(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 779 603**

(51) Int. CI.:

H04S 3/00 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **17.11.2016 PCT/US2016/062497**

(87) Fecha y número de publicación internacional: **26.05.2017 WO17087650**

(96) Fecha de presentación y número de la solicitud europea: **17.11.2016 E 16806384 (0)**

(97) Fecha y número de publicación de la concesión europea: **19.02.2020 EP 3378239**

(54) Título: **Sistema y método de salida binaural paramétrico**

(30) Prioridad:

17.11.2015 US 201562256462 P
14.12.2015 EP 15199854

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
18.08.2020

(73) Titular/es:

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

(72) Inventor/es:

BREEBAART, DIRK JEROEN;
COOPER, DAVID MATTHEW;
DAVIS, MARK F.;
MCGRATH, DAVID S.;
KJOERLING, KRISTOFER;
MUNDT, HARALD y
WILSON, RHONDA J.

(74) Agente/Representante:

ELZABURU, S.L.P

DESCRIPCIÓN

Sistema y método de salida binaural paramétrico

Campo de la invención

La presente invención proporciona sistemas y métodos para la forma mejorada de salida binaural paramétrica cuando se utiliza opcionalmente el seguimiento de la cabeza.

Referencias

Gundry, K., "A New Matrix Decoder for Surround Sound," AES 19th International Conf., Schloss Elmau, Alemania, 2001.

Vinton, M., McGrath, D., Robinson, C., Brown, P., "Next generation surround decoding and up-mixing for consumer and professional applications", AES 57th International Conf., Hollywood, CA, EE.UU., 2015.

Wightman, F. L., y Kistler, D. J. (1989). "Headphone simulation of free-field listening. I. Stimulus synthesis," J. Acoust. Soc. Am. 85, 858-867.

Norma ISO/IEC 14496-3:2009 – Tecnología de la información – Codificación de objetos audiovisuales - - Parte 3: Audio, 2009.

Mania, Katerina, *et al.* "Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity." Proceedings of the 1st Symposium on Applied perception in graphics and visualization. ACM, 2004.

Allison, R. S., Harris, L. R., Jenkin, M., Jasiobedzka, U., y Zacher, J. E. (2001, March). Tolerance of temporal delay in virtual environments. In Virtual Reality, 2001. Proceedings. IEEE (págs. 247-254). IEEE.

Van de Par, Steven, y Armin Kohlrausch. "Sensitivity to auditory-visual asynchrony and to jitter in auditory-visual timing." Electronic Imaging. International Society for Optics and Photonics, 2000.

Antecedentes de la invención

Cualquier discusión sobre la técnica anterior a lo largo de la especificación no debe considerarse de ninguna manera como una admisión de que dicha técnica es ampliamente conocida o forma parte del conocimiento general común en el campo.

La creación, codificación, distribución y reproducción de contenido de audio se basa tradicionalmente en canales. Es decir, se prevé un sistema de reproducción de destino específico para el contenido en todo el ecosistema de contenido. Ejemplos de tales sistemas de reproducción de destino son mono, estéreo, 5.1, 7.1, 7.1.4 y similares.

Si el contenido se va a reproducir en un sistema de reproducción diferente al previsto, se puede aplicar una mezcla descendente o ascendente. Por ejemplo, el contenido 5.1 se puede reproducir en un sistema de reproducción estéreo mediante el uso de ecuaciones específicas de mezcla descendente conocidas. Otro ejemplo es la reproducción de contenido estéreo a través de una configuración de altavoces 7.1, que puede comprender un llamado proceso de mezcla ascendente que podría o no ser guiado por la información presente en la señal estéreo, tal como la utilizada por los llamados codificadores matriciales como Dolby Pro Logic. Para guiar el proceso de mezcla ascendente, la información sobre la posición original de las señales antes de la mezcla descendente puede señalizarse implícitamente mediante la inclusión de relaciones de fase específicas en las ecuaciones de mezcla descendente, o dicho de otra manera, aplicando ecuaciones de mezcla descendente de valor complejo. Un ejemplo bien conocido de tal método de mezcla descendente utilizando coeficientes de mezcla descendente de valor complejo para contenido con altavoces colocados en dos dimensiones es LtRt (Vinton *et al.* 2015).

La señal de mezcla descendente (estéreo) resultante se puede reproducir en un sistema de altavoces estéreo, o se puede mezclar en configuraciones de altavoces con altavoces envolventes y/ o de altura. La ubicación prevista de la señal puede derivarse mediante un mezclador ascendente a partir de las relaciones de fase entre canales. Por ejemplo, en una representación estéreo LtRt, una señal que está fuera de fase (por ejemplo, tiene un coeficiente de correlación cruzada normalizado de forma de onda entre canales cerca de -1) idealmente debería ser reproducida por uno o más altavoces envolventes, mientras que un coeficiente de correlación positivo (cercano a +1) indica que la señal debe ser reproducida por los altavoces frente al oyente.

Se han desarrollado una variedad de algoritmos y estrategias de mezcla ascendente que difieren en sus estrategias para recrear una señal multicanal a partir de la mezcla descendente estéreo. En mezcladores ascendentes relativamente simples, el coeficiente de correlación cruzada normalizado de las señales de forma de onda estéreo se rastrea en función del tiempo, mientras que la(s) señal(es) son dirigidas a los altavoces delanteros o traseros dependiendo del valor del coeficiente de correlación cruzada normalizado. Este enfoque funciona bien para un contenido relativamente simple en el que solo un objeto auditivo está presente simultáneamente. Los mezcladores ascendentes más avanzados se basan en información estadística que se deriva de regiones de frecuencia específicas

para controlar el flujo de señal desde la entrada estéreo a la salida multicanal (Gundry 2001, Vinton et al. 2015). Específicamente, un modelo de señal basado en un componente dirigido o dominante y una señal residual estéreo (difusa) se puede emplear en mosaicos individuales de tiempo/frecuencia como se describe en el documento EP1070438. Además de la estimación del componente dominante y las señales residuales, también se calcula un ángulo de dirección (en acimut, posiblemente aumentado con elevación), y posteriormente la señal del componente dominante se dirige a uno o más altavoces para reconstruir la posición (calculada) durante la reproducción.

El uso de codificadores matriciales y decodificadores/mezcladores ascendentes no se limita al contenido basado en canales. Los desarrollos recientes en la industria del audio se basan en objetos de audio en lugar de canales, en los que uno o más objetos consisten en una señal de audio y metadatos asociados que indican, entre otras cosas, su posición prevista en función del tiempo. Para dicho contenido de audio basado en objetos, también se pueden usar codificadores matriciales, como se describe en Vinton et al. 2015. En dicho sistema, las señales de los objetos se mezclan de forma descendente en una representación de señal estéreo con coeficientes de mezcla descendente que dependen de los metadatos posicionales del objeto.

La mezcla ascendente y la reproducción de contenido codificado en matriz no se limitan necesariamente a la reproducción en altavoces. La representación de un componente dirigido o dominante que consiste en una señal de componente dominante y una posición (prevista) permite la reproducción en auriculares mediante convolución con respuestas de impulso relacionadas con la cabeza (HRIR) (Wightman *et al*, 1989). Un esquema simple de un sistema que implementa este método se muestra 1 en la figura 1. La señal de entrada 2, en un formato codificado en matriz, se analiza primero 3 para determinar la dirección y magnitud de componente dominante. La señal de componente dominante se convoluciona 4, 5 por medio de un par de HRIRs derivados de una búsqueda 6 basada en la dirección del componente dominante, para calcular una señal de salida para la reproducción de auriculares 7 de modo que la señal de reproducción se perciba como proveniente de la dirección que fue determinada por la etapa de análisis de componente dominante 3. Este esquema puede aplicarse en señales de banda ancha así como en subbandas individuales, y puede aumentarse con procesamiento dedicado de señales residuales (o difusas) de varias maneras.

El uso de codificadores matriciales es muy adecuado para la distribución y reproducción en receptores AV, pero puede ser problemático para aplicaciones móviles que requieren bajas velocidades de transmisión de datos y bajo consumo de energía.

Independientemente de si se utiliza contenido basado en canales o en objetos, los codificadores y decodificadores de matriz dependen de relaciones de fase entre canales bastante precisas de las señales que se distribuyen desde el codificador de matriz al decodificador. En otras palabras, el formato de distribución debe en gran medida preservar la forma de onda. Tal dependencia de la preservación de la forma de onda puede ser problemática en condiciones restringidas de velocidad de bits, en las que los códecs de audio emplean métodos paramétricos en lugar de herramientas de codificación de forma de onda para obtener una mejor calidad de audio. Se conocen generalmente ejemplos de tales herramientas paramétricas que no conservan la forma de onda a menudo que se denominan replicación de banda espectral, estéreo paramétrico, codificación de audio espacial y similares, tal como se implementa en los códecs de audio MPEG-4 (ISO/ IEC 14496-3: 2009)

Como se expone en la sección anterior, el mezclador ascendente consiste en análisis y dirección (o convolución HRIR) de señales. Para dispositivos con alimentación, como los receptores AV, esto generalmente no causa problemas, pero para los dispositivos que funcionan con baterías, como teléfonos móviles y tabletas, la complejidad computacional y los requisitos de memoria correspondientes asociados con estos procesos a menudo no son deseables debido a su impacto negativo en la vida de la batería.

El análisis anteriormente mencionado normalmente también introduce latencia de audio adicional. Tal latencia de audio no es deseable porque (1) requiere demoras de video para mantener la sincronización de labios de audio y video que requiere una cantidad significativa de memoria y potencia de procesamiento, y (2) puede causar asincronía/ latencia entre los movimientos de la cabeza y la reproducción de audio en el caso del seguimiento de la cabeza.

La mezcla descendente codificada en matriz también puede no sonar de forma óptima en altavoces estéreo o auriculares, debido a la posible presencia de componentes de señal fuertes fuera de fase.

Compendio de la invención

Un objeto de la invención es proporcionar una forma mejorada de salida binaural paramétrica.

Según un primer aspecto de la presente invención, se proporciona un método de acuerdo con la reivindicación 1, de codificación de audio de entrada basado en canal o objeto para la reproducción, el método incluye las etapas de: (a) renderizar inicialmente la entrada basada en canal u objeto audio en una presentación de salida inicial (por ejemplo, representación de salida inicial); (b) determinar una estimación del componente de audio dominante a partir del audio de entrada basado en canal u objeto y determinar una serie de factores de ponderación del componente de audio dominante para mapear la presentación de salida inicial en el componente de audio dominante; (c) determinar una estimación de la dirección o posición del componente de audio dominante; y (d) codificar la presentación de salida inicial, los factores de ponderación del componente de audio dominante, la dirección o posición del componente de audio dominante como la señal codificada para la reproducción, en el que dicha presentación de salida inicial

comprende una mezcla descendente estéreo. Proporcionar la serie de factores de ponderación de componentes de audio dominantes para mapear la presentación de salida inicial en el componente de audio dominante puede permitir utilizar los factores de ponderación de componentes de audio dominantes y la presentación de salida inicial para determinar la estimación del componente dominante.

En algunas realizaciones, el método incluye además determinar una estimación de una mezcla residual que es la presentación de salida inicial menos una representación del componente de audio dominante o la estimación del mismo. El método también puede incluir generar una mezcla binaural anecoica del canal o el audio de entrada basado en objetos, y determinar una estimación de una mezcla residual, en donde la estimación de la mezcla residual puede ser la mezcla binaural anecoica menos una representación del componente de audio dominante o la estimación de los mismos. Además, el método puede incluir la determinación de una serie de coeficientes de matriz residuales para mapear la presentación de salida inicial para la estimación de la mezcla residual.

La presentación de salida inicial puede comprender una presentación de auriculares o altavoces. El audio de entrada basado en canal u objeto puede estar en mosaico de tiempo y frecuencia y la etapa de codificación puede repetirse para una serie de etapas de tiempo y una serie de bandas de frecuencia. La presentación de salida inicial puede comprender una mezcla de altavoces estéreo.

Según un aspecto adicional de la presente invención, se proporciona un método para decodificar una señal de audio codificada según la reivindicación 7, incluyendo la señal de audio codificada: una presentación de salida inicial; una dirección de componente de audio dominante y factores de ponderación de componente de audio dominante, en el que dicha presentación de salida inicial comprende una mezcla descendente estéreo; el método comprende las etapas de: (a) utilizar los factores de ponderación de componente de audio dominante y la presentación de salida inicial para determinar un componente dominante estimado; (b) renderizar el componente dominante estimado con una binauralización en una ubicación espacial relativa a un oyente previsto según la dirección del componente de audio dominante para formar un componente dominante estimado binauralizado renderizado; (c) reconstruir una estimación de componente residual a partir de la presentación de salida inicial; y (d) combinar el componente dominante estimado binauralizado renderizado y el componente residual estimado para formar una señal codificada de audio espacializada de salida.

La señal de audio codificada puede incluir además una serie de coeficientes de matriz residuales que representan una señal de audio residual y la etapa (c) puede comprender además (c1) aplicar los coeficientes de matriz residual a la presentación de salida inicial para reconstruir la estimación del componente residual.

En algunas realizaciones, la estimación del componente residual puede reconstruirse restando el componente dominante estimado binauralizado renderizado de la presentación de salida inicial. La etapa (b) puede incluir una rotación inicial del componente dominante estimado según una señal de entrada de seguimiento de la cabeza que indica la orientación de la cabeza de un oyente previsto.

Breve descripción de los dibujos

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

la figura 1 ilustra esquemáticamente un decodificador de auriculares para contenido codificado en matriz;

la figura 2 ilustra esquemáticamente un codificador según una realización;

la figura 3 es un diagrama de bloques esquemático del decodificador.

la figura 4 es una visualización detallada de un codificador; y

la figura 5 ilustra una forma del decodificador con más detalle.

Descripción detallada

Las realizaciones proporcionan un sistema y un método para representar contenido de audio basado en objetos o canales que es (1) compatible con la reproducción estéreo, (2) permite la reproducción binaural incluyendo el seguimiento de la cabeza, (3) es de una baja complejidad de decodificador y (4) no se basa en, pero es compatible con la codificación matricial.

Esto se logra combinando el análisis del lado del codificador de uno o más componentes dominantes (u objeto dominante o combinación de los mismos) incluyendo ponderaciones para predecir estos componentes dominantes a partir de una mezcla descendente, en combinación con parámetros adicionales que minimizan el error entre un renderizado binaural basado solo en los componentes dirigidos o dominantes, y la presentación binaural deseada del contenido completo.

En una realización, se proporciona un análisis del componente dominante (o componentes dominantes múltiples) en el codificador en lugar del decodificador/renderizador. La cadena de audio se aumenta con metadatos que indican la

dirección del componente dominante e información sobre cómo se puede(n) obtener el/los componente(s) dominante(s) de una señal de mezcla descendente asociada.

La figura 2 ilustra una forma de un codificador 20 de la realización preferida. El contenido 21 basado en objeto o canal se somete a un análisis 23 para determinar uno o más componentes dominantes. Este análisis puede tener lugar en función del tiempo y la frecuencia (suponiendo que el contenido de audio se divida en mosaicos de tiempo y subtítulos de frecuencia). El resultado de este proceso es una señal de componente dominante 26 (o múltiples señales de componente dominante), y la información asociada de posición (s) o dirección (s) 25. Posteriormente, se estiman 24 las ponderaciones y la salida 27 para permitir la reconstrucción de la señal de componente dominante (s) a partir de una mezcla descendente transmitida. Este generador de mezcla descendente 22 no necesariamente tiene que cumplir con las reglas de mezcla descendente LtRt, pero podría ser una mezcla descendente estándar ITU (LoRo) que utiliza coeficientes de mezcla descendente no negativos y de valor real. Por último, la señal de mezcla descendente de salida 29, las ponderaciones 27 y los datos de posición 25 son empaquetados por un codificador de audio 28 y preparados para su distribución.

Volviendo ahora a la figura 3, se ilustra un decodificador correspondiente 30 de la realización preferida. El decodificador de audio reconstruye la señal de mezcla descendente. La señal es introducida 31 y desempaquetada por el decodificador de audio 32 en señal de mezcla descendente, ponderaciones y dirección de los componentes dominantes. Posteriormente, las ponderaciones de estimación de componentes dominantes se utilizan para reconstruir 34 el/los componente(s) dirigido(s), que son renderizados 36 usando datos de posición o dirección transmitidos. Los datos de posición pueden modificarse opcionalmente 33 dependiendo de la información de rotación y translación de la cabeza 38. Además, los componentes dominantes reconstruidos pueden sustraerse 35 de la mezcla descendente. Opcionalmente, hay una sustracción del/de los componente(s) dominante(s) dentro de la ruta de mezcla descendente, pero alternativamente, esta sustracción también puede ocurrir en el codificador, como se describe a continuación.

Para mejorar la eliminación o cancelación del componente dominante reconstruido en el sustractor 35, la salida del componente dominante puede representarse primero usando los datos de posición o dirección transmitidos antes de la sustracción. Esta etapa de representación opcional 39 se muestra en la figura 3.

Volviendo ahora a describir inicialmente el codificador con más detalle, la figura 4 muestra una forma de codificador 40 para procesar contenido de audio basado en objetos (por ejemplo, Dolby Atmos). Los objetos de audio se almacenan originalmente como objetos Atmos 41 y se dividen inicialmente en mosaicos de tiempo y frecuencia usando un banco 42 de filtro de espejo en cuadratura de valor complejo híbrido (HCQMF). Las señales de los objetos de entrada se pueden denotar por $x_i[n]$ cuando se omiten los índices de tiempo y frecuencia correspondientes; la posición correspondiente dentro del cuadro actual viene dada por el vector unitario $\vec{p}_1$, y el índice i se refiere al número de objeto, y el índice n se refiere al tiempo (por ejemplo, índice de muestra de subbanda). El objeto de entrada señala $x_i[n]$ son un ejemplo de audio de entrada basado en canal u objeto.

Una mezcla binaural anecoica, sub-banda Y (y_l, y_r) se crea 43 utilizando escalares de valor complejo $H_{l,i}$, $H_{r,i}$ (por ejemplo, HRTF 48 de un toque) que representan la representación de sub-banda de los HRIRs correspondientes a la posición $\vec{p}_i$:

$$y_l[n] = \sum_i H_{l,i} x_i[n]$$

$$y_r[n] = \sum_i H_{r,i} x_i[n]$$

Alternativamente, la mezcla binaural Y (y_l, y_r) puede crearse por convolución utilizando respuestas de impulso relacionadas con la cabeza (HRIRs). Además, una mezcla descendente estéreo $z_l z_r$ (que incorpora a modo de ejemplo una presentación de salida inicial) se crea 44 utilizando coeficientes de ganancia de panoramización de amplitud $g_{l,i} g_{r,i}$:

$$z_l[n] = \sum_i g_{l,i} x_i[n]$$

$$z_r[n] = \sum_i g_{r,i} x_i[n]$$

El vector de dirección del componente dominante $\vec{p}_D$ (que encarna a modo de ejemplo una dirección o posición de componente de audio dominante) puede estimarse calculando el componente dominante 45 calculando inicialmente una suma ponderada de vectores de dirección de unidad para cada objeto:

$$\vec{p}_D = \frac{\sum_i \sigma_i^2 \vec{p}_i}{\sum_i \sigma_i^2}$$

con σ_i^2 la energía de la señal $x_i[n]$:

$$\sigma_i^2 = \sum_n x_i[n] x_i^*[n]$$

y con (.)* siendo el operador de conjugación compleja.

La señal dominante/dirigida, d[n] (que encarna a modo de ejemplo un componente de audio dominante) viene dada por:

$$d[n] = \sum_i x_i[n] \mathcal{F}(\vec{p}_D, \vec{p}_i)$$

con $\mathcal{F}(\vec{p}_1, \vec{p}_2)$ una función que produce una ganancia que disminuye al aumentar la distancia entre los vectores unitarios $(\vec{p}_1, \vec{p}_2)$. Por ejemplo, para crear un micrófono virtual con un patrón de direccionalidad basado en armónicos esféricos de orden superior, una implementación correspondería a:

$$\mathcal{F}(\vec{p}_1, \vec{p}_2) = (a + b\vec{p}_1^T . \vec{p}_2)^c$$

con $\vec{p}_1$ representando un vector de dirección unitario en un sistema de coordenadas bidimensional o tridimensional, (.) el operador del producto de puntos para dos vectores y con parámetros a modo de ejemplo a, b, c (por ejemplo a = b = 0,5; c = 1).

Las ponderaciones o coeficientes de predicción $w_{l,d}$ $w_{r,d}$ se calculan 46 y se usan para calcular 47 una señal dirigida estimada $\hat{d}[n]$:

$$\hat{d}[n] = w_{l,d} z_l + w_{r,d} z_r$$

con ponderaciones $w_{l,d}$ $w_{r,d}$ minimizando el error medio cuadrático entre d[n] y $\hat{d}[n]$ dadas las señales de mezcla descendente z_l,z_r. Las ponderaciones $w_{l,d}$ $w_{r,d}$ son un ejemplo de factores de ponderación de componentes de audio dominantes para mapear la presentación de salida inicial (por ejemplo, z_l,z_r) al componente de audio dominante (por ejemplo, $\hat{d}[n]$). Un método conocido para derivar estas ponderaciones es mediante la aplicación de un predictor mínimo de error medio cuadrático (MMSE):

$$\begin{bmatrix} w_{l,d} \\ w_{r,d} \end{bmatrix} = (R_{zz} + \epsilon I)^{-1} R_{zd}$$

con R_{ab} la matriz de covarianza entre las señales para las señales a y las señales b, y ϵ un parámetro de regularización.

Posteriormente, se puede restar 49 la estimación representada de la señal del componente dominante $\hat{d}$ [n] de la mezcla binaural anecoica y_l, y_r para crear una mezcla binaural residual $\tilde{y}_l$, $\tilde{y}_r$ utilizando HRTF (HRIR) $H_{l,D}$, $H_{r,D}$ 50 asociado con la dirección/ posición $\vec{p}_D$ de la señal componente dominante $\hat{d}$:

$$\tilde{y}_l[n] = y_l[n] - H_{l,D}\, \hat{d}[n]$$

$$\tilde{y}_r[n] = y_r[n] - H_{r,D}\, \hat{d}[n]$$

Por último, se calcula 51 otro conjunto de coeficientes de predicción o ponderaciones $w_{i,j}$ que permite la reconstrucción de la mezcla binaural residual $\tilde{y}_l$, $\tilde{y}_r$ de la mezcla estéreo z_l,z_r utilizando estimaciones de error medio cuadrático mínimo:

$$\begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{bmatrix} = (R_{zz} + \epsilon I)^{-1} R_{z\tilde{y}}$$

con R_{ab} la matriz de covarianza entre señales para representación a y representación b, y ϵ un parámetro de regularización. Los coeficientes de predicción o ponderaciones $w_{i,j}$ son un ejemplo de coeficientes de matriz residual para mapear la presentación de salida inicial (por ejemplo, z_l, z_r) a la estimación de la mezcla binaural residual $\tilde{y}_l$, $\tilde{y}_r$. La expresión anterior puede estar sujeta a restricciones de nivel adicionales para superar cualquier pérdida de predicción. El codificador emite la siguiente información:

La mezcla estéreo z_l,z_r (que encarna a modo de ejemplo la presentación de salida inicial);

Los coeficientes para estimar el componente dominante $w_{l,d}$ $w_{r,d}$ (que encarna a modo de ejemplo los factores de ponderación de componentes de audio dominantes);

La posición o dirección del componente dominante $\vec{p}_D$;

Y opcionalmente, las ponderaciones residuales $w_{i,j}$ (que encarnan a modo de ejemplo los coeficientes de la matriz residuales).

Aunque la descripción anterior se refiere al renderizado basado en un único componente dominante, en algunas realizaciones el codificador puede estar adaptado para detectar múltiples componentes dominantes, determinar ponderaciones y direcciones para cada uno de los múltiples componentes dominantes, renderizar y restar cada uno de los múltiples componentes dominantes de la mezcla binaural anecoica Y, y luego determinar las ponderaciones residuales después de que cada uno de los múltiples componentes dominantes haya sido sustraído de la mezcla binaural anecoica Y.

Decodificador/renderizador

La figura 5 ilustra una forma de decodificador/renderizador 60 con más detalle. El decodificador/renderizador 60 aplica un proceso destinado a reconstruir la mezcla binaural y_l, y_r para salida al oyente 71 desde la información de entrada desempaquetada z_l,z_r; $w_{l,d}$ $w_{r,d}$; $\vec{p}_D$; $w_{i,j}$. Aquí, la mezcla estéreo z_l,z_r es un ejemplo de una primera representación de audio, y los coeficientes de predicción o ponderaciones $w_{i,j}$ y/o la dirección/posición $\vec{p}_D$ de la señal de componente dominante $\hat{d}$ son ejemplos de datos adicionales de transformación de audio.

Inicialmente, la mezcla descendente estéreo esta dividida en mosaicos de tiempo/frecuencia utilizando un banco de filtros adecuado o una transformación 61, como el banco de análisis 61 HCQMF. Otras transformaciones, como una transformada discreta de Fourier, una transformación de coseno o seno (modificada), banco de filtros de dominio de tiempo, o transformadas wavelet también se pueden aplicar igualmente. Posteriormente, la señal de componente dominante estimada $\hat{d}[n]$ se calcula 63 utilizando ponderaciones de coeficiente de predicción $w_{l,d}$ $w_{r,d}$:

$$\hat{d}[n] = w_{l,d} z_l + w_{r,d} z_r$$

La señal de componente dominante calculada $\hat{d}$ [n] es un ejemplo de una señal auxiliar. Por lo tanto, se puede decir que esta etapa corresponde a la creación de una o más señales auxiliares basadas en dicha primera representación de audio y datos de transformación recibidos.

Esta señal de componente dominante se procesa posteriormente 65 y se modifica 68 con HRTF 69 en función de los datos de posición/dirección transmitidos $\vec{p}_D$, posiblemente modificada (girada) en base a la información obtenida de un seguidor de la cabeza 62. Finalmente, la salida binaural anecoica total consiste en la señal de componente dominante renderizada sumada 66 con los residuos reconstruidos $\tilde{y}_l$, $\tilde{y}_r$ basados en las ponderaciones de coeficientes de predicción $w_{i,j}$:

$$\begin{bmatrix} \tilde{y}_l \\ \tilde{y}_r \end{bmatrix} = \left(\begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{bmatrix} \right) \begin{bmatrix} z_l \\ z_r \end{bmatrix}$$

$$\begin{bmatrix} \hat{y}_l \\ \hat{y}_r \end{bmatrix} = \left(\begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{bmatrix} + \begin{bmatrix} H_{l,D} \\ H_{r,D} \end{bmatrix} \begin{bmatrix} w_{l,d} & w_{r,d} \end{bmatrix} \right) \begin{bmatrix} z_l \\ z_r \end{bmatrix}$$

La salida binaural anecoica total es un ejemplo de una segunda representación de audio. Por lo tanto, se puede decir que esta etapa corresponde a la creación de una segunda representación de audio que consiste en una combinación de dicha primera representación de audio y dichas señal(es) auxiliares, en las que una o más de dichas señal(es) auxiliares se han modificado en respuesta a dichos datos de orientación de la cabeza.

Debe observarse además que si se recibe información sobre más de una señal dominante, cada señal dominante se puede representar y agregar a la señal residual reconstruida.

Mientras no se aplique rotación o traslación de la cabeza, las señales de salida $\hat{y}_l$, $\hat{y}_r$ deben estar muy cerca (en términos de error cuadrático medio) de las señales binaurales de referencia y_l, y_r siempre y cuando

$$\hat{d}[n] \approx d[n]$$

Propiedades clave

Como se puede observar de la formulación de la ecuación anterior, la operación efectiva para construir la presentación binaural anecoica a partir de la presentación estéreo consiste en una matriz 70 de 2x2, en la cual los coeficientes de la matriz dependen de la información transmitida $w_{l,d}$ $w_{r,d}$; $\vec{p}_D$; $w_{i,j}$ y la rotación y/o traslación del rastreador de la cabeza. Esto indica que la complejidad del proceso es relativamente baja, ya que el análisis de los componentes dominantes se aplica en el codificador en lugar de en el decodificador.

Si no se estima un componente dominante (por ejemplo, $w_{l,d}$,$w_{r,d}$ = 0), la solución descrita es equivalente a un método binaural paramétrico.

En los casos en que se desee excluir ciertos objetos de la rotación/seguimiento de la cabeza, estos objetos se pueden excluir del (1) análisis de dirección del componente dominante y (2) la predicción de la señal del componente dominante. Como resultado, estos objetos se convertirán de estéreo a binaural a través de los coeficientes $w_{i,j}$ y, por lo tanto, no se verán afectados por ninguna rotación o traslación de la cabeza.

En una línea de pensamiento similar, los objetos se pueden configurar en un modo de "paso a través", lo que significa que en la presentación binaural, estarán sujetos a un cribado de amplitud en lugar de una convolución HRIR. Esto se puede obtener simplemente usando ganancias de amplitud panorámica para los coeficientes $H._{,i}$ en lugar de los HRTF de un toque o cualquier otro procesamiento binaural adecuado.

Extensiones

Las realizaciones que no forman parte de la invención no se limitan al uso de mezclas estéreo descendentes, ya que también se pueden emplear otros recuentos de canales.

El decodificador 60 descrito con referencia a la figura 5 tiene una señal de salida que consta de una dirección de componente dominante representada más la señal de entrada formando una matriz mediante los coeficientes de matriz $w_{i,j}$. Los últimos coeficientes se pueden derivar de varias maneras, por ejemplo:

1. Los coeficientes $w_{i,j}$ se puede determinar en el codificador mediante la reconstrucción paramétrica de las señales $\tilde{y}_l$, $\tilde{y}_r$. En otras palabras, en esta implementación, los coeficientes $w_{i,j}$ tienen el objetivo de la reconstrucción fiel de las señales binaurales y_l, y_r eso se habría obtenido al representar los objetos/canales de entrada originales de forma binaural; en otras palabras, los coeficientes $w_{i,j}$ son contenidos dirigidos.

2. Los coeficientes $w_{i,j}$ se pueden enviar desde el codificador al decodificador para representar HRTF para posiciones espaciales fijas, por ejemplo en ángulos de acimut de +/- 45 grados. En otras palabras, la señal residual se procesa para simular la reproducción a través de dos altavoces virtuales en ciertos lugares. Como estos coeficientes que representan los HRTF se transmiten del codificador al decodificador, las ubicaciones de los altavoces virtuales pueden cambiar con el tiempo y la frecuencia. Si se emplea este enfoque utilizando altavoces virtuales estáticos para representar la señal residual, los coeficientes $w_{i,j}$ no necesitan transmisión del codificador al decodificador, y en su lugar pueden estar cableados en el decodificador. Una variación de este enfoque consistiría en un conjunto limitado de posiciones estáticas disponibles en el decodificador, con sus correspondientes coeficientes $w_{i,j}$, y la selección de qué posición estática se usa para procesar la señal residual se señala desde el codificador al decodificador.

Las señales $\tilde{y}_l$, $\tilde{y}_r$ pueden ser sometidas a un denominado mezclador ascendente, que reconstruye más de 2 señales por medio del análisis estadístico de estas señales en el decodificador, seguido de una renderización binaural de las señales mixtas ascendentes resultantes.

Los métodos descritos también se pueden aplicar en un sistema en el que la señal transmitida Z es una señal binaural. En ese caso particular, el decodificador 60 de la figura 5 permanece como está, mientras que el bloque etiquetado “Generar mezcla estéreo (LoRo)” 44 en la figura 4 debe reemplazarse por “Generar mezcla binaural anecoica”43 (figura 4) que es lo mismo que el bloque que produce el par de señales Y. Además, se pueden generar otras formas de mezclas según los requisitos.

Este enfoque puede ampliarse con métodos para reconstruir una o más señales de entrada FDN de la mezcla estéreo transmitida que consiste en un subconjunto específico de objetos o canales.

El enfoque se puede extender con múltiples componentes dominantes que se predicen a partir de la mezcla estéreo transmitida y se renderizan en el lado del decodificador. No existe una limitación fundamental de predecir solo un componente dominante para cada mosaico de tiempo/frecuencia. En particular, el número de componentes dominantes puede diferir en cada mosaico de tiempo/frecuencia.

Interpretación

La referencia a lo largo de esta especificación a "algunas realizaciones" o "una realización" significa que una característica, estructura o característica particular descrita en relación con la realización está incluida en al menos una realización de la presente invención. Por lo tanto, las apariencias de las frases "en algunas realizaciones" o "en una realización" en varios lugares a lo largo de esta especificación no se refieren necesariamente a la misma realización, pero puede ser así. Además, las propiedades, estructuras o características particulares se pueden combinar de cualquier manera adecuada en una o más realizaciones, como resultaría evidente para un experto en la materia a partir de esta descripción.

Como se usa en este 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 están haciendo referencia a diferentes menciones de objetos similares, y no pretende implicar que los objetos así descritos deben estar en una secuencia dada, ya sea temporal, espacial, en clasificación o de cualquier otra manera.

En las reivindicaciones que siguen y en la descripción en este documento, cualquiera de los términos que comprende, compuesto por son términos abiertos que significan incluir al menos los elementos/características que siguen, pero sin excluir otros. Por lo tanto, el término que comprende, cuando se usa en las reivindicaciones, no debe interpretarse como limitativo de los medios o elementos o etapas enumerados a continuación. Por ejemplo, el alcance de la expresión un dispositivo que comprende A y B no debe limitarse a dispositivos que consisten solo en los elementos A y B. Cualquiera de los términos incluyendo o que incluye como se usa 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. Por lo tanto, incluir es sinónimo y significa comprender.

Como se usa en el presente documento, el término "a modo de ejemplo" se usa en el sentido de proporcionar ejemplos, en lugar de indicar calidad. Es decir, una "realización a modo de ejemplo" es una realización proporcionada como un ejemplo, en oposición a ser necesariamente una realización de calidad a modo de ejemplo.

Debería apreciarse que en la descripción anterior de realizaciones a modo de ejemplo de la invención, algunas características de la invención a veces se agrupan en una sola realización, figura o descripción de las mismas con el fin de hacer más eficiente la divulgación y ayudar a comprender uno o más de los diversos aspectos inventivos. Sin embargo, este método de divulgación no debe interpretarse como que refleja una intención de que la invención reivindicada requiera más características de las que se mencionan expresamente en cada reivindicación. Más bien, como reflejan las siguientes reivindicaciones, los aspectos inventivos se encuentran en menos de todas las características de una sola realización descrita anteriormente. Por lo tanto, las reivindicaciones que siguen a la Descripción detallada se incorporan expresamente en esta Descripción detallada, y cada una de las reivindicaciones se presenta como una realización separada de esta invención.

Además, aunque algunas realizaciones descritas en el presente documento incluyen algunas pero no otras características incluidas en otras realizaciones, las combinaciones de características de diferentes realizaciones están destinadas a estar dentro del alcance de la invención y forman diferentes realizaciones, como entenderán los expertos en la materia dentro del alcance definido por las reivindicaciones adjuntas.

Además, algunas de las realizaciones se describen en el presente documento como un método o combinación de elementos de un método que pueden ser implementados mediante un procesador de un sistema informático o mediante otros medios para llevar a cabo la función. Por lo tanto, un procesador con las instrucciones necesarias para llevar a cabo dicho método o elemento de un método forma un medio para llevar a cabo el método o elemento de un método. Además, un elemento descrito en la presente memoria de una realización de aparato es un ejemplo de un medio para llevar a cabo la función realizada por el elemento con el fin de llevar a cabo la invención.

En la descripción proporcionada en la presente memoria, se exponen numerosos detalles específicos. Sin embargo, se entiende que las realizaciones de la invención se pueden llevar a la práctica sin estos detalles específicos. En otros casos, los métodos, estructuras y técnicas bien conocidos no se han mostrado en detalle para no oscurecer la comprensión de esta descripción.

De manera similar, se debe observar que el término "acoplado", cuando se usa en las reivindicaciones, no debe interpretarse como limitado solo a conexiones directas. Se pueden usar los términos "acoplado" y "conectado", junto con sus derivados. Debe entenderse que estos términos no pretenden ser sinónimos entre sí. Por lo tanto, el alcance de la expresión de un dispositivo A acoplado a un dispositivo B no debe limitarse a dispositivos o sistemas en los que una salida del dispositivo A está directamente conectada a una entrada del dispositivo B. Esto significa que existe una ruta entre una salida de A y una entrada de B que puede ser una ruta que incluye otros dispositivos o medios. "Acoplado" puede significar que dos o más elementos están en contacto físico o eléctrico directo, o que dos o más elementos no están en contacto directo entre sí, pero aún cooperan o interactúan entre sí.

Por lo tanto, aunque se han descrito realizaciones de la invención, los expertos en la materia reconocerán que se pueden hacer otras modificaciones adicionales sin apartarse del alcance definido por las reivindicaciones adjuntas, y que está destinado a reivindicar que todos esos cambios y modificaciones están dentro del alcance de la invención.

REIVINDICACIONES

1. Un método para codificar audio de entrada basado en un canal u objeto (21) para la reproducción, incluyendo el método las etapas de:

(a) renderizar inicialmente el audio de entrada basado en un canal u objeto (21) en una presentación de salida inicial;

(b) determinar (23) una estimación de una señal de componente de audio dominante (26) a partir del audio de entrada basado en un canal u objeto (21) y determinar (24) una serie de factores de ponderación de componente de audio dominante (27) para mapear la presentación de salida inicial en la señal de componente de audio dominante, para permitir la utilización de los factores de ponderación de componente de audio dominante (27) y la presentación de salida inicial para determinar la estimación de la señal de componente de audio dominante;

(c) determinar una estimación de la dirección o posición del componente de audio dominante (25); y

(d) codificar la presentación de salida inicial, los factores de ponderación de componente de audio dominante (27), la dirección o posición de componente de audio dominante (25) como la señal codificada para la reproducción,

en el que dicha presentación de salida inicial comprende una señal estéreo de mezcla descendente (29).

2. Un método de acuerdo con la reivindicación 1, que comprende además determinar una estimación de una mezcla residual que es la presentación de salida inicial menos una renderización de la señal de componente de audio dominante o la estimación de la misma.

3. Un método de acuerdo con la reivindicación 1, que comprende además generar (43) una mezcla binaural anecoica del audio de entrada basado en el un canal u objeto (21) y determinar (49) una estimación de una mezcla residual, en el que la estimación de la mezcla residual es la mezcla binaural anecoica menos una renderización de la señal de componente de audio dominante o la estimación de la misma.

4. Un método de acuerdo con la reivindicación 2 o 3, que comprende además determinar una serie de coeficientes de matriz residuales para mapear la presentación de salida inicial a la estimación de la mezcla residual.

5. El método de acuerdo con cualquiera de las reivindicaciones anteriores, en el que dicha presentación de salida inicial comprende una presentación de auriculares o altavoz.

6. El método de acuerdo con cualquier reivindicación anterior, en el que dicho audio de entrada basado en un canal u objeto (21) está en mosaico de tiempo y frecuencia y dicha etapa de codificación se repite para una serie de etapas de tiempo y una serie de bandas de frecuencia.

7. Un método para decodificar una señal de audio codificada, incluyendo la señal de audio codificada:

- una presentación de salida inicial;

- una dirección de componente de audio dominante y factores de ponderación de componente de audio dominante,

en donde dicha presentación de salida inicial comprende una señal estéreo de mezcla descendente (29);

comprendiendo el método comprende las etapas de:

(a) utilizar (63) los factores de ponderación de componente de audio dominante y la presentación de salida inicial para determinar una señal de componente dominante estimada;

(b) renderizar (65) la señal del componente dominante estimado con una binauralización en una ubicación espacial relativa a un oyente en cuestión según la dirección del componente de audio dominante para formar un componente dominante estimado binauralizado renderizado;

(c) reconstruir una estimación de componente residual a partir de la presentación de salida inicial; y

(d) combinar (66) la señal de componente dominante estimada binauralizada renderizada y la estimación de componente residual para formar una señal codificada de audio espacializada de salida.

8. Un método de acuerdo con la reivindicación 7, en el que dicha señal de audio codificada incluye además una serie de coeficientes de matriz residuales que representan una señal de audio residual y dicha etapa (c) comprende además:

(c1) aplicar (64) dichos coeficientes de matriz residuales a la presentación de salida inicial para reconstruir la estimación de componente residual.

9. Un método de acuerdo con la reivindicación 7, en el que la estimación de componente residual se reconstruye restando el componente dominante estimado binauralizado renderizado de la presentación de salida inicial.

10. Un método de acuerdo con una cualquiera de las reivindicaciones 7 a 9, en el que dicha etapa (b) incluye una rotación inicial de la señal de componente dominante estimada de acuerdo con una señal de entrada de seguimiento de la cabeza que indica la orientación de la cabeza de un oyente en cuestión.

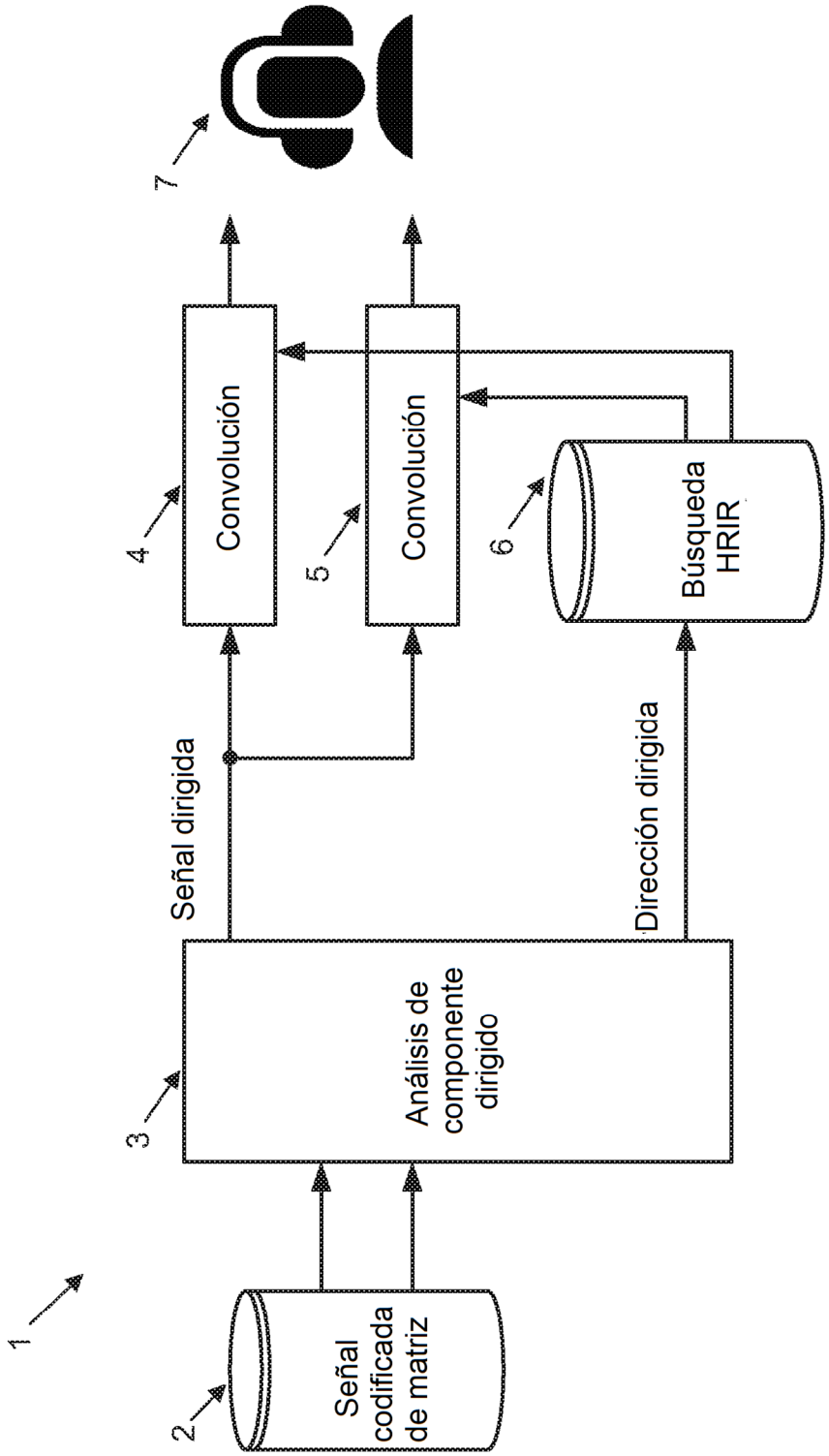
1
2
Señal codificada de matriz
3
Análisis de componente dirigido
Señal dirigida
Dirección dirigida
4
Convolución
5
Convolución
6
Búsqueda HRIR
7

FIG. 1

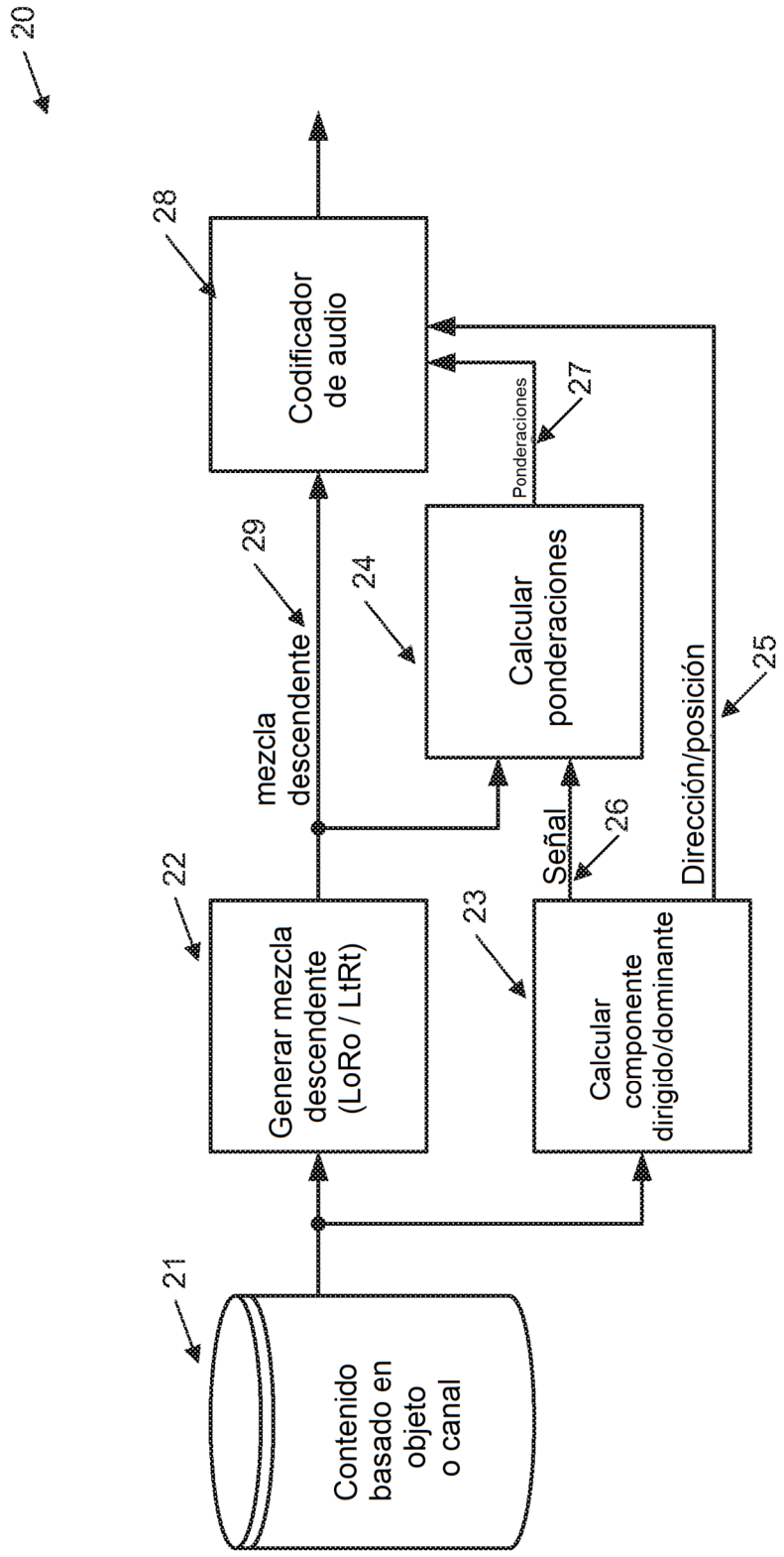
20
28
Codificador de audio
27
Ponderaciones
29
mezcla descendente
24
Calcular ponderaciones
25
Dirección/posición
Señal
26
22
Generar mezcla descendente (LoRo / LtRt)
23
Calcular componente dirigido/dominante
21
Contenido basado en objeto o canal

FIG. 2

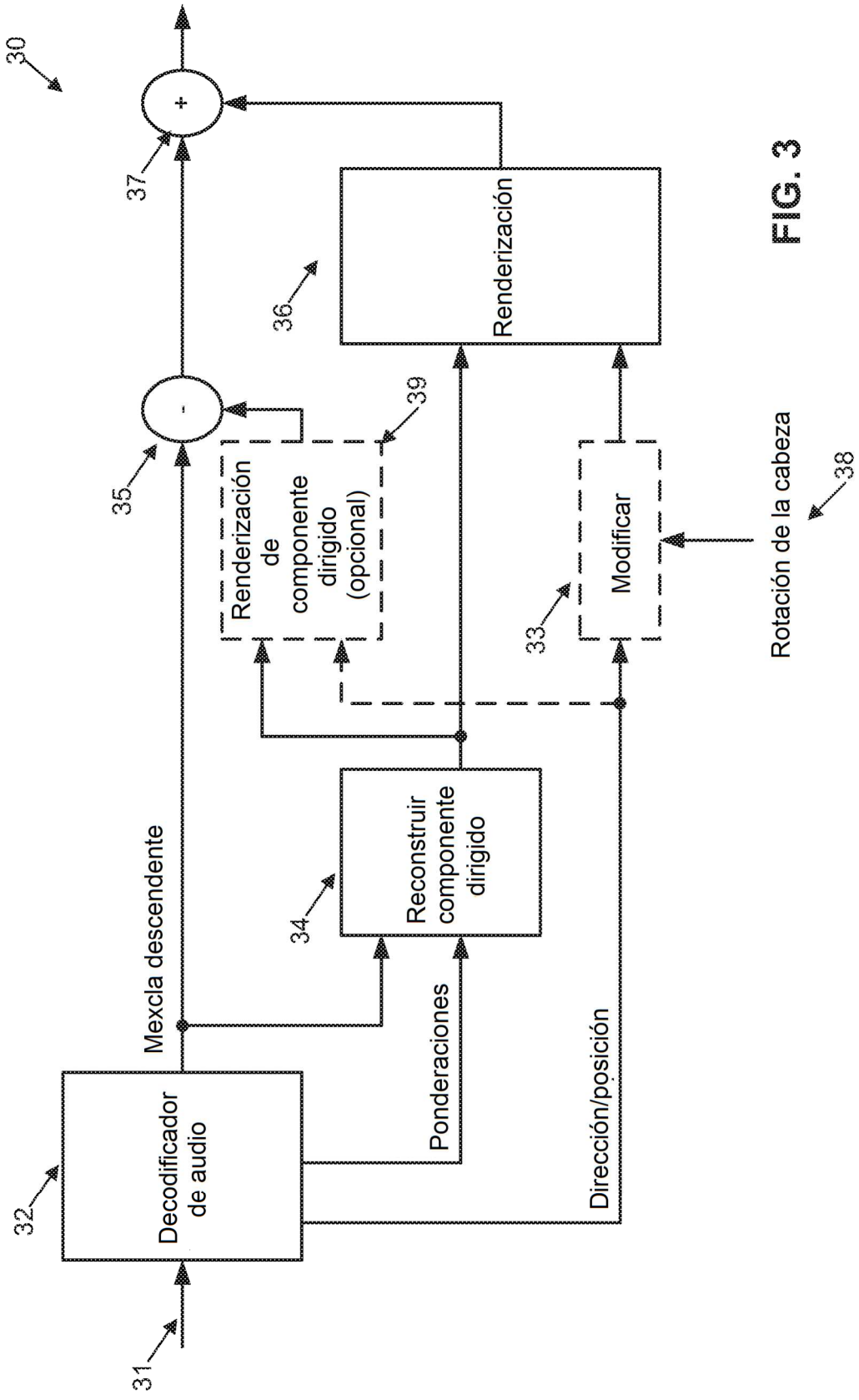
30
37
+
36
Renderización
35
-
39
Renderización de componente dirigido (opcional)
33
Modificar
38
Rotación de la cabeza
Mexcla descendente
34
Reconstruir componente dirigido
Ponderaciones
Dirección/posición
32
Decodificador de audio
31

FIG. 3

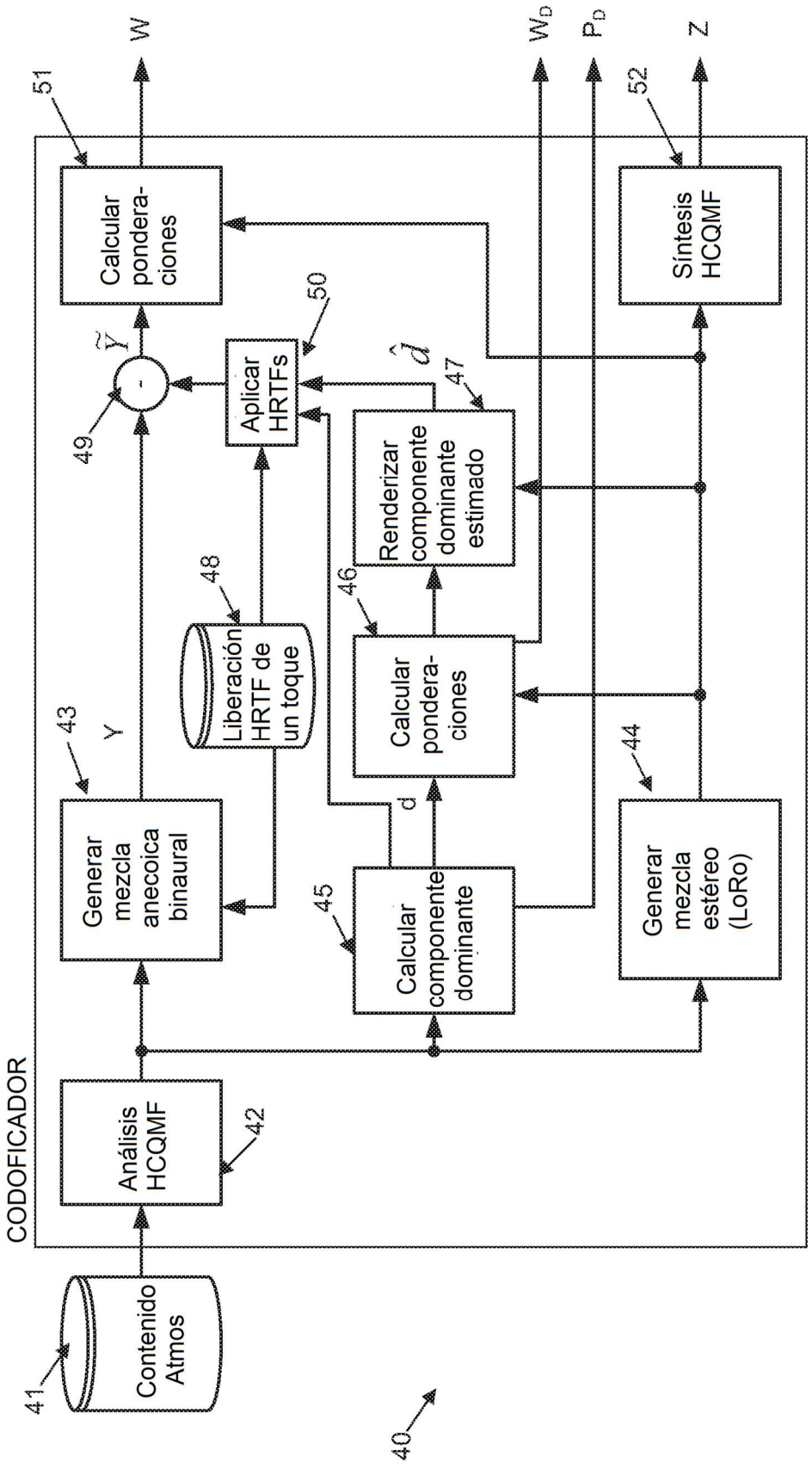
CODOFICADOR
Contenido Atmos
41
Análisis HCQMF
42
Generar mezcla anecoica binaural
43
Y
Liberación HRTF de un toque
48
Aplicar HRTFs
50
Calcular pondera-ciones
51
W
Calcular componente dominante
45
d
Calcular pondera-ciones
46
Renderizar componente dominante estimado
47
$\hat{d}$
$\tilde{Y}$
49
-
Generar mezcla estéreo (LoRo)
44
Síntesis HCQMF
52
W_D
P_D
Z
40

FIG. 4

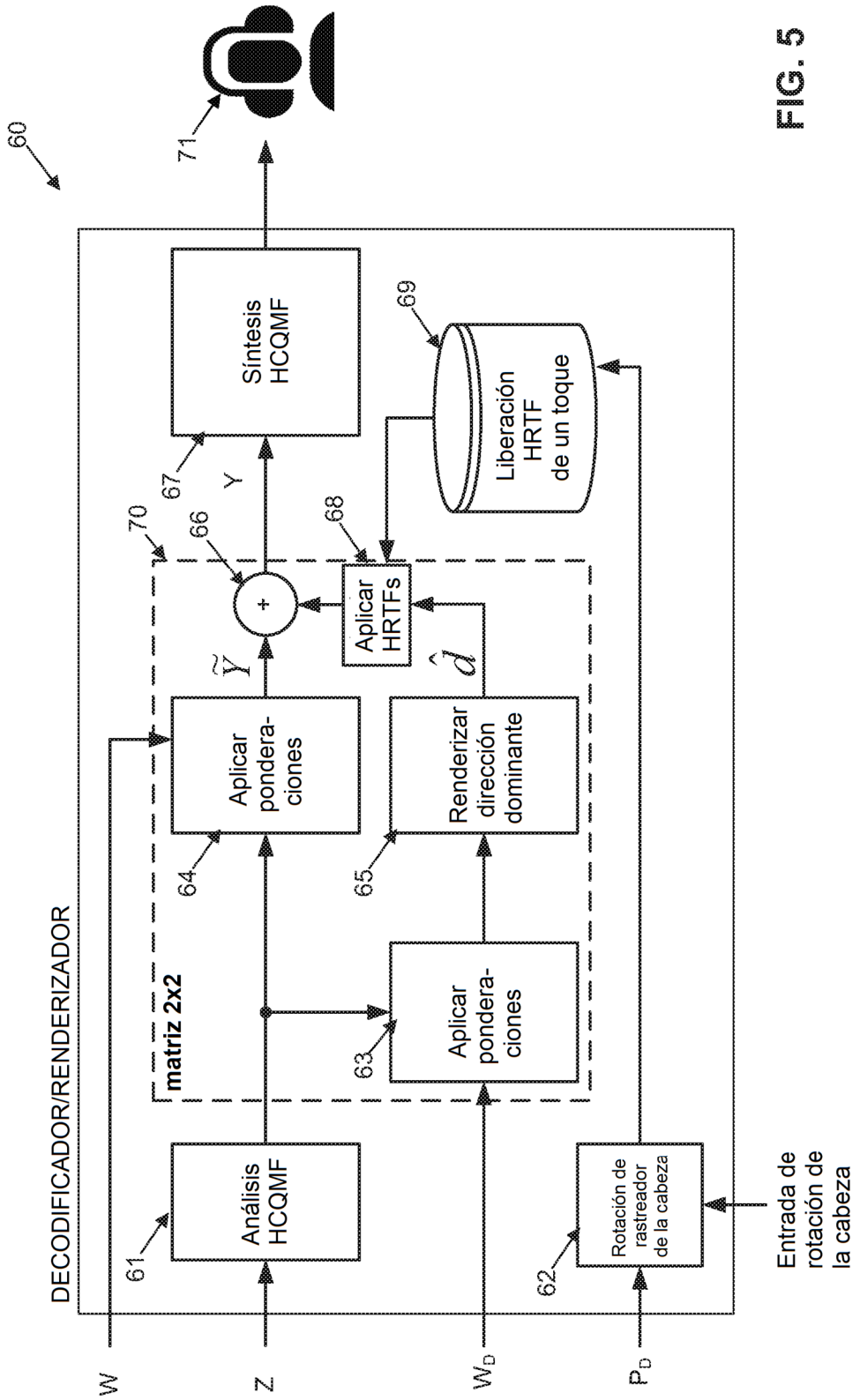
60
71
DECODIFICADOR/RENDERIZADOR
Sintesis HCQMF
67
Y
69
Liberación HRTF de un toque
70
66
68
Aplicar HRTFs
$\tilde{Y}$
$\hat{d}$
Aplicar pondera-ciones
Renderizar dirección dominante
64
65
matriz 2x2
63
Aplicar pondera-ciones
Análisis HCQMF
61
62
Rotación de rastreador de la cabeza
Entrada de rotación de la cabeza
W
Z
W_D
P_D

FIG. 5