(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 700 246**

(51) Int. CI.:

G10L 19/008 (2013.01)
G10L 19/20 (2013.01)
G10L 19/22 (2013.01)
G10L 21/02 (2013.01)
G10L 21/0324 (2013.01)
G10L 21/0364 (2013.01)
H04R 5/04 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **27.08.2014 PCT/US2014/052962**

(87) Fecha y número de publicación internacional: **05.03.2015 WO15031505**

(96) Fecha de presentación y número de la solicitud europea: **27.08.2014 E 14762180 (9)**

(97) Fecha y número de publicación de la concesión europea: **03.10.2018 EP 3039675**

(54) Título: **Mejora paramétrica de la voz**

(30) Prioridad:

28.08.2013 US 201361870933 P
25.10.2013 US 201361895959 P
25.11.2013 US 201361908664 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
14.02.2019

(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:

KOPPENS, JEROEN y
MUESCH, HANNES

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

DESCRIPCIÓN

Mejora paramétrica de la voz.

Referencia cruzada a solicitudes relacionadas

La presente solicitud reivindica la prioridad de la Solicitud de Patente Provisional de los Estados Unidos No. 61/870,933, presentada el 28 de agosto de 2013, Solicitud de Patente Provisional de los Estados Unidos No. 61/895,959, presentada el 25 octubre de 2013 y Solicitud de Patente Provisional de los Estados Unidos No. 61/908,664, presentada el 25 de noviembre de 2013.

Tecnología

La presente descripción pertenece al procesamiento de señales de audio y, más concretamente, a la mejora del contenido de voz de un programa de audio con respecto a otro contenido del programa, en el cual la mejora de la voz es "híbrida" en el sentido de que incluye una mejora con forma de onda codificada (o una mejora con forma de onda relativamente más codificada) en algunas condiciones de señal y mejora paramétrica codificada (o mejora paramétricamente relativamente más codificada) en otras condiciones de señal. Otros aspectos son la codificación, decodificación y reproducción de programas de audio que incluyen datos suficientes para permitir dicha mejora de voz híbrida.

Antecedentes

En las películas y en la televisión, el diálogo y la narrativa se presentan, con frecuencia, junto con otro audio diferente de la voz como, por ejemplo, música, efectos, o ambiente de eventos deportivos. En muchos casos, los sonidos de voz y no voz se capturan de forma separada y se mezclan juntos bajo el control de un ingeniero de sonido. El ingeniero de sonido selecciona el nivel de la voz en relación con el nivel de la no voz en una manera que es apropiada para la mayoría de los oyentes. Sin embargo, algunos oyentes, p.ej., aquellos con una lesión auditiva, experimentan dificultades para comprender el contenido de voz de programas de audio (que tienen relaciones de mezcla de voz-no voz determinadas por un ingeniero) y preferirían que la voz se mezclara en un nivel relativo más alto.

Existe un problema a resolver para permitir que dichos oyentes aumenten la audibilidad del contenido de voz de programa de audio con respecto a la del contenido de audio de no voz.

Un enfoque actual es proveer a los oyentes dos corrientes de audio de alta calidad. Una corriente lleva audio de contenido primario (principalmente voz) y la otra lleva audio de contenido secundario (el programa de audio restante, que excluye la voz) y se provee al usuario el control sobre el proceso de mezcla. Desafortunadamente, el presente esquema no es práctico porque no se basa en la práctica actual de transmitir un programa de audio totalmente mixto. Además, requiere aproximadamente dos veces el ancho de banda de la práctica de radiodifusión actual ya que dos corrientes de audio independientes, cada una de calidad de radiodifusión, deben entregarse al usuario.

Otro método de mejora de la voz (al que, en la presente memoria, se hará referencia como mejora "con forma de onda codificada") se describe en la Publicación de Solicitud de Patente de los Estados Unidos No. 2010/0106507 A1, publicada el 29 de abril de 2010, asignada a Dolby Laboratories, Inc., y que nombra a Hannes Muesch como inventor. En la mejora con forma de onda codificada, la relación voz-fondo (no voz) de una mezcla de audio original de contenido de voz y no voz (a la que a veces se hace referencia como una mezcla principal) aumenta por la incorporación a la mezcla principal de una versión de calidad reducida (copia de baja calidad) de la señal de voz limpia que se ha enviado al receptor junto a la mezcla principal. Con el fin de reducir la sobrecarga del ancho de banda, la copia de baja calidad se codifica normalmente a una velocidad binaria muy baja. Debido a la codificación a velocidad binaria baja, los artefactos de codificación se asocian a la copia de baja calidad, y los artefactos de codificación son claramente audibles cuando la copia de baja calidad se reproduce y se escucha de manera aislada. Por consiguiente, la copia de baja calidad tiene una calidad objetable cuando se escucha de manera aislada. La mejora con forma de onda codificada intenta ocultar dichos artefactos de codificación mediante la incorporación de la copia de baja calidad a la mezcla principal solamente durante los tiempos en los cuales el nivel de los componentes de no voz es alto de modo que los artefactos de codificación se enmascaran por los componentes de no voz. Como se detallará más adelante, las limitaciones del presente enfoque incluyen lo siguiente: la cantidad de mejora de la voz normalmente no puede ser constante en el tiempo, y los artefactos de audio pueden convertirse en audibles cuando los componentes de fondo (no voz) de la mezcla principal son débiles o su espectro frecuencia-amplitud difiere radicalmente de aquel del ruido de codificación.

Según la mejora con forma de onda codificada, un programa de audio (para la entrega a un decodificador para la decodificación y subsiguiente reproducción) se codifica como un tren de bits que incluye la copia de voz de baja calidad (o una versión codificada de aquella) como un tren paralelo de la mezcla principal. El tren de bits puede incluir metadatos indicativos de un parámetro de escalamiento que determina la cantidad de mejora de la voz con forma de onda codificada que se llevará a cabo (a saber, el parámetro de escalamiento determina un factor de

escalamiento que se aplicará a la copia de voz de baja calidad antes de que la copia de voz de baja calidad escalada se combine con la mezcla principal, o un valor máximo de dicho factor de escalamiento que asegurará el enmascaramiento de artefactos de codificación). Cuando el valor actual del factor de escalamiento es cero, el decodificador no lleva a cabo la mejora de la voz en el segmento correspondiente de la mezcla principal. El valor actual del parámetro de escalamiento (o el valor máximo actual que puede alcanzar) se determina normalmente en el codificador (dado que se genera normalmente por un modelo psicoacústico computacionalmente intensivo), pero puede generarse en el decodificador. En el último caso, ningún metadato indicativo del parámetro de escalamiento necesitará enviarse del codificador al decodificador, y el decodificador, en su lugar, puede determinar a partir de la mezcla principal una relación de potencia del contenido de voz de la mezcla con respecto a la potencia de la mezcla e implementar un modelo para determinar el valor actual del parámetro de escalamiento en respuesta al valor actual de la relación de potencia.

