)$ es el vector de panoramización para el p-ésimo elemento de audio con la dirección de llegada asociada dir_p que panoramizaría el p-ésimo elemento de audio hacia los N canales de la

señal basada en canal, con $p = 1, \dots, P$ indicando uno respectivo de entre el uno o más elementos de audio y P indicando el número total del uno o más elementos de audio. Las barras verticales en la ecuación (21) indican una operación de aumento de matriz. Por consiguiente, la matriz E es una matriz $N \times P$.

- 5 Además, la matriz E puede determinarse para cada uno de una pluralidad de segmentos de tiempo k . En ese caso, la matriz E y las direcciones de llegada dir_p tendrían un índice k que indica el segmento de tiempo, por ejemplo

$$E_k = (I_N | Pan_{down}(dir_{k,1}) | \dots | Pan_{down}(dir_{k,P})) \quad (21a)$$

- 10 Si el método propuesto funciona en bandas, la matriz E puede ser la misma para todas las subbandas de frecuencia.

- 15 De acuerdo con el paso S1420, la matriz E_k es la matriz de mezcla que se usaría para mapear la señal de audio residual y las señales de audio del uno o más elementos de audio a los canales de la señal de audio basada en canal. Como se puede ver en las ecuaciones (21) y (21a), la matriz E_k se basa en los vectores de panoramización $Pan_{down}(dir)$ determinados en el paso S1410.

- 20 En la ecuación (20), la matriz S es una matriz diagonal $(N + P) \times (N + P)$. Puede verse como una matriz de covarianza para la representación intermedia. Sus coeficientes se pueden calcular basándose en la información de energía, de acuerdo con el paso S1430. Los primeros N elementos diagonales están dados por

$$\{S\}_{n,n} = rms(Pan_{down})_n \left(1 - \sum_{p=1}^P e_p \right) \quad (22)$$

- 25 para $1 \leq n \leq N$, y los P elementos diagonales restantes están dados por

$$\{S\}_{N+p,N+p} = e_p \quad (23)$$

- 30 para $1 \leq p \leq P$, donde e_p es la potencia de señal asociada con la dirección de llegada del p -ésimo elemento de audio.

- La matriz de covarianza S puede determinarse para cada uno de una pluralidad de segmentos de tiempo k , y/o para cada una de una pluralidad de subbandas de frecuencia b . En ese caso, la matriz de covarianza S y las potencias de señal e_p tendrían un índice k que indica el segmento de tiempo y/o un índice b que indica la subbanda de frecuencia. Los primeros N elementos diagonales vendrían dados por

$$\{S_{k,b}\}_{n,n} = rms(Pan_{down})_n \left(1 - \sum_{p=1}^P e_{k,p,b} \right) \quad (1 \leq n \leq N) \quad (22a)$$

- 40 y los P elementos diagonales restantes vendrían dados por

$$\{S_{k,b}\}_{N+p,N+p} = e_{k,p,b} \quad (1 \leq p \leq P) \quad (23a)$$

- En una realización preferida, la matriz de desmezcla $M_{k,b}$ se aplica, por el desmezclador 302, para producir un flujo 70 de audio espacial separado (como un ejemplo de la representación intermedia), de acuerdo con la implementación descrita anteriormente del paso S1310, en donde los primeros N canales son el flujo residual 80 y los P canales restantes representan los componentes acústicos dominantes.

- El flujo espacial 70 separado del canal $N + P$, $Y_k(f)$, las P señales 90 de objeto dominantes de canal (como ejemplos de las señales de audio del uno o más elementos de audio generados en el paso S1310), $O_k(f)$, y el flujo residual 80 del canal N (como un ejemplo de la señal de audio residual generada en el paso S1320), $R_k(f)$,

se computan a partir de la mezcla 30 de audio del canal N, $X_k(f)$, de acuerdo con:

$$\begin{aligned} Y_k(f) &= \sum_{b=1}^B FR_b(f) M_{k,b} \times X_k \\ R_k(f) &= \{Y_k(f)\}_{1..N} \\ O_k(f) &= \{Y_k(f)\}_{N+1..N+P} \end{aligned} \quad (24)$$

donde las señales se representan en forma de STFT, la expresión $\{Y_k(f)\}_{1..N}$ indica una señal de canal N formada a partir de los canales 1..N de $Y_k(f)$, y $\{Y_k(f)\}_{N+1..N+P}$ indica una señal de canal P formada a partir de los canales N+1..N+P de $Y_k(f)$. Los expertos en la técnica apreciarán que la aplicación de la matriz $M_{k,b}$ puede lograrse de acuerdo con métodos alternativos, conocidos en la técnica, que proporcionan una función aproximada equivalente a la de la ecuación (24).

Además de lo anterior, en algunas realizaciones, el número de componentes acústicos dominantes P puede adaptarse para tomar un valor diferente para cada segmento de tiempo, de modo que P_k puede depender del índice del segmento de tiempo, k. Por ejemplo, el análisis 202 de escena en el codificador 200 de escena puede determinar un valor de P_k para cada segmento de tiempo. En general, el número de componentes acústicos dominantes P puede depender del tiempo. La elección de P (o P_k) puede incluir un equilibrio entre la tasa de datos de los metadatos y la calidad de la escena de audio reconstruida.

Volviendo a la figura 12, el decodificador espacial 300 produce una escena 50 de audio reconstituida de canal M en donde el flujo de canal M está asociado con un panoramizador de salida $Pan_{out}(\theta)$. Esto puede realizarse de acuerdo con el paso S1340 descrito anteriormente. Los ejemplos de panoramizadores de salida incluyen funciones de panoramización estéreo, funciones de panoramización de amplitud basadas en vector, como se conocen en la técnica, y funciones de panoramización ambisónica de orden superior, como se conocen en la técnica.

Por ejemplo, el panoramizador 91 de objetos en la figura 12 puede estar adaptado para crear el flujo 92 de objetos panoramizados de canal M, Z_p , de acuerdo con

$$Z_p(f) = \sum_{p=1}^P Pan_{out}(\theta_p) O_{k,p}(f) \quad (25)$$

La figura 15 es un diagrama de flujo que proporciona una alternativa, que no forma parte de la invención, de un método 1500 de decodificación de una escena de audio espacial compacta para producir una escena de audio reconstituida. El método 1500 comprende los pasos S1510 a S1580.

En el paso S1510, se recibe una escena de audio espacial compacta y se extraen el flujo de mezcla descendente codificada y el flujo de metadatos codificados.

En el paso S1520, el flujo de mezcla descendente codificada se decodifica para formar un flujo de mezcla descendente.

En el paso S1530, el flujo de metadatos codificados se decodifica para formar la información de dirección y la información de fracción de energía.

En el paso S1540, se forma una matriz de desmezcla por banda a partir de la información de dirección y la información de fracción de energía.

En el paso S1550, el flujo de mezcla descendente se procesa de acuerdo con la matriz de desmezcla para formar un flujo separado.

En el paso S1560, las señales de objeto son extraídas del flujo separado y son panoramizadas para producir señales de objeto panoramizadas de acuerdo con la información de dirección y un formato de salida deseado.

En el paso S1570, se extraen señales residuales del flujo separado y se procesan para crear señales residuales decodificadas de acuerdo con el formato de salida deseado.

Finalmente, en el paso S1580, las señales de objeto panoramizadas y las señales residuales decodificadas se combinan para formar una escena de audio reconstituida.

Se entiende que los pasos anteriores pueden realizarse en cualquier orden o en paralelo entre sí, siempre que el orden de los pasos garantice que esté disponible la entrada necesaria para cada paso.