Otro método (al que se hará referencia, en la presente memoria, como mejora "paramétrica codificada") para mejorar la inteligibilidad de la voz en presencia de audio en competencia (fondo) es segmentar el programa de audio original (normalmente, una pista de audio) en losas tiempo/frecuencia e impulsar las losas según la relación de la potencia (o nivel) de su contenido de voz y fondo, para lograr un impulso del componente de voz con respecto al fondo. La idea subyacente del presente enfoque es similar a la de la supresión de ruido de sustracción espectral guiada. En un ejemplo extremo del presente enfoque, en el cual todas las losas con SNR (a saber, la relación de potencia, o nivel, del componente de voz con respecto a la del contenido de sonido en competencia) debajo de un umbral predeterminado se suprimen completamente, se ha demostrado que provee mejoras de inteligibilidad de la voz robustas. En la aplicación del presente método a la radiodifusión, la relación voz-fondo (SNR) puede inferirse mediante comparación de la mezcla de audio original (de contenido de voz y no voz) y el componente de voz de la mezcla. La SNR inferida puede entonces transformarse en un conjunto apropiado de parámetros de mejora que se transmiten junto a la mezcla de audio original. En el receptor, dichos parámetros pueden (opcionalmente) aplicarse a la mezcla de audio original para derivar una señal indicativa de voz mejorada. Como se detallará más adelante, la mejora paramétrica codificada funciona mejor cuando la señal de voz (el componente de voz de la mezcla) domina a la señal de fondo (el componente de no voz de la mezcla).

La mejora con forma de onda codificada requiere que una copia de baja calidad del componente de voz de un programa de audio entregado se encuentre disponible en el receptor. Con el fin de limitar la sobrecarga de datos incurrida al transmitir dicha copia junto a la mezcla de audio principal, dicha copia se codifica a una velocidad binaria muy baja y muestra distorsiones de codificación. Dichas distorsiones de codificación se enmascararán probablemente por el audio original cuando el nivel de los componentes de no voz sea alto. Cuando las distorsiones de codificación se enmascaran, la calidad resultante del audio mejorado es muy buena.

La mejora paramétrica codificada se basa en el análisis de la señal de mezcla de audio principal en losas tiempo/frecuencia y la aplicación de ganancias/atenuaciones apropiadas a cada una de dichas losas, similar a la normalización de diálogo basada en parámetros, según se describe en el Comité de Sistemas de Televisión Avanzada (ATSC, por sus siglas en inglés), "*ATSC Standard: Digital Audio Compression (AC-3, E-AC-3)*", 17 diciembre 2012, XP055295947. La velocidad de datos que se necesita para retransmitir dichas ganancias al receptor es baja cuando se compara con la de la mejora con forma de onda codificada. Sin embargo, debido a la resolución temporal-espectral limitada de los parámetros, la voz, cuando se mezcla con audio de no voz, no puede manipularse sin también afectar el audio de no voz. La mejora paramétrica codificada del contenido de voz de una mezcla de audio introduce, por consiguiente, la modulación en el contenido de no voz de la mezcla y dicha modulación ("modulación de fondo") puede convertirse en objetable después de la reproducción de la mezcla de voz mejorada. Las modulaciones de fondo son más probablemente objetables cuando la relación voz-fondo es muy baja.

Los enfoques descritos en la presente sección son enfoques que pueden perseguirse, pero no necesariamente enfoques que se han concebido o perseguido previamente. Por lo tanto, salvo que se indique lo contrario, no debe suponerse que cualquiera de los enfoques descritos en la presente sección puede considerarse una técnica anterior meramente en virtud de su inclusión en la presente sección. De manera similar, no debe suponerse que las cuestiones identificadas con respecto a uno o más enfoques se han reconocido en cualquier técnica anterior según la presente sección, a menos que se indique lo contrario.

Breve descripción de los dibujos

La presente invención se ilustra a modo de ejemplo, y no a modo de restricción, en las figuras de los dibujos anexos y en los cuales iguales numerales de referencia se refieren a elementos similares y en los cuales:

La Figura 1 es un diagrama de bloques de un sistema configurado para generar parámetros de predicción para reconstruir el contenido de voz de una señal de contenido mixto de un solo canal (que tiene contenido de voz y no voz).

La Figura 2 es un diagrama de bloques de un sistema configurado para generar parámetros de predicción para reconstruir el contenido de voz de una señal de contenido mixto multicanal (que tiene contenido de voz y no voz).

La Figura 3 es un diagrama de bloques de un sistema que incluye un codificador configurado para llevar a cabo un ejemplo del método de codificación descrito para generar un tren de bits de audio codificado indicativo de un programa de audio, y un decodificador configurado para decodificar y llevar a cabo la mejora de la voz (según una realización del método descrito) en el tren de bits de audio codificado.

La Figura 4 es un diagrama de bloques de un sistema configurado para reproducir una señal de audio de contenido mixto multicanal, que incluye llevar a cabo la mejora de la voz convencional en aquella.

La Figura 5 es un diagrama de bloques de un sistema configurado para reproducir una señal de audio de contenido mixto multicanal, que incluye llevar a cabo la mejora de la voz paramétrica codificada convencional en aquella.

La Figura 6 y Figura 6A son diagramas de bloques de sistemas configurados para reproducir una señal de audio de contenido mixto multicanal, que incluye llevar a cabo una realización del método de mejora de la voz descrito en aquella.

La Figura 7 es un diagrama de bloques de un sistema para llevar a cabo un ejemplo del método de codificación descrito mediante el uso de un modelo de enmascaramiento auditivo;

la Figura 8A y Figura 8B ilustran flujos de proceso a modo de ejemplo; y

la Figura 9 ilustra una plataforma de hardware a modo de ejemplo en la cual un ordenador o un dispositivo informático según se describe en la presente memoria pueden implementarse.

Descripción de realizaciones a modo de ejemplo

Las realizaciones a modo de ejemplo, que se refieren a la mejora de la voz con forma de onda codificada y paramétrica codificada híbrida, se describen en la presente memoria. En la siguiente descripción, en aras de la explicación, se establecen numerosos detalles específicos con el fin de proveer una comprensión exhaustiva de la presente invención. Será aparente, sin embargo, que la presente invención puede practicarse sin dichos detalles específicos. En otras instancias, estructuras y dispositivos conocidos no se describen en detalle exhaustivo, con el fin de evitar la oclusión, oscurecimiento u ofuscación innecesarias de la presente invención.

Las realizaciones a modo de ejemplo se describen en la presente memoria según el siguiente resumen:

1. Resumen general

2. Notación y Nomenclatura

3. Generación de parámetros de predicción

4. Funciones de mejora de la voz

5. Reproducción de la voz

6. Representación media/lateral

7. Flujos de proceso a modo de ejemplo

8. Mecanismos de implementación - Resumen de hardware

9. Equivalentes, extensiones, alternativas y varios

1. Resumen general

La invención se define por las reivindicaciones independientes anexas. Realizaciones a modo de ejemplo se proveen por las reivindicaciones dependientes. El presente resumen presenta una descripción básica de algunos aspectos de una realización de la presente invención. Debe notarse que el presente resumen no es un resumen extensivo o exhaustivo de aspectos de la realización. Además, debe notarse que el presente resumen no pretende comprenderse como uno que identifica aspectos o elementos particularmente significativos de la realización, o que delinea el alcance de la realización en particular, o de la invención en general. El presente resumen meramente presenta algunos conceptos que se refieren a la realización a modo de ejemplo en un formato condensado y simplificado, y debe comprenderse meramente como un preludio conceptual a una descripción más detallada de realizaciones a modo de ejemplo que sigue más abajo. Es preciso notar que, aunque realizaciones separadas se describen en la presente memoria, cualquier combinación de realizaciones y/o realizaciones parciales descritas en la presente memoria pueden combinarse para formar realizaciones adicionales

Las fortalezas y debilidades individuales de la mejora paramétrica codificada y mejora con forma de onda codificada pueden compensarse entre sí, y la mejora de la voz convencional puede, en algunas realizaciones, mejorarse

sustancialmente por un método de mejora híbrido que emplea la mejora paramétrica codificada (o una mezcla de mejora paramétrica codificada y mejora con forma de onda codificada) en algunas condiciones de señal y mejora con forma de onda codificada (o una mezcla diferente de mejora paramétrica codificada y con forma de onda codificada) en otras condiciones de señal. Realizaciones típicas del método de mejora híbrido proveen una mejora de la voz más coherente y de mejor calidad que la que puede lograrse por la mejora paramétrica codificada o con forma de onda codificada solas.

En una clase de realizaciones, el método incluye las etapas de: (a) recibir un tren de bits indicativo de un programa de audio que incluye voz que tiene una forma de onda no mejorada y otro contenido de audio, en donde el tren de bits incluye: datos de audio indicativos de la voz y el otro contenido de audio, datos de forma de onda indicativos de una versión de calidad reducida de la voz (donde los datos de audio se han generado mediante la mezcla de datos de voz con datos de no voz, los datos de forma de onda comprenden, normalmente, menos bits que los datos de voz, en donde la versión de calidad reducida tiene una segunda forma de onda similar (p.ej., al menos sustancialmente similar) a la forma de onda no mejorada, y la versión de calidad reducida tendrá una calidad objetable si se escucha de manera aislada, y datos paramétricos, en donde los datos paramétricos con los datos de audio determinan la voz paramétricamente construida, y la voz paramétricamente construida es una versión paramétricamente reconstruida de la voz que concuerda al menos parcialmente con (p.ej., es una buena aproximación de) la voz; y (b) llevar a cabo la mejora de la voz en el tren de bits en respuesta a un indicador de mezcla y, de esta manera, generar datos indicativos de un programa de audio de voz mejorada, incluso mediante la combinación de los datos de audio con una combinación de datos de voz de baja calidad determinados a partir de datos de forma de onda, y datos de voz reconstruida, en donde la combinación se determina por el indicador de mezcla (p.ej., la combinación tiene una secuencia de estados determinada por una secuencia de valores actuales del indicador de mezcla), los datos de voz reconstruida se generan en respuesta a al menos algunos de los datos paramétricos y al menos algunos de los datos de audio, y el programa de audio de voz mejorada tiene artefactos de mejora de la voz menos audibles (p.ej., artefactos de mejora de la voz que se enmascaran mejor y, por consiguiente, son menos audibles cuando el programa de audio de voz mejorada se reproduce y escucha) que lo que serían en un programa de audio de voz mejorada con forma de onda puramente codificada determinado mediante la combinación de solamente los datos de voz de baja calidad (que son indicativos de la versión de calidad reducida de la voz) con los datos de audio o un programa de audio de voz mejorada paramétrico codificado puramente determinado a partir de los datos paramétricos y los datos de audio.

En la presente memoria, "artefacto de mejora de la voz" (o "artefacto de codificación de mejora de la voz") denota una distorsión (normalmente, una distorsión medible) de una señal de audio (indicativa de una señal de voz y de una señal de audio de no voz) provocada por una representación de la señal de voz (p.ej., señal de voz con forma de onda codificada, o datos paramétricos en conjunto con la señal de contenido mixto).

En algunas realizaciones, el indicador de mezcla (que puede tener una secuencia de valores, p.ej., uno para cada uno de una secuencia de segmentos de tren de bits) se incluye en el tren de bits recibido en la etapa (a). Algunas realizaciones incluyen una etapa de generación del indicador de mezcla (p.ej., en un receptor que recibe y decodifica el tren de bits) en respuesta al tren de bits recibido en la etapa (a).

Debe comprenderse que la expresión "indicador de mezcla" no pretende requerir que el indicador de mezcla sea un solo parámetro o valor (o una secuencia de parámetros o valores únicos) para cada segmento del tren de bits. Más bien, se contempla que, en algunas realizaciones, un indicador de mezcla (para un segmento del tren de bits) puede ser un conjunto de dos o más parámetros o valores (p.ej., para cada segmento, un parámetro de control de mejora paramétrica codificada, y un parámetro de control de mejora con forma de onda codificada) o una secuencia de conjuntos de parámetros o valores.

En algunas realizaciones, el indicador de mezcla para cada segmento puede ser una secuencia de valores que indican la mezcla por banda de frecuencia del segmento.

Los datos de forma de onda y los datos paramétricos no necesitan proveerse para (p.ej., incluirse en) cada segmento del tren de bits, y tanto los datos de forma de onda como los datos paramétricos no necesitan usarse para llevar a cabo la mejora de la voz en cada segmento del tren de bits. Por ejemplo, en algunos casos, al menos un segmento puede incluir datos de forma de onda solamente (y la combinación determinada por el indicador de mezcla para cada segmento puede consistir en datos de forma de onda solamente) y al menos otro segmento puede incluir datos paramétricos solamente (y la combinación determinada por el indicador de mezcla para cada segmento puede consistir en datos de voz reconstruida solamente).

Se contempla que, normalmente, un codificador genera el tren de bits incluso mediante la codificación (p.ej., compresión) de los datos de audio, pero no mediante la aplicación de la misma codificación a los datos de forma de onda o datos paramétricos. Por consiguiente, cuando el tren de bits se entrega a un receptor, el receptor normalmente analizará el tren de bits para extraer los datos de audio, los datos de forma de onda y los datos paramétricos (y el indicador de mezcla si se entrega en el tren de bits), pero decodificará solamente los datos de audio. El receptor normalmente llevará a cabo la mejora de la voz en los datos de audio decodificados (mediante el

uso de datos de forma de onda y/o datos paramétricos) sin aplicar a los datos de forma de onda o datos paramétricos el mismo proceso de decodificación que se aplica a los datos de audio.

Normalmente, la combinación (indicada por el indicador de mezcla) de los datos de forma de onda y los datos de voz reconstruida cambia con el tiempo, cada estado de la combinación perteneciendo al contenido de voz y otro contenido de audio de un segmento correspondiente del tren de bits. El indicador de mezcla se genera de modo que el estado actual de la combinación (de datos de forma de onda y datos de voz reconstruida) se determina al menos parcialmente por propiedades de señal de la voz y otro contenido de audio (p.ej., una relación de la potencia del contenido de voz y la potencia de otro contenido de audio) en el segmento correspondiente del tren de bits. En algunas realizaciones, el indicador de mezcla se genera de modo que el estado actual de la combinación se determina por propiedades de señal de la voz y otro contenido de audio en el segmento correspondiente del tren de bits. En algunas realizaciones, el indicador de mezcla se genera de modo que el estado actual de la combinación se determina tanto por propiedades de señal de la voz y otro contenido de audio en el segmento correspondiente del tren de bits como por una cantidad de artefactos de codificación en los datos de forma de onda.

La etapa (b) puede incluir una etapa de llevar a cabo la mejora de la voz con forma de onda codificada mediante la combinación (p.ej., mezcla) de al menos algunos de los datos de voz de baja calidad con los datos de audio de al menos un segmento del tren de bits, y llevar a cabo la mejora de la voz paramétrica codificada mediante la combinación de los datos de voz reconstruida con los datos de audio de al menos un segmento del tren de bits. Una combinación de la mejora de la voz con forma de onda codificada y mejora de la voz paramétrica codificada se lleva a cabo en al menos un segmento del tren de bits mediante la mezcla tanto de los datos de voz de baja calidad como de la voz paramétricamente construida para el segmento con los datos de audio del segmento. En algunas condiciones de señal, solo una (pero no ambas) de la mejora de la voz con forma de onda codificada y mejora de la voz paramétrica codificada se lleva a cabo (en respuesta al indicador de mezcla) en un segmento (o en cada uno de más de un segmento) del tren de bits.

En la presente memoria, la expresión "SNR" (relación señal-ruido) se usará para denotar la relación de potencia (o diferencia en nivel) del contenido de voz de un segmento de un programa de audio (o de todo el programa) con respecto a la del contenido de no voz del segmento o programa, o del contenido de voz de un segmento del programa (o todo el programa) con respecto a la de todo el contenido (voz y no voz) del segmento o programa.

En una clase de realizaciones, el método implementa la conmutación basada en SNR temporal "a ciegas" entre la mejora paramétrica codificada y la mejora con forma de onda codificada de segmentos de un programa de audio. En el presente contexto, "a ciegas" denota que la conmutación no se guía perceptualmente por un modelo de enmascaramiento auditivo complejo (p.ej., de un tipo que se describirá en la presente memoria), sino que se guía por una secuencia de valores SRN (indicadores de mezcla) correspondientes a segmentos del programa. En una realización en la presente clase, la mejora de la voz híbrida codificada se logra por la conmutación temporal entre mejora paramétrica codificada y mejora con forma de onda codificada, de modo que la mejora paramétrica codificada o mejora con forma de onda codificada (pero no la mejora paramétrica codificada y mejora con forma de onda codificada) se lleva a cabo en cada segmento de un programa de audio en el cual la mejora de la voz se lleva a cabo. Mediante el reconocimiento de que la mejora con forma de onda codificada tiene un mejor rendimiento en la condición de SNR baja (en segmentos que tienen valores bajos de SNR) y la mejora paramétrica codificada tiene un mejor rendimiento en SNR favorables (en segmentos que tienen valores altos de SNR), la decisión de conmutación se basa normalmente en la relación de voz (diálogo) con respecto al audio restante en una mezcla de audio original.

Las realizaciones que implementan la conmutación basada en SNR temporal "a ciegas" normalmente incluyen las etapas de: segmentar la señal de audio no mejorada (mezcla de audio original) en segmentos de tiempo consecutivos, y determinar para cada segmento la SNR entre el contenido de voz y el otro contenido de audio (o entre el contenido de voz y el contenido de audio total) del segmento; y para cada segmento, comparar la SNR con un umbral y proveer un parámetro de control de mejora paramétrica codificada para el segmento (a saber, el indicador de mezcla para el segmento indica que la mejora paramétrica codificada debe llevarse a cabo) cuando la SNR es mayor que el umbral o proveer un parámetro de control de mejora con forma de onda codificada para el segmento (a saber, el indicador de mezcla para el segmento indica que la mejora con forma de onda codificada debe llevarse a cabo) cuando la SNR no es mayor que el umbral. Normalmente, la señal de audio no mejorada se entrega (p.ej., se transmite) con los parámetros de control incluidos como metadatos a un receptor, y el receptor lleva a cabo (en cada segmento) el tipo de mejora de la voz indicado por el parámetro de control para el segmento. Por consiguiente, el receptor lleva a cabo la mejora paramétrica codificada en cada segmento para el cual el parámetro de control es un parámetro de control de mejora paramétrica codificada, y la mejora con forma de onda codificada en cada segmento para el cual el parámetro de control es un parámetro de control de mejora con forma de onda codificada.

Si una persona desea incurrir en el coste de transmisión (con cada segmento de una mezcla de audio original) tanto datos de forma de onda (para implementar la mejora de la voz con forma de onda codificada) como parámetros de mejora paramétrica codificada con una mezcla original (no mejorada), un grado más alto de mejora de la voz puede lograrse mediante la aplicación tanto de la mejora con forma de onda codificada como de la mejora paramétrica

codificada a segmentos individuales de la mezcla. Por consiguiente, en una clase de realizaciones, el método implementa la mezcla basada en SNR temporal "a ciegas" entre la mejora paramétrica codificada y la mejora con forma de onda codificada de segmentos de un programa de audio. En el presente contexto, "a ciegas" denota que la conmutación no se guía perceptualmente por un modelo de enmascaramiento auditivo complejo (p.ej., de un tipo que se describirá en la presente memoria), sino que se guía por una secuencia de valores SRN correspondientes a segmentos del programa.

Las realizaciones que implementan la mezcla basada en SNR temporal "a ciegas" normalmente incluyen las etapas de: segmentar la señal de audio no mejorada (mezcla de audio original) en segmentos de tiempo consecutivos, y determinar para cada segmento la SNR entre el contenido de voz y el otro contenido de audio (o entre el contenido de voz y el contenido de audio total) del segmento; y para cada segmento, proveer un indicador de control de mezcla, donde el valor del indicador de control de mezcla se determina por (es una función de) la SNR para el segmento.

En algunas realizaciones, el método incluye una etapa de determinar (p.ej., recibir una solicitud de) una cantidad total ("T") de mejora de la voz, y el indicador de control de mezcla es un parámetro, α, para cada segmento de modo que $T = \alpha Pw + (1-\alpha)Pp$, donde Pw es una mejora con forma de onda codificada para el segmento que producirá la cantidad total predeterminada de mejora, T, si se aplica al contenido de audio no mejorado del segmento mediante el uso de datos de forma de onda provistos para el segmento (donde el contenido de voz del segmento tiene una forma de onda no mejorada, los datos de forma de onda para el segmento son indicativos de una versión de calidad reducida del contenido de voz del segmento, la versión de calidad reducida tiene una forma de onda similar (p.ej., al menos sustancialmente similar) a la forma de onda no mejorada, y la versión de calidad reducida del contenido de voz es de una calidad objetable cuando se reproduce y percibe de manera aislada), y Pp es una mejora paramétrica codificada que producirá la cantidad total predeterminada de mejora, T, si se aplica a contenido de audio no mejorado del segmento mediante el uso de datos paramétricos provistos para el segmento (donde los datos paramétricos para el segmento, con el contenido de audio no mejorado del segmento, determinan una versión paramétricamente reconstruida del contenido de voz del segmento). En algunas realizaciones, el indicador de control de mezcla para cada uno de los segmentos es un conjunto de dichos parámetros, incluido un parámetro para cada banda de frecuencia del segmento relevante.

Cuando la señal de audio no mejorada se entrega (p.ej., se transmite) con los parámetros de control como metadatos a un receptor, el receptor puede llevar a cabo (en cada segmento) la mejora de la voz híbrida indicada por los parámetros de control para el segmento. De manera alternativa, el receptor genera los parámetros de control a partir de la señal de audio no mejorada.

En algunas realizaciones, el receptor lleva a cabo (en cada segmento de la señal de audio no mejorada) una combinación de mejora paramétrica codificada (en una cantidad determinada por la mejora Pp escalada por el parámetro α para el segmento) y mejora con forma de onda codificada (en una cantidad determinada por la mejora Pw escalada por el valor (1 - α) para el segmento), de modo que la combinación de mejora paramétrica codificada y mejora con forma de onda codificada genera la cantidad total predeterminada de mejora:

$$T = \alpha Pw + (1-\alpha)Pp \qquad (1)$$

En otra clase de realizaciones, la combinación de mejora con forma de onda codificada y paramétrica codificada que se llevará a cabo en cada segmento de una señal de audio se determina por un modelo de enmascaramiento auditivo. En algunas realizaciones en la presente clase, la relación de mezcla óptima para una mezcla de mejora con forma de onda codificada y mejora paramétrica codificada que se llevará a cabo en un segmento de un programa de audio usa la cantidad más alta de mejora con forma de onda codificada que simplemente evita que el ruido de codificación se convierta en audible. Debe apreciarse que la disponibilidad del ruido de codificación en un decodificador es siempre en la forma de un cálculo estadístico y no puede determinarse de forma exacta.

En algunas realizaciones en la presente clase, el indicador de mezcla para cada segmento de los datos de audio es indicativo de una combinación de mejora con forma de onda codificada y paramétrica codificada que se llevará a cabo en el segmento, y la combinación es al menos sustancialmente igual a una combinación de maximización con forma de onda codificada determinada para el segmento por el modelo de enmascaramiento auditivo, donde la combinación de maximización con forma de onda codificada especifica una cantidad relativa más grande de mejora con forma de onda codificada que asegura que el ruido de codificación (debido a la mejora con forma de onda codificada) en el segmento correspondiente del programa de audio de voz mejorada no sea audible de manera objetable (p.ej., no es audible). En algunas realizaciones, la cantidad relativa más grande de mejora con forma de onda codificada que asegura que el ruido de codificación en un segmento del programa de audio de voz mejorada no sea audible de manera objetable es la cantidad relativa más grande que asegura que la combinación de mejora con forma de onda codificada y mejora paramétrica codificada que se llevarán a cabo (en un segmento correspondiente de datos de audio) genera una cantidad total predeterminada de mejora de la voz para el segmento, y/o (donde artefactos de la mejora paramétrica codificada se incluyen en la evaluación llevada a cabo por el modelo de enmascaramiento auditivo) puede permitir que artefactos de codificación (debido a la mejora con forma de onda

codificada) sean audibles (cuando es favorable) con respecto a artefactos de la mejora paramétrica codificada (p.ej., cuando los artefactos de codificación audibles (debido a la mejora con forma de onda codificada) son menos objetables que los artefactos audibles de la mejora paramétrica codificada).

El aporte de la mejora con forma de onda codificada en el esquema de codificación híbrido puede aumentarse mientras se asegura que el ruido de codificación no se convierte en audible de manera objetable (p.ej., no se convierte en audible) mediante el uso de un modelo de enmascaramiento auditivo para predecir, de manera más exacta, cómo el ruido de codificación en la copia de voz de calidad reducida (que se usará para implementar la mejora con forma de onda codificad) se enmascara por la mezcla de audio del programa principal y para seleccionar la relación de mezcla de manera acorde.

Algunas realizaciones que emplean un modelo de enmascaramiento auditivo incluyen las etapas de: segmentar la señal de audio no mejorada (mezcla de audio original) en segmentos de tiempo consecutivos, y proveer una copia de calidad reducida de la voz en cada segmento (para su uso en la mejora con forma de onda codificada) y parámetros de mejora paramétrica codificada (para su uso en la mejora paramétrica codificada) para cada segmento; para cada uno de los segmentos, usar el modelo de enmascaramiento auditivo para determinar una cantidad máxima de mejora con forma de onda codificada que puede aplicarse sin que los artefactos de codificación se conviertan en audibles de manera objetable; y generar un indicador (para cada segmento de la señal de audio no mejorada) de una combinación de mejora con forma de onda codificada (en una cantidad que no supera la cantidad máxima de la mejora con forma de onda codificada determinada mediante el uso del modelo de enmascaramiento auditivo para el segmento, y que al menos sustancialmente concuerda con la cantidad máxima de la mejora con forma de onda codificada determinada mediante el uso del modelo de enmascaramiento auditivo para el segmento) y mejora paramétrica codificada, de modo que la combinación de la mejora con forma de onda codificada y mejora paramétrica codificada genera una cantidad total predeterminada de mejora de la voz para el segmento.

En algunas realizaciones, cada indicador se incluye (p.ej., por un codificador) en un tren de bits que también incluye datos de audio codificados indicativos de la señal de audio no mejorada.

En algunas realizaciones, la señal de audio no mejorada se segmenta en segmentos de tiempo consecutivos y cada segmento de tiempo se segmenta en bandas de frecuencia, para cada una de las bandas de frecuencia de cada uno de los segmentos de tiempo, el modelo de enmascaramiento auditivo se usa para determinar una cantidad máxima de mejora con forma de onda codificada que puede aplicarse sin que los artefactos de codificación se conviertan en audibles de manera objetable, y un indicador se genera para cada banda de frecuencia de cada segmento de tiempo de la señal de audio no mejorada.

De manera opcional, el método también incluye una etapa de llevar a cabo (en cada segmento de la señal de audio no mejorada) en respuesta al indicador para cada segmento, la combinación de mejora con forma de onda codificada y mejora paramétrica codificada determinadas por el indicador, de modo que la combinación de la mejora con forma de onda codificada y mejora paramétrica codificada genera la cantidad total predeterminada de mejora de la voz para el segmento.

En algunas realizaciones, el contenido de audio se codifica en una señal de audio codificada para una configuración de canal de audio de referencia (o representación) como, por ejemplo, una configuración de sonido envolvente, una configuración de altavoz 5.1, una configuración de altavoz 7.1, una configuración de altavoz 7.2, etc. La configuración de referencia puede comprender canales de audio como, por ejemplo, canales estéreo, canales frontales izquierdo y derecho, canales envolventes, canales de altavoz, canales de objeto, etc. Uno o más de los canales que llevan contenido de voz pueden no ser canales de una representación de canal de audio Media/Lateral (M/S, por sus siglas en inglés). Según su uso en la presente memoria, una representación de canal de audio M/S (o, simplemente, representación M/S) comprende al menos un canal medio y un canal lateral. En una realización a modo de ejemplo, el canal medio representa una suma de canales izquierdo y derecho (p.ej., igualmente ponderados), mientras que el canal lateral representa una diferencia de canales izquierdo y derecho, en donde los canales izquierdo y derecho pueden considerarse cualquier combinación de dos canales, p.ej., canales frontal-central y frontal-izquierdo.

En algunas realizaciones, el contenido de voz de un programa puede mezclarse con contenido de no voz y puede distribuirse en dos o más canales no M/S como, por ejemplo, canales izquierdo y derecho, canales frontales izquierdo y derecho, etc., en la configuración de canal de audio de referencia. El contenido de voz puede, pero no se requiere que así sea, representarse en un centro imaginario en contenido estéreo en el cual el contenido de voz es igualmente alto en dos canales no M/S como, por ejemplo, canales izquierdo y derecho, etc. El contenido estéreo puede contener contenido de no voz que no es necesariamente igual de alto o que se encuentra incluso presente en los dos canales.

En algunos enfoques, múltiples conjuntos de datos de control no M/S, parámetros de control, etc., para la mejora de la voz correspondientes a múltiples canales de audio no M/S en los cuales el contenido de voz se distribuye se transmiten como parte de metadatos de audio totales de un codificador de audio a decodificadores de audio corriente abajo. Cada uno de los múltiples conjuntos de datos de control no M/S, parámetros de control, etc., para la

mejora de la voz corresponde a un canal de audio específico de los múltiples canales de audio no M/S en los cuales el contenido de voz se distribuye y puede usarse por un decodificador de audio corriente abajo para controlar funciones de mejora de la voz relacionadas con el canal de audio específico. Según su uso en la presente memoria, un conjunto de datos de control no M/S, parámetros de control, etc., se refiere a datos de control, parámetros de control, etc., para funciones de mejora de la voz en un canal de audio de una representación no M/S como, por ejemplo, la configuración de referencia en la cual una señal de audio según se describe en la presente memoria se codifica.

En algunas realizaciones, los metadatos de mejora de la voz M/S se transmiten -además de o en lugar de uno o más conjuntos de los datos de control no M/S, parámetros de control, etc., - como parte de metadatos de audio de un codificador de audio a decodificadores de audio corriente abajo. Los metadatos de mejora de la voz M/S pueden comprender uno o más conjuntos de datos de control M/S, parámetros de control, etc., para la mejora de la voz. Según su uso en la presente memoria, un conjunto de datos de control M/S, parámetros de control, etc., se refiere a datos de control, parámetros de control, etc., para funciones de mejora de la voz en un canal de audio de la representación M/S. En algunas realizaciones, los metadatos de mejora de la voz M/S para la mejora de la voz se transmiten por un codificador de audio a decodificadores de audio corriente abajo con el contenido mixto codificado en la configuración de canal de audio de referencia. En algunas realizaciones, el número de conjuntos de datos de control M/S, parámetros de control, etc., para la mejora de la voz en los metadatos de mejora de la voz M/S puede ser menor que el número de múltiples canales de audio no M/S en la representación de canal de audio de referencia en los cuales el contenido de voz en el contenido mixto se distribuye. En algunas realizaciones, incluso cuando el contenido de voz en el contenido mixto se distribuye en dos o más canales de audio no M/S como, por ejemplo, canales izquierdo y derecho, etc., en la configuración de canal de audio de referencia, solo un conjunto de datos de control M/S, parámetros de control, etc., para la mejora de la voz -p.ej., correspondiente al canal medio de la representación M/S- se envía como los metadatos de mejora de la voz M/S por un codificador de audio a decodificadores corriente abajo. El único conjunto de datos de control M/S, parámetros de control, etc., para la mejora de la voz puede usarse para lograr funciones de mejora de la voz para los dos o más canales de audio no M/S como, por ejemplo, los canales izquierdo y derecho, etc. En algunas realizaciones, matrices de transformación entre la configuración de referencia y la representación M/S pueden usarse para aplicar funciones de mejora de la voz según los datos de control M/S, parámetros de control, etc., para la mejora de la voz según se describe en la presente memoria.

Las técnicas según se describe en la presente memoria pueden usarse en escenarios en los cuales el contenido de voz queda abarcado en el centro imaginario de los canales izquierdo y derecho, el contenido de voz no queda abarcado completamente en el centro (p.ej., no es igualmente alto en los canales izquierdo y derecho, etc.), etc. En un ejemplo, dichas técnicas pueden usarse en escenarios en los cuales un gran porcentaje (p.ej., 70+%, 80+%, 90+%, etc.) de la energía del contenido de voz se encuentra en la señal media o canal medio de la representación M/S. En otro ejemplo, transformaciones (p.ej., espaciales, etc.) como, por ejemplo, panoramización, rotaciones, etc., pueden usarse para transformar contenido de voz no igualado en la configuración de referencia para que esa igual o sustancialmente igual en la configuración M/S. Los vectores de reproducción, matrices de transformación, etc., que representan el panoramización, rotaciones, etc., pueden usarse como parte de, o en conjunto con, funciones de mejora de la voz.

En algunas realizaciones (p.ej., un modo híbrido, etc.), una versión (p.ej., una versión reducida, etc.) del contenido de voz se envía a un decodificador de audio corriente abajo como solamente una señal de canal medio o como señales de canal medio y canal lateral en la representación M/S, junto con el contenido mixto enviado en la configuración de canal de audio de referencia posiblemente con una representación no M/S. En algunas realizaciones, cuando la versión del contenido de voz se envía a un decodificador de audio corriente abajo solamente como una señal de canal medio en la representación M/S, un vector de reproducción correspondiente que funciona (p.ej., lleva a cabo la transformación, etc.) en la señal de canal medio para generar porciones de señal en uno o más canales no M/S de una configuración de canal de audio no M/S (p.ej., la configuración de referencia, etc.) según la señal de canal medio se envía también al decodificador de audio corriente abajo.

En algunas realizaciones, un algoritmo de mejora de diálogo/voz (p.ej., en un decodificador de audio corriente abajo, etc.) que implementa la conmutación basada en SNR temporal "a ciegas" entre la mejora paramétrica codificada (p.ej., predicción de diálogo independiente de canal, predicción de diálogo multicanal, etc.) y mejora con forma de onda codificada de segmentos de un programa de audio funciona al menos en parte en la representación M/S.

Las técnicas según se describen en la presente memoria que implementan funciones de mejora de la voz al menos parcialmente en la representación M/S pueden usarse con predicción independiente de canal (p.ej., en el canal medio, etc.), predicción multicanal (p.ej., en el canal medio y canal lateral, etc.), etc. Dichas técnicas pueden usarse también para soportar la mejora de la voz para uno, dos o más diálogos al mismo tiempo. Cero, uno o más conjuntos adicionales de parámetros de control, datos de control, etc., como, por ejemplo, parámetros de predicción, ganancias, vectores de reproducción, etc., pueden proveerse en la señal de audio codificada como parte de los metadatos de mejora de la voz M/S para soportar diálogos adicionales.

En algunas realizaciones, la sintaxis de la señal de audio codificada (p.ej., emitida desde el codificador, etc.) soporta una transmisión de una bandera M/S de un codificador de audio corriente arriba a decodificadores de audio corriente abajo. La bandera M/S se encuentra presenta/se establece cuando funciones de mejora de la voz se llevan a cabo al menos en parte con datos de control M/S, parámetros de control, etc., que se transmiten con la bandera M/S. Por ejemplo, cuando la bandera M/S se establece, una señal estéreo (p.ej., de canales izquierdo y derecho, etc.) en canales no M/S puede transformarse primero por un decodificador de audio receptor en el canal medio y el canal lateral de la representación M/S antes de aplicar funciones de mejora de la voz M/S con los datos de control M/S, parámetros de control, etc., como recibidos con la bandera M/S, según uno o más de los algoritmos de mejora de la voz (p.ej., predicción de diálogo independiente de canal, predicción de diálogo multicanal, basados en forma de onda, híbridos con forma de onda paramétrica, etc.). Después de llevar a cabo las funciones de mejora de la voz M/S, las señales de voz mejorada en la representación M/S pueden transformarse otra vez en los canales no M/S.

En algunas realizaciones, el programa de audio cuyo contenido de voz se mejorará según las realizaciones de la invención incluye canales de altavoz, pero no canales de objeto. En otras realizaciones, el programa de audio cuyo contenido de voz se mejorará según realizaciones de la invención es un programa de audio basado en objeto (normalmente, un programa de audio basado en objeto multicanal) que comprende al menos un canal de objeto y opcionalmente también al menos un canal de altavoz.

Otro aspecto de la presente descripción es un sistema que incluye un codificador configurado (p.ej., programado) para llevar a cabo cualquier ejemplo del método de codificación descrito para generar un tren de bits que incluye datos de audio codificados, datos de forma de onda y datos paramétricos (y, de forma opcional, también un indicador de mezcla (p.ej., datos que indican la mezcla) para cada segmento de los datos de audio) en respuesta a datos de audio indicativos de un programa que incluye contenido de voz y no voz, y un decodificador configurado para analizar el tren de bits para recuperar los datos de audio codificados (y, de forma opcional, también cada indicador de mezcla) y para decodificar los datos de audio codificados para recuperar los datos de audio. De manera alternativa, el decodificador se configura para generar un indicador de mezcla para cada segmento de los datos de audio, en respuesta a los datos de audio recuperados. El decodificador se configura para llevar a cabo la mejora de la voz híbrida en los datos de audio recuperados en respuesta a cada indicador de mezcla.

Otro aspecto de la presente descripción es un decodificador configurado para llevar a cabo cualquier realización del método descrito. En otra clase de realizaciones, se provee un decodificador que incluye una memoria intermedia que almacena (p.ej., en una manera no transitoria) al menos un segmento (p.ej., trama) de un tren de bits de audio codificado que se ha generado por cualquier ejemplo del método.

Otros aspectos de la presente descripción incluyen un sistema o dispositivo (p.ej., un codificador, un decodificador o un procesador) configurado (p.ej., programado) para llevar a cabo cualquier realización del método descrito, y un medio legible por ordenador (p.ej., un disco) que almacena el código para implementar cualquier realización del método o etapas de aquel. Por ejemplo, el sistema puede ser o incluir un procesador de propósito general programable, procesador digital de señales, o microprocesador, programado con software o firmware y/o de otra forma configurado para llevar a cabo cualquiera de una variedad de funciones en datos, incluida una realización del método o etapas de aquel. Dicho procesador de propósito general puede ser o incluir un sistema informático que incluye un dispositivo de entrada, una memoria y circuitos de procesamiento programados (y/o de otra forma configurados) para llevar a cabo una realización del método (o sus etapas) en respuesta a datos aseverados a aquel.

En algunas realizaciones, los mecanismos según se describe en la presente memoria forman una parte de un sistema de procesamiento de medios, incluidos, pero sin limitación a ello: un dispositivo audiovisual, un TV de pantalla plana, un dispositivo portátil, una máquina de juegos, televisión, sistema de cine en casa, tableta, dispositivo móvil, ordenador portátil, *netbook*, radioteléfono celular, lector de libros electrónicos, terminal de punto de venta, ordenador de sobremesa, estación de trabajo de ordenador, estación de ordenadores, otros varios tipos de terminales y unidades de procesamiento de medios, etc.

Varias modificaciones a las realizaciones preferidas y principios y características genéricas descritas en la presente memoria serán inmediatamente aparentes para las personas con experiencia en la técnica. Por consiguiente, la descripción no pretende limitarse a las realizaciones que se muestran, sino que se le otorgará el más amplio alcance congruente con los principios y las características descritas en la presente memoria.

2. Notación y Nomenclatura

A lo largo de la presente descripción, incluso en las reivindicaciones, los términos "diálogo" y "voz" se usan de manera intercambiable como sinónimos para denotar contenido de señal de audio percibido como una forma de comunicación por un ser humano (o personaje en un mundo virtual).

A lo largo de la presente descripción, incluso en las reivindicaciones, la expresión llevar a cabo una función "en" una señal o datos (p.ej., filtrar, escalar, transformar o aplicar ganancia a, la señal o datos) se usa en un sentido amplio para denotar llevar a cabo la función directamente en la señal o datos, o en una versión procesada de la señal o

datos (p.ej., en una versión de la señal que ha experimentado el filtrado preliminar o preprocesamiento previo a llevar a cabo la función en aquellos).

A lo largo de la presente descripción, incluso en las reivindicaciones, la expresión "sistema" se usa en un sentido amplio para denotar un dispositivo, sistema o subsistema. Por ejemplo, puede hacerse referencia a un subsistema que implementa un decodificador como un sistema de decodificador, y también puede hacerse referencia a un sistema que incluye dicho subsistema (p.ej., un sistema que genera X señales de salida en respuesta a múltiples entradas, en el cual el subsistema genera M de las entradas y las otras X - M entradas se reciben de una fuente externa) como un sistema de decodificador.

A lo largo de la presente descripción, incluso en las reivindicaciones, el término "procesador" se usa en un sentido amplio para denotar un sistema o dispositivo programable o de otra forma configurable (p.ej., con software o firmware) para llevar a cabo funciones en los datos (p.ej., audio o vídeo u otros datos de imagen). Ejemplos de procesadores incluyen una matriz de puertas programable por campo (u otro circuito integrado o conjunto de chips configurables), un procesador digital de señales programado y/o de otra forma configurado para llevar a cabo el procesamiento canalizado en datos de audio u otros datos de sonido, un procesador u ordenador de propósito general programable y un chip de microprocesador o conjunto de chips programables.

A lo largo de la presente descripción, incluso en las reivindicaciones, las expresiones "procesador de audio" y "unidad de procesamiento de audio" se usan de manera intercambiable y en un sentido amplio para denotar un sistema configurado para procesar datos de audio. Ejemplos de unidades de procesamiento de audio incluyen, pero sin limitación, codificadores (p.ej., transcodificadores), decodificadores, códecs, sistemas de preprocesamiento, sistemas de postprocesamiento y sistemas de procesamiento de tren de bits (a los que algunas veces se hace referencia como herramientas de procesamiento de tren de bits).

A lo largo de la presente descripción, incluso en las reivindicaciones, la expresión "metadatos" se refiere a datos separados y diferentes de los datos de audio correspondientes (contenido de audio de un tren de bits que también incluye metadatos). Los metadatos se asocian a datos de audio e indican al menos una característica de los datos de audio (p.ej., qué tipo de procesamiento ya se ha llevado a cabo, o debe llevarse a cabo, en los datos de audio, o la trayectoria de un objeto indicada por los datos de audio). La asociación de los metadatos a los datos de audio es síncrona en el tiempo. Por consiguiente, los metadatos presentes (más recientemente recibidos o actualizados) pueden indicar que los datos de audio correspondientes tienen, de manera contemporánea, tienen una característica indicada y/o comprenden los resultados de un tipo indicado de procesamiento de datos de audio.

A lo largo de la presente descripción, incluso en las reivindicaciones, el término "se acopla(n)" o "acoplada/o/(s)" se usa para significar una conexión directa o indirecta. Por consiguiente, si un primer dispositivo se acopla a un segundo dispositivo, dicha conexión puede ser a través de una conexión directa o a través de una conexión indirecta mediante otros dispositivos y conexiones.

A lo largo de la presente descripción, incluso en las reivindicaciones, las siguientes expresiones tienen las siguientes definiciones:

- altavoz se usa para denotar cualquier transductor de emisión de sonido. La presente definición incluye altavoces implementados como múltiples transductores (p.ej., altavoz de graves y altavoz especializado en altas frecuencias);

- alimentación de altavoz: una señal de audio que se aplicará directamente a un altavoz, o una señal de audio que se aplicará a un amplificador y altavoz en serie;

- canal (o "canal de audio"): una señal de audio monofónica. Dicha señal puede normalmente reproducirse de manera tal que es equivalente a la aplicación de la señal directamente a un altavoz en una posición deseada o nominal. La posición deseada puede ser estática, como es normalmente el caso con altavoces físicos, o dinámica;

- programa de audio: un conjunto de uno o más canales de audio (al menos un canal de altavoz y/o al menos un canal de objeto) y opcionalmente también metadatos asociados (p.ej., metadatos que describen una presentación de audio espacial deseada);

- canal de altavoz (o "canal de alimentación de altavoz"): un canal de audio que se asocia a un altavoz nombrado (en una posición deseada o nominal), o con una zona de altavoz nombrada dentro de una configuración de altavoz definida. Un canal de altavoz se reproduce de manera tal que es equivalente a la aplicación de la señal de audio directamente al altavoz nombrado (en la posición deseada o nominal) o a un altavoz en la zona de altavoz nombrada;

- canal de objeto: un canal de audio indicativo de sonido emitido por una fuente de audio (a la que a veces se hace referencia como un "objeto" de audio). Normalmente, un canal de objeto determina una descripción de fuente de audio paramétrica (p.ej., metadatos indicativos de la descripción de fuente de audio paramétrica se incluyen en o se proveen con el canal de objeto). La descripción de fuente puede determinar el sonido emitido por el fuente (como

una función de tiempo), la posición aparente (p.ej., coordenadas espaciales 3D) de la fuente como una función de tiempo, y opcionalmente al menos un parámetro adicional (p.ej., tamaño o ancho de la fuente aparente) que caracteriza a la fuente;

- programa de audio basado en objeto: un programa de audio que comprende un conjunto de uno o más canales de objeto (y, de forma opcional, que también comprende al menos un canal de altavoz) y, de forma opcional, también metadatos asociados (p.ej., metadatos indicativos de una trayectoria de un objeto de audio que emite sonido indicado por un canal de objeto, o metadatos de otra forma indicativos de una presentación de audio espacial deseada de sonido indicado por un canal de objeto, o metadatos indicativos de una identificación de al menos un objeto de audio que es una fuente de sonido indicada por un canal de objeto); y

- reproducción: el proceso de conversión de un programa de audio en una o más alimentaciones de altavoz, o el proceso de conversión de un programa de audio en una o más alimentaciones de altavoz y conversión de la alimentación de altavoz en sonido mediante el uso de uno o más altavoces (en el último caso, a veces se hace referencia a la reproducción en la presente memoria como reproducción "por" el altavoz). Un canal de audio puede reproducirse de manera trivial ("en" una posición deseada) mediante la aplicación de la señal directamente a un altavoz físico en la posición deseada, o uno o más canales de audio pueden reproducirse mediante el uso de una de una variedad de técnicas de virtualización diseñadas para que sean sustancialmente equivalentes (para el oyente) a dicha reproducción trivial. En este último caso, cada canal de audio puede convertirse en una o más alimentaciones de altavoz que se aplicarán a los altavoces en ubicaciones conocidas, que son, en general, diferentes de la posición deseada, de modo que el sonido emitido por el altavoz en respuesta a la alimentación se percibirá como una emisión desde la posición deseada. Ejemplos de dichas técnicas de virtualización incluyen la reproducción binaural mediante auriculares (p.ej., mediante el uso del procesamiento de Auriculares Dolby que simulan hasta canales de sonido envolvente 7.1 para el usuario de los auriculares) y síntesis de campo de onda.

Las realizaciones de los métodos de codificación, decodificación y mejora de la voz y sistemas configurados para implementar los métodos se describirán con referencia a la Figura 3, Figura 6 y Figura 7.

3. Generación de parámetros de predicción

Con el fin de llevar a cabo la mejora de la voz (incluida la mejora de la voz híbrida según realizaciones de la invención), es necesario tener acceso a la señal de voz que se mejorará. Si la señal de voz no se encuentra disponible (de manera separada de una mezcla del contenido de voz y no voz de la señal mixta que se mejorará) en el momento en el que la mejora de la voz se lleva a cabo, pueden usarse técnicas paramétricas para crear una reconstrucción de la voz de la mezcla disponible.

Un método para la reconstrucción paramétrica del contenido de voz de una señal de contenido mixto (indicativa de una mezcla de contenido de voz y no voz) se basa en la reconstrucción de la potencia de la voz en cada losa tiempo-frecuencia de la señal y genera parámetros según:

$$p_{n,b} = \sqrt{\sum_{s \in n, f \in b} \frac{D_{s,f}^2}{M_{s,f}^2}} \qquad (2)$$

donde $p_{n,b}$ es el parámetro (valor de mejora de la voz paramétrica codificada) para la losa que tiene el índice temporal n y el índice de banda de frecuencia b, el valor $D_{s,f}$ representa la señal de voz en el intervalo de