5 Se han descrito anteriormente métodos de procesamiento de una señal de audio espacial para generar una representación comprimida de la señal de audio espacial, así como métodos de procesamiento de una representación comprimida de una señal de audio espacial para generar una representación reconstruida de la señal de audio espacial. Además, la presente divulgación también se refiere a un aparato para llevar a cabo estos métodos. Un ejemplo de dicho aparato 1600 se ilustra esquemáticamente en la figura 16. El aparato 1600
10 puede comprender un procesador 1610 (por ejemplo, una unidad central de procesamiento (CPU), una unidad de procesamiento de gráficos (GPU), un procesador de señal digital (DSP), uno o más circuitos integrados de aplicación específica (ASIC), uno o más circuitos integrados de radiofrecuencia (RFIC), o cualquier combinación de estos) y una memoria 1620 acoplada al procesador 1610. El procesador puede estar adaptado para llevar a cabo algunos o todos los pasos de los métodos descritos a lo largo de la divulgación. Si el aparato 1600 actúa
15 como un codificador (por ejemplo, codificador de escena), puede recibir, como entrada 1630, la señal de audio espacial (es decir, la escena de audio espacial), por ejemplo. El aparato 1600 puede entonces generar, como salida 1640, la representación comprimida de la señal de audio espacial. Si el aparato 1600 actúa como un decodificador (por ejemplo, un decodificador de escena), puede recibir, como entrada 1630, la representación comprimida. El aparato puede entonces generar, como salida 1640, la escena de audio reconstituida.

20 El aparato 1600 puede ser un ordenador servidor, un ordenador cliente, un ordenador personal (PC), una tableta PC, un decodificador (STB), un asistente digital personal (PDA), un teléfono celular, un teléfono inteligente, un dispositivo web, un enrutador, conmutador o puente de red, o cualquier máquina capaz de ejecutar instrucciones (secuenciales o de otro tipo) que especifiquen acciones que debe realizar ese aparato. Además, aunque en la
25 figura 16 solo se ilustra un único aparato 1600, la presente divulgación se relacionará con cualquier colección de aparatos que ejecutan individual o conjuntamente instrucciones para realizar una o más de las metodologías discutidas en el presente documento.

30 La presente divulgación se refiere además a un programa (por ejemplo, un programa de ordenador) que comprende instrucciones que, cuando son ejecutadas por un procesador, hacen que el procesador lleve a cabo algunos o todos los pasos de los métodos descritos en el presente documento.

Aún además, la presente divulgación se refiere a un medio de almacenamiento legible por ordenador (o legible por máquina) que almacena el programa mencionado anteriormente. Aquí, el término "medio de almacenamiento legible por ordenador" incluye, entre otros, repositorios de datos en forma de memorias de estado sólido, medios
35 ópticos y medios magnéticos, por ejemplo.

Consideraciones de configuración adicionales

40 A menos que se indique específicamente lo contrario, como se desprende de las siguientes discusiones, se aprecia que a lo largo de la divulgación las discusiones que utilizan términos como "procesar", "computar", "calcular", "determinar", "analizar" o similares, se refieren a la acción y/o procesos de un ordenador o sistema informático, o dispositivos informáticos electrónicos similares, que manipulan y/o transforman datos representados como cantidades físicas, tales como electrónicas, en otros datos representados de manera similar
45 como cantidades físicas.

De manera similar, el término "procesador" puede referirse a cualquier dispositivo o porción de un dispositivo que procesa datos electrónicos, por ejemplo, de registros y/o memoria para transformar esos datos electrónicos en otros datos electrónicos que, por ejemplo, pueden almacenarse en registros y/o memoria. Un "ordenador", una
50 "máquina de computación" o una "plataforma informática" pueden incluir uno o más procesadores.

Las metodologías descritas en el presente documento son, en una realización de ejemplo, ejecutables por uno o más procesadores que aceptan código legible por ordenador (también llamado legible por máquina) que contiene un conjunto de instrucciones que, cuando son ejecutadas por uno o más de los procesadores, llevan a cabo al
55 menos uno de los métodos descritos en el presente documento. Se incluye cualquier procesador capaz de ejecutar un conjunto de instrucciones (secuenciales o de otro tipo) que especifiquen acciones a realizar. De este modo, un ejemplo es un sistema de procesamiento típico que incluye uno o más procesadores. Cada procesador puede incluir uno o más de los siguientes: una CPU, una unidad de procesamiento gráfico y una unidad DSP programable. El sistema de procesamiento puede incluir además un subsistema de memoria que incluye una
60 RAM principal y/o una RAM estática y/o una ROM. Puede incluirse un subsistema de bus para la comunicación entre los componentes. El sistema de procesamiento también puede ser un sistema de procesamiento distribuido con procesadores acoplados por una red. Si el sistema de procesamiento requiere una pantalla, tal pantalla puede incluirse, por ejemplo una pantalla de cristal líquido (LCD) o una pantalla de tubo de rayos catódicos (CRT). Si se requiere la entrada manual de datos, el sistema de procesamiento también incluye un dispositivo de
65 entrada tal como una o más unidades de entrada alfanumérica tales como un teclado, un dispositivo de control señalador tal como un ratón, etc. El sistema de procesamiento también puede incluir un sistema de

almacenamiento tal como una unidad de disco. El sistema de procesamiento en algunas configuraciones puede incluir un dispositivo de salida de sonido y un dispositivo de interfaz de red. El subsistema de memoria incluye de este modo un medio portador legible por ordenador que lleva un código legible por ordenador (por ejemplo, software) que incluye un conjunto de instrucciones para provocar la ejecución, cuando son ejecutadas por uno o más procesadores, de uno o más de los métodos descritos en el presente documento. Téngase en cuenta que, cuando el método incluye varios elementos, por ejemplo varios pasos, no se implica ningún orden de dichos elementos, a menos que se indique específicamente. El software puede residir en el disco duro, o también puede residir, total o al menos parcialmente, en la memoria RAM y/o en el procesador durante su ejecución por el sistema de ordenador. De este modo, la memoria y el procesador también constituyen un medio portador legible por ordenador que transporta un código legible por ordenador. Además, un medio portador legible por ordenador puede formar o estar incluido en un producto de programa de ordenador.

En realizaciones de ejemplo alternativas, el uno o más procesadores funcionan como un dispositivo independiente o pueden estar conectados, por ejemplo en red con otro procesador o procesadores, en un despliegue en red, el uno o más procesadores pueden funcionar en la capacidad de un servidor o una máquina de usuario en un entorno de red servidor-usuario, o como una máquina par en un entorno de red distribuido o de igual a igual. El uno o más procesadores pueden formar un ordenador personal (PC), una tableta PC, un asistente digital personal (PDA), un teléfono celular, un dispositivo web, un enrutador, conmutador o puente de red, o cualquier máquina capaz de ejecutar un conjunto de instrucciones (secuenciales o de otro tipo) que especifiquen acciones que debe realizar esa máquina.

Téngase en cuenta que el término "máquina" también se considerará que incluye cualquier colección de máquinas que individual o conjuntamente ejecutan un conjunto (o múltiples conjuntos) de instrucciones para realizar una o más de las metodologías discutidas en el presente documento.

De este modo, una realización de ejemplo de cada uno de los métodos descritos en el presente documento es en forma de un medio portador legible por ordenador que lleva un conjunto de instrucciones, por ejemplo un programa de ordenador que se va a ejecutar en uno o más procesadores, por ejemplo uno o más procesadores que son parte de una disposición de servidor web. De este modo, como apreciarán los expertos en la técnica, las realizaciones de ejemplo de la presente divulgación pueden realizarse como un método, un aparato tal como un aparato de propósito especial, un aparato tal como un sistema de procesamiento de datos o un medio portador legible por ordenador, por ejemplo un producto de programa de ordenador. El medio portador legible por ordenador transporta un código legible por ordenador que incluye un conjunto de instrucciones que, cuando se ejecutan en uno o más procesadores, hacen que el procesador o los procesadores implementen un método. Por consiguiente, los aspectos de la presente divulgación pueden tomar la forma de un método, una realización de ejemplo completamente de hardware, una realización de ejemplo completamente de software o una realización de ejemplo que combina aspectos de software y hardware. Además, la presente divulgación puede tomar la forma de un medio portador (por ejemplo, un producto de programa de ordenador en un medio de almacenamiento legible por ordenador) que lleva un código de programa legible por ordenador incorporado en el medio.

El software también puede transmitirse o recibirse a través de una red mediante un dispositivo de interfaz de red. Si bien en una realización de ejemplo el medio portador es un medio único, el término "medio portador" debe entenderse como que incluye un medio único o múltiples medios (por ejemplo, una base de datos centralizada o distribuida, y/o cachés y servidores asociados) que almacenan el uno o más conjuntos de instrucciones. El término "medio portador" también se entenderá que incluye cualquier medio que sea capaz de almacenar, codificar o transportar un conjunto de instrucciones para su ejecución por uno o más de los procesadores y que haga que el uno o más procesadores realicen una o más de las metodologías de la presente divulgación. Un medio portador puede adoptar muchas formas, incluidas, entre otras, medios no volátiles, medios volátiles y medios de transmisión. Los medios no volátiles incluyen, por ejemplo, discos ópticos, magnéticos y discos magneto-ópticos. Los medios volátiles incluyen la memoria dinámica, como la memoria principal. Los medios de transmisión incluyen cables coaxiales, cables de cobre y fibra óptica, incluidos los cables que componen un subsistema de bus. Los medios de transmisión también pueden tomar la forma de ondas acústicas o de luz, como las que se generan durante las comunicaciones de datos por ondas de radio e infrarrojos. Por ejemplo, el término "medio portador" se entenderá que incluye, entre otros, memorias de estado sólido, un producto de ordenador incorporado en medios ópticos y magnéticos; un medio que lleva una señal propagada detectable por al menos un procesador o uno o más procesadores y que representa un conjunto de instrucciones que, cuando se ejecutan, implementan un método; y un medio de transmisión en una red que lleva una señal propagada detectable por al menos un procesador del uno o más procesadores y que representa el conjunto de instrucciones.

Se entenderá que los pasos de los métodos discutidos se llevan a cabo en una realización de ejemplo mediante un procesador (o procesadores) apropiados de un sistema de procesamiento (por ejemplo, un ordenador) que ejecuta instrucciones (código legible por ordenador) almacenadas en la memoria. También se entenderá que la divulgación no está limitada a ninguna técnica de implementación o programación particular y que la divulgación puede implementarse usando cualquier técnica apropiada para implementar la funcionalidad descrita en el

presente documento. La divulgación no se limita a ningún lenguaje de programación o sistema operativo en particular.

5 La referencia a lo largo de esta divulgación a "una realización de ejemplo", "algunas realizaciones de ejemplo" o "una realización de ejemplo" significa que una característica, estructura o característica particular descrita en relación con la realización de ejemplo está incluida en al menos una realización de ejemplo de la presente divulgación. De este modo, las apariciones de las frases "en una realización de ejemplo", "en algunas realizaciones de ejemplo" o "en una realización de ejemplo" en diversos lugares a lo largo de esta divulgación no necesariamente se refieren todas a la misma realización de ejemplo. Además, las características, estructuras o
10 características particulares se pueden combinar de cualquier manera adecuada, como sería evidente para una persona con conocimientos ordinarios en la técnica a partir de esta divulgación, en una o más realizaciones de ejemplo.

15 Como se usa en el presente documento, a menos que se especifique lo contrario, el uso de los adjetivos ordinales "primero", "segundo", "tercero", etc., para describir un objeto común, simplemente indica que se hace referencia a diferentes instancias de objetos similares y no pretende implicar que los objetos así descritos deben estar en una secuencia dada, ya sea temporalmente, espacialmente, en clasificación o de cualquier otra manera.

20 En las reivindicaciones a continuación y en la descripción incluida en el presente documento, cualquiera de los términos comprendiendo, comprendido por o que comprende es un término abierto que significa incluir al menos los elementos/características que siguen, pero sin excluir otros. De este modo, el término que comprende, cuando se usa en las reivindicaciones, no debe interpretarse como limitativo de los medios, elementos o pasos enumerados a continuación. Por ejemplo, el alcance de la expresión un dispositivo que comprende A y B no debe limitarse a dispositivos que constan únicamente de los elementos A y B. Cualquiera de los términos
25 incluyendo o que incluye o que incluye como se usan en el presente documento también es un término abierto que también significa incluir al menos los elementos/características que siguen al término, pero sin excluir otros. De este modo, incluir es sinónimo de y significa comprender.

30 Se debe apreciar que, en la descripción anterior de las realizaciones de ejemplo de la divulgación, varias características de la divulgación a veces se agrupan juntas en una sola realización, figura o descripción de ejemplo de la misma con el propósito de simplificar la divulgación y ayudar en la comprensión de uno o más de los diversos aspectos inventivos. Sin embargo, este método de divulgación no debe interpretarse como un reflejo de la intención de que las reivindicaciones requieran más características que las que se mencionan expresamente en cada reivindicación. Más bien, como reflejan las siguientes reivindicaciones, los aspectos
35 inventivos residen en menos de todas las características de una única realización de ejemplo divulgada anteriormente. De este modo, las reivindicaciones que siguen a la descripción se incorporan expresamente mediante la presente a esta descripción, siendo cada reivindicación por sí misma una realización de ejemplo separada de esta divulgación.

40 Además, si bien algunas realizaciones de ejemplo descritas en el presente documento incluyen algunas pero no otras características incluidas en otras realizaciones de ejemplo, las combinaciones de características de diferentes realizaciones de ejemplo están destinadas a estar dentro del alcance de la divulgación y formar diferentes realizaciones de ejemplo, como lo entenderían los expertos en la técnica.

45 En la descripción que se proporciona en el presente documento se establecen numerosos detalles específicos. Sin embargo, se entiende que se pueden practicar realizaciones de ejemplo de la divulgación sin estos detalles específicos. En otros casos, no se han mostrado en detalle métodos, estructuras y técnicas bien conocidos para no dificultar la comprensión de esta descripción.

50 De este modo, si bien se han descrito los que se cree que son los mejores modos de la divulgación, los expertos en la técnica reconocerán que se pueden realizar otras modificaciones adicionales. Por ejemplo, las fórmulas dadas anteriormente son meramente representativas de los procedimientos que pueden usarse. Se pueden agregar funcionalidades a los diagramas de bloques y se pueden intercambiar operaciones entre bloques funcionales. Se pueden agregar pasos a los métodos descritos dentro del alcance de la presente divulgación.
55

REIVINDICACIONES

- 1.- Un método (1300) para procesar una representación comprimida de una señal de audio espacial para generar una representación reconstruida de la señal de audio espacial, en donde la representación comprimida comprende una señal de audio basada en canal con un número predefinido de canales y metadatos, comprendiendo los metadatos información de dirección e información de energía, comprendiendo la información de dirección indicaciones de direcciones de llegada de uno o más elementos de audio en una escena de audio y comprendiendo la información de energía, para al menos una subbanda de frecuencia, respectivas indicaciones de potencia de señal asociadas con las direcciones de llegada, comprendiendo el método:
- generar (S1310) señales de audio del uno o más elementos de audio basándose en la señal de audio basada en canal, la información de dirección y la información de energía; y
- generar (S1320) una señal de audio residual de la cual el uno o más elementos de audio están sustancialmente ausentes, basándose en la señal de audio basada en canal, la información de dirección y la información de energía,
- en donde la generación de señales de audio del uno o más elementos de audio comprende:
- determinar (1400) coeficientes de una matriz de mezcla inversa M para mapear la señal de audio basada en canal a una representación intermedia que comprende la señal de audio residual y las señales de audio del uno o más elementos de audio, basándose en la información de dirección y la información de energía,
- en donde la determinación de los coeficientes de la matriz de mezcla inversa M comprende:
- determinar (S1410), para cada uno del uno o más elementos de audio, un vector de panoramización $Pan_{down}(dir)$ para panoramizar el elemento de audio a los canales de la señal de audio basada en canal, basándose en la dirección de llegada dir del elemento de audio;
- determinar (S1420) una matriz de mezcla E que se usaría para mapear la señal de audio residual y las señales de audio del uno o más elementos de audio a los canales de la señal de audio basada en canal, basándose en los vectores de panoramización determinados;
- determinar (S1430) una matriz de covarianza S para la representación intermedia basándose en la información de energía; y
- determinar (S1440) los coeficientes de la matriz de mezcla inversa M basándose en la matriz de mezcla E y la matriz de covarianza S.
- 2.- El método de acuerdo con la reivindicación 1, en donde una indicación de potencia de señal se refiere, para una dirección de llegada dada y una subbanda de frecuencia dada, a una fracción de potencia de señal en la subbanda de frecuencia dada para la dirección de llegada dada en relación con la potencia de señal total en la subbanda de frecuencia dada.
- 3.- El método de acuerdo con la reivindicación 1 o 2, en donde la información de energía incluye indicaciones de potencia de señal para cada una de una pluralidad de subbandas de frecuencia.
- 4.- El método de acuerdo con una cualquiera de las reivindicaciones 1 a 3, que comprende además:
- panoramizar (S1330) las señales de audio del uno o más elementos de audio a un conjunto de canales de un formato de audio de salida; y
- generar (S1340) una señal de audio multicanal reconstruida en el formato de audio de salida basándose en el uno o más elementos de audio panoramizados y la señal residual.
- 5.- El método de acuerdo con una cualquiera de las reivindicaciones 1 a 4, en donde la matriz de mezcla E se determina de acuerdo con

$$E = (I_N | Pan_{down}(dir_1) | \dots | Pan_{down}(dir_P))$$

- donde I_N es una matriz de identidad $N \times N$, con N indicando el número de canales de la señal basada en canal, $Pan_{down}(dir_p)$ es el vector de panoramización para el p-ésimo elemento de audio con dirección de llegada asociada dir_p que panoramizaría el p-ésimo elemento de audio a los N canales de la señal basada en canal, con $p = 1, \dots, P$ indicando uno respectivo de entre el uno o más elementos de audio y P indicando el número total del uno o más elementos de audio.

6.- El método de acuerdo con la reivindicación 5, en donde la matriz de covarianza S se determina como una matriz diagonal de acuerdo con

$$\{S\}_{n,n} = \text{rms}(Pan_{down})_n \left(1 - \sum_{p=1}^P e_p \right)$$

para $1 \leq n \leq N$, y

$$\{S\}_{N+p,N+p} = e_p$$

para $1 \leq p \leq P$,

donde e_p es la potencia de señal asociada con la dirección de llegada del p-ésimo elemento de audio.

7.- El método de acuerdo con una cualquiera de las reivindicaciones 1 a 6, en donde la determinación de los coeficientes de la matriz de mezcla inversa basándose en la matriz de mezcla y la matriz de covarianza implica la determinación de una pseudoinversa basándose en la matriz de mezcla y la matriz de covarianza.

8.- El método de acuerdo con una cualquiera de las reivindicaciones 1 a 7, en donde la matriz de mezcla inversa M se determina de acuerdo con

$$M = S \times E^* \times (E \times S \times E^*)^{-1}$$

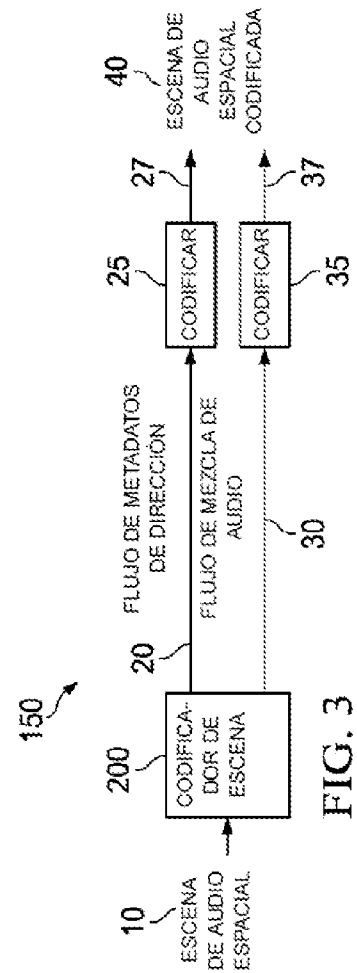
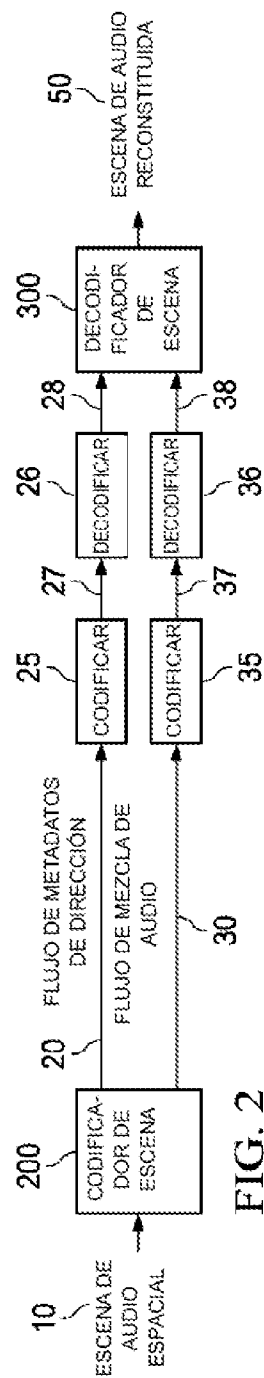
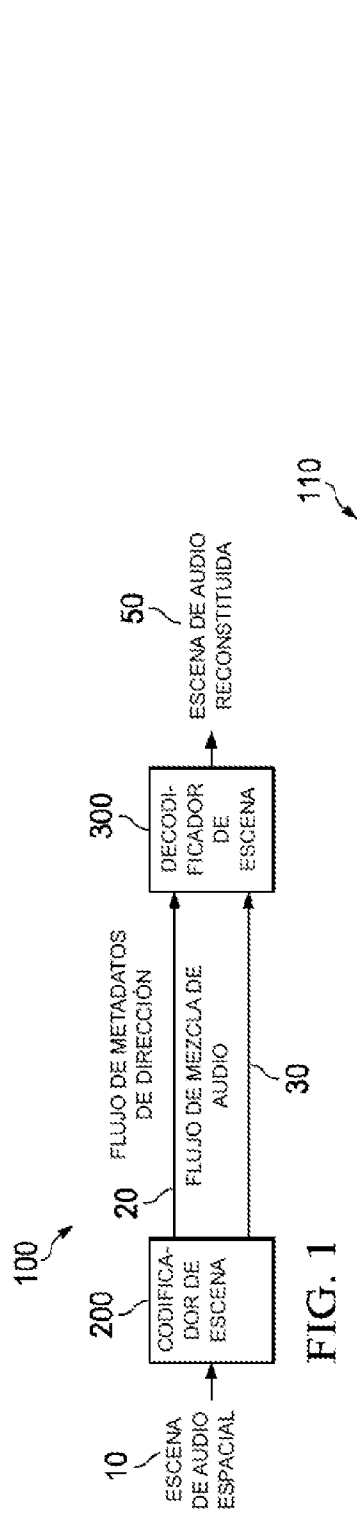
donde "x" indica el producto de matriz y "*" indica la transpuesta conjugada de una matriz.

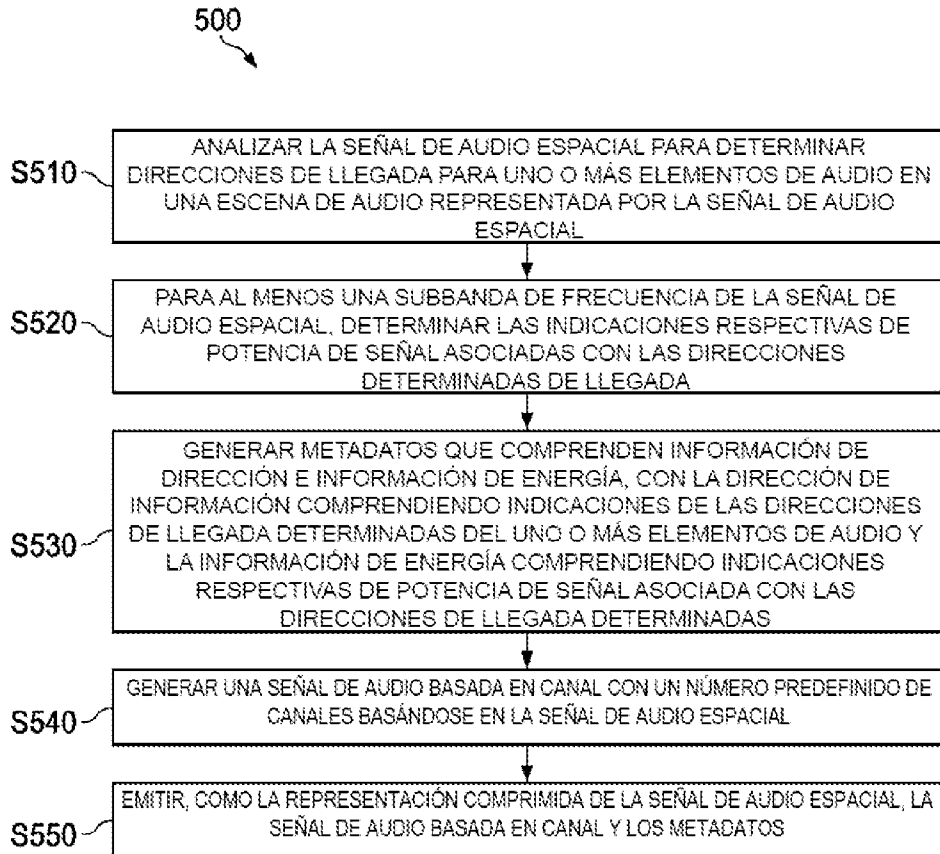
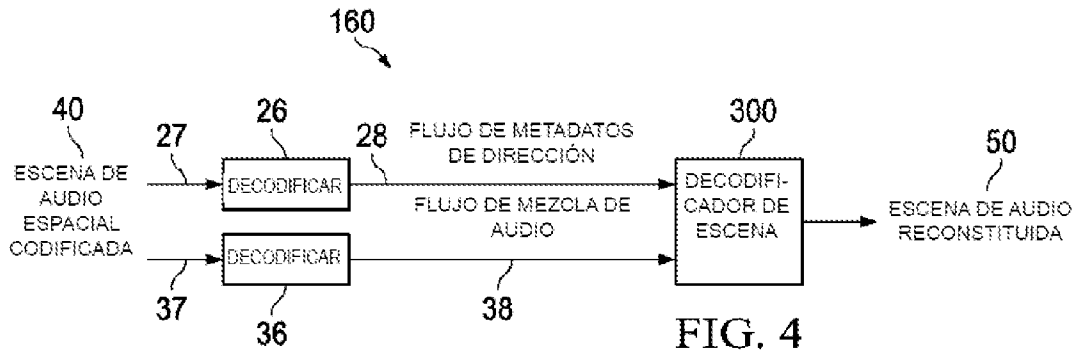
9.- El método de acuerdo con una cualquiera de las reivindicaciones anteriores, en donde la señal de audio basada en canal es una señal ambisónica de primer orden.

10.- Un programa que comprende instrucciones que, cuando son ejecutadas por un procesador, hacen que el procesador lleve a cabo todos los pasos del método de acuerdo con una cualquiera de las reivindicaciones anteriores.

11.- Un medio de almacenamiento legible por ordenador que almacena el programa de acuerdo con la reivindicación 10.

12.- Un aparato que comprende un procesador y una memoria acoplada al procesador, en donde el procesador está adaptado para llevar a cabo todos los pasos del método de acuerdo con una cualquiera de las reivindicaciones 1 a 9.





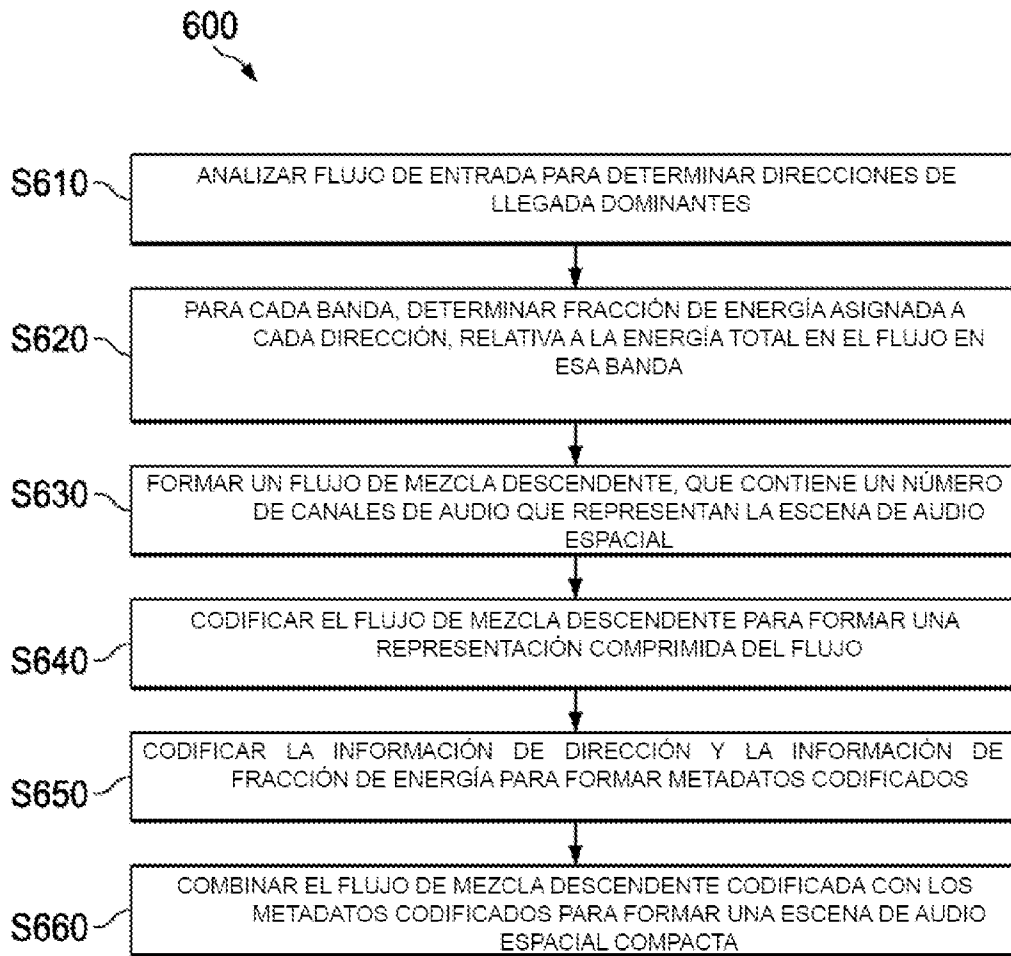


FIG. 6

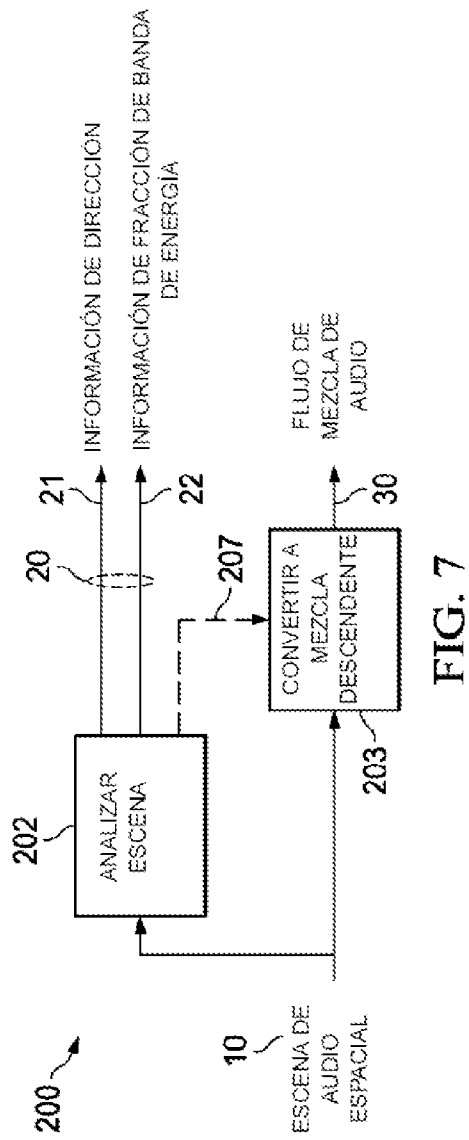


FIG. 7

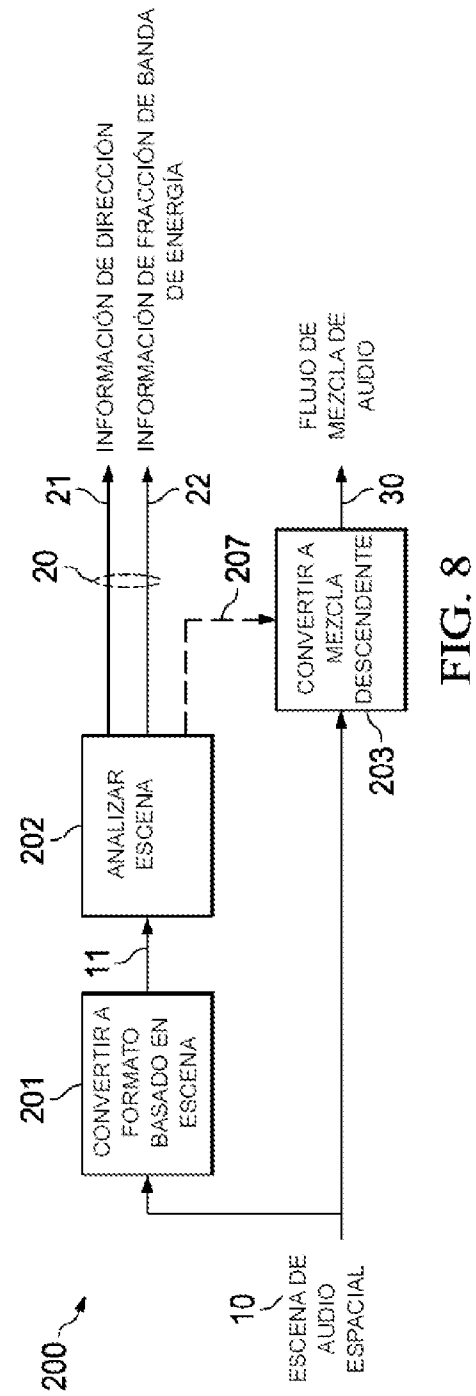
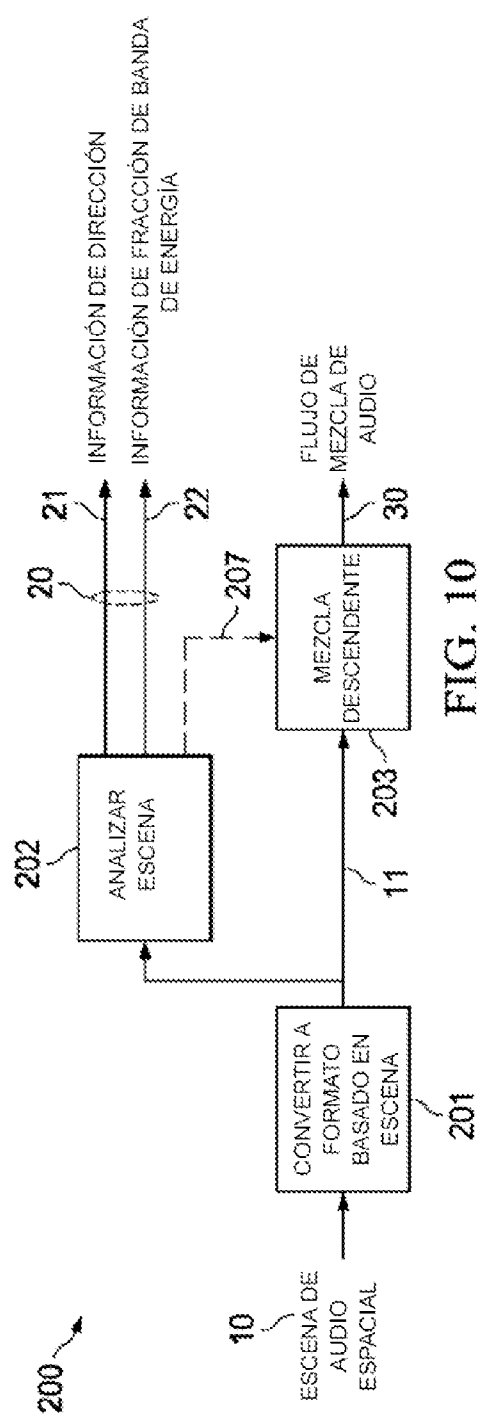
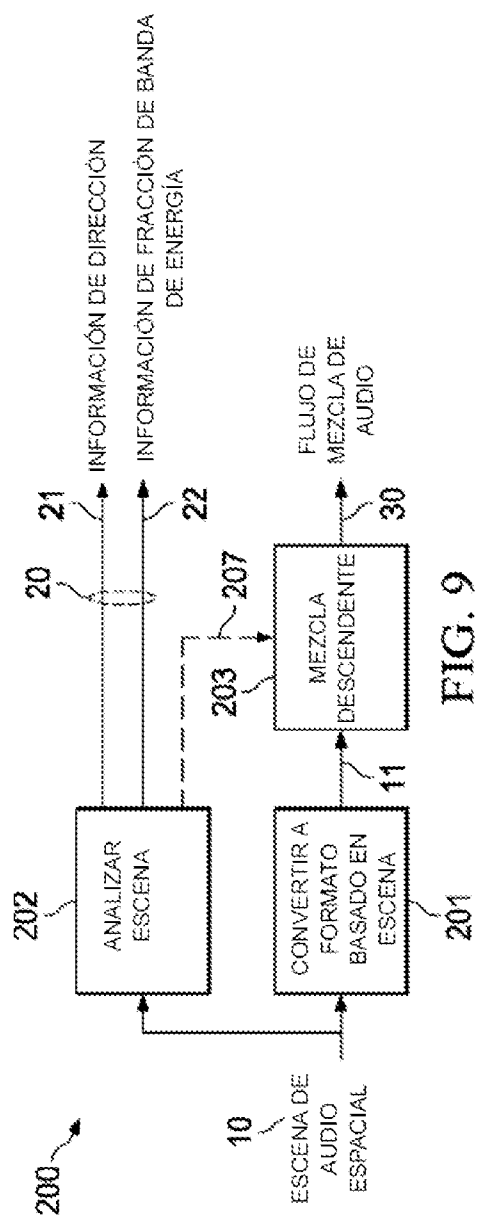
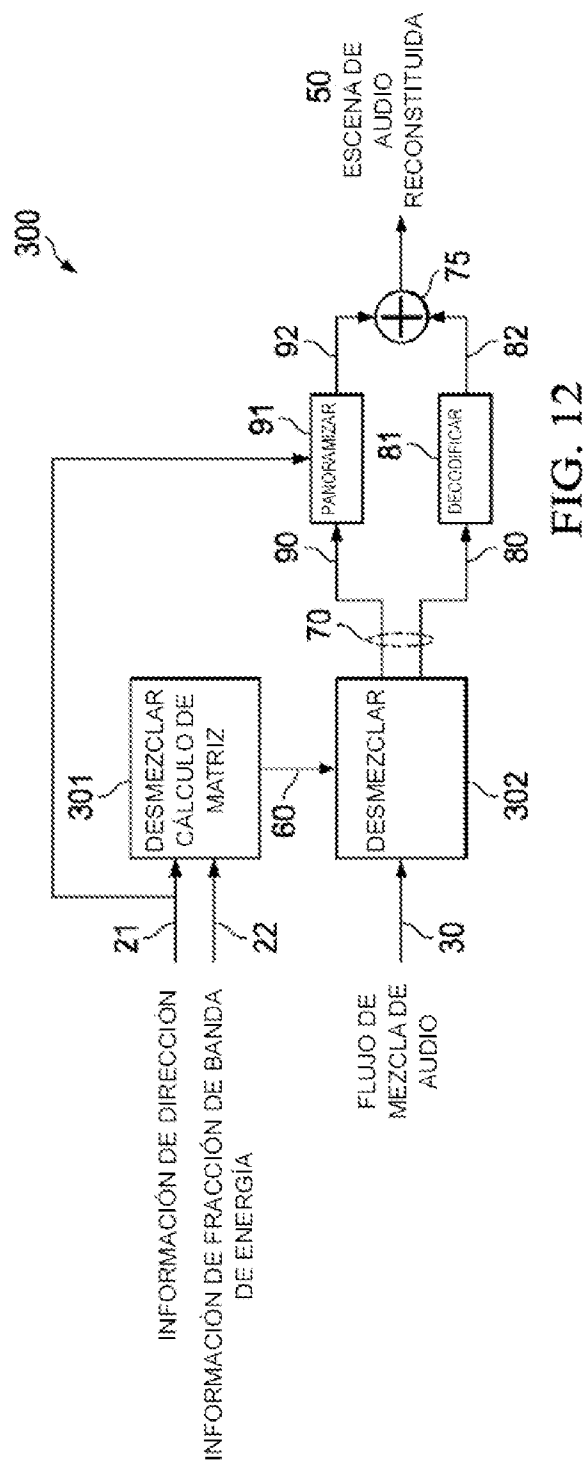
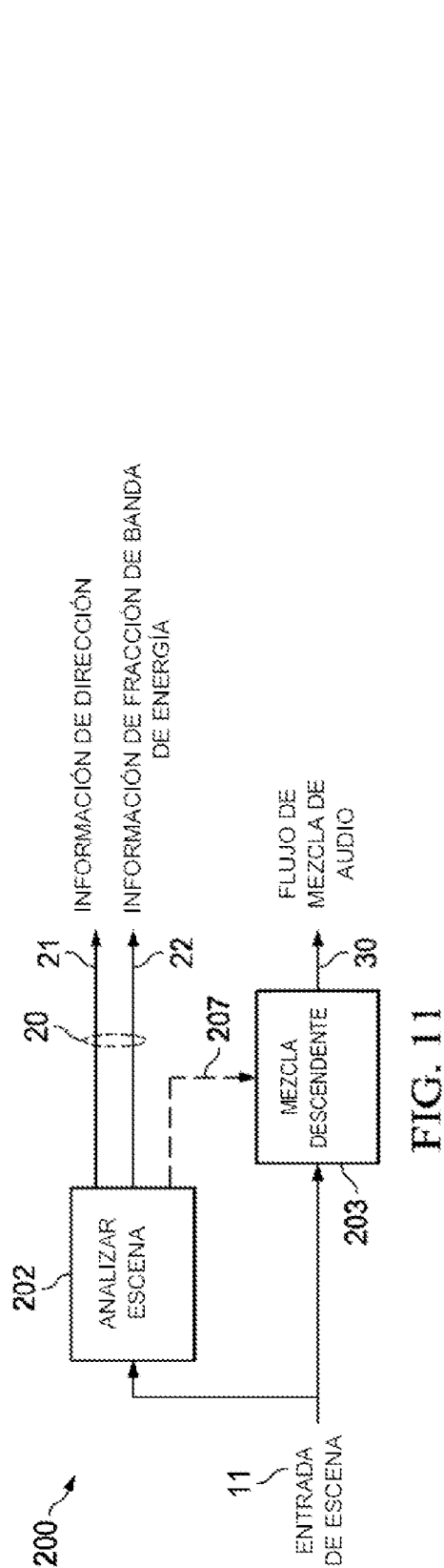


FIG. 8





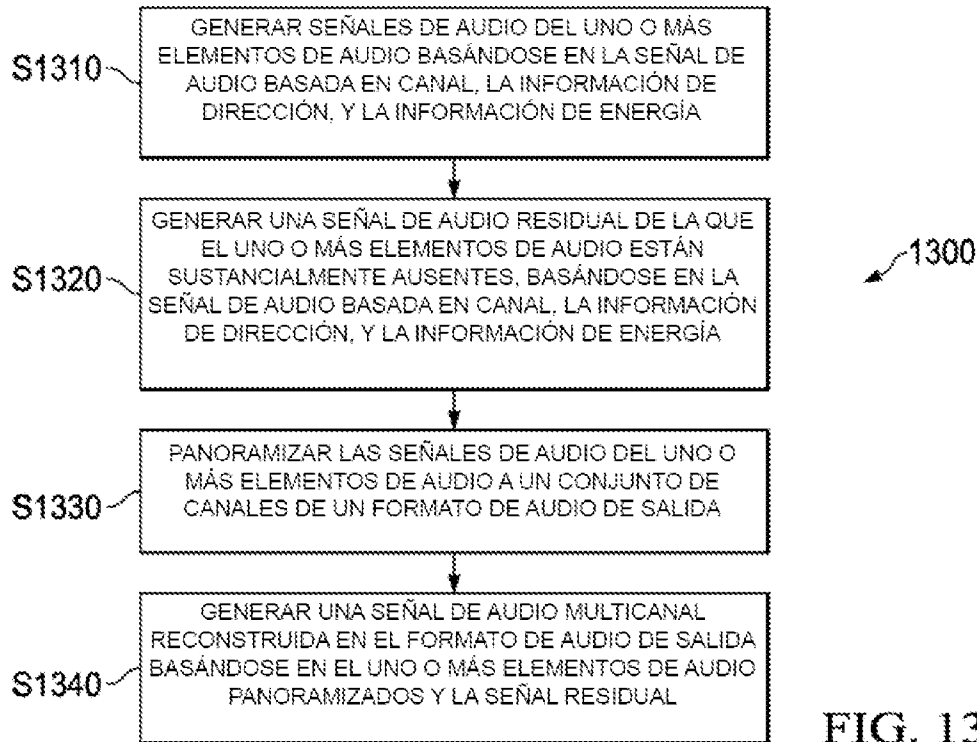


FIG. 13

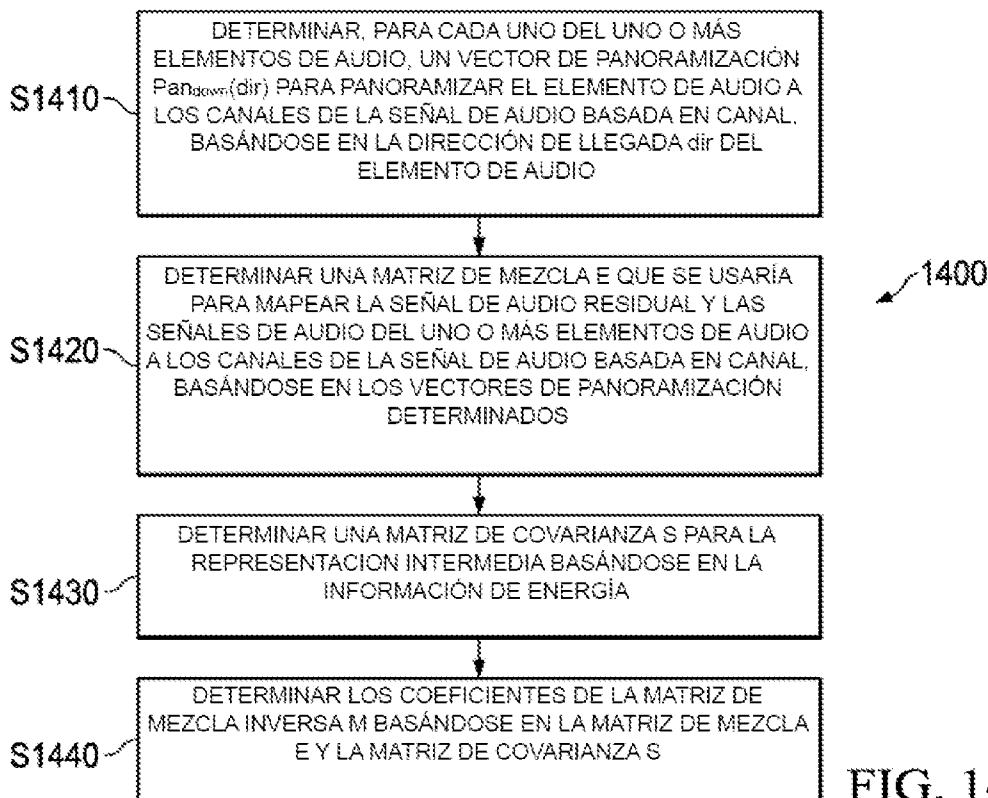


FIG. 14

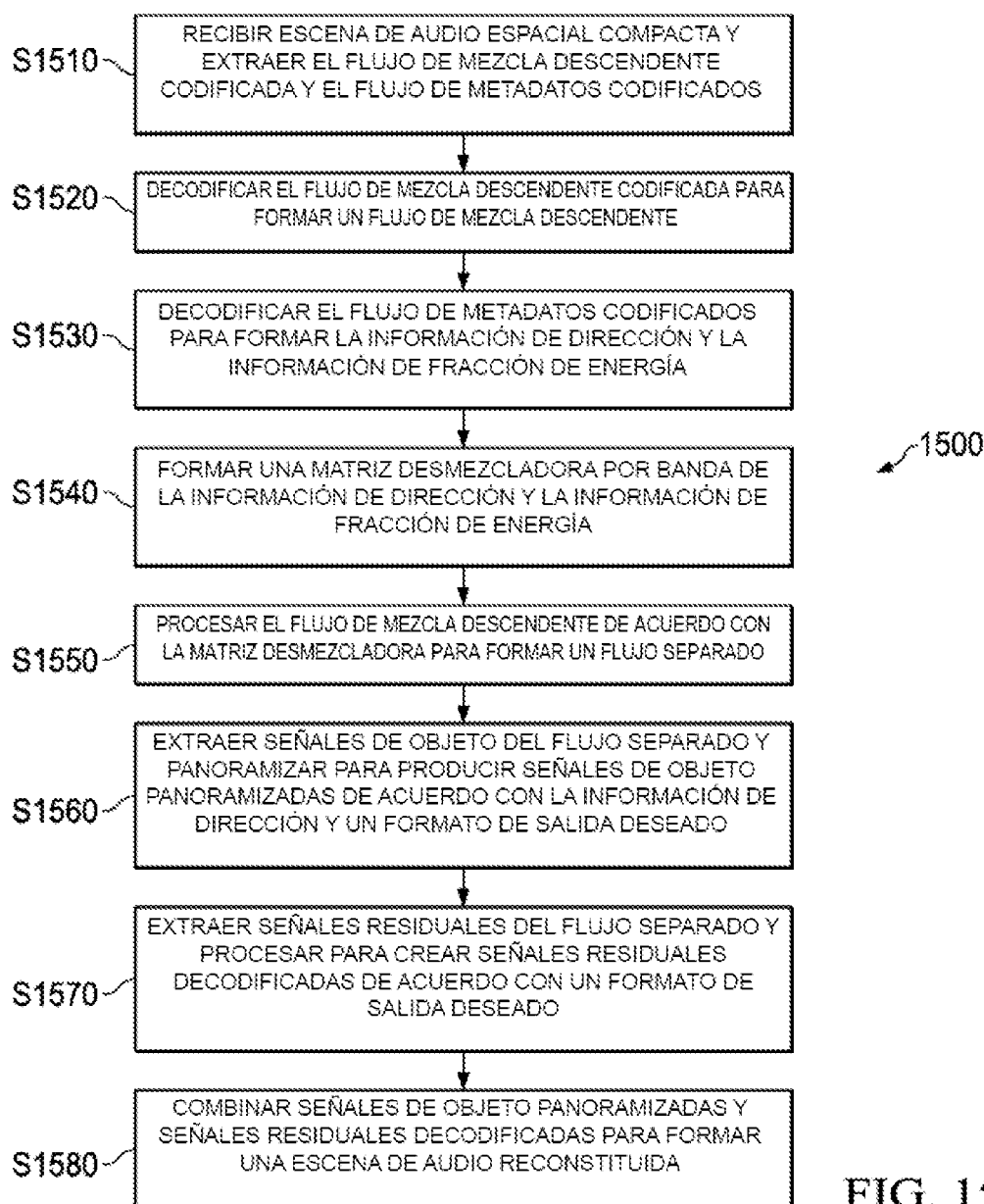


FIG. 15

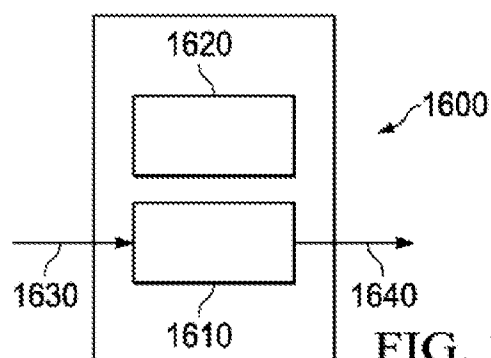


FIG. 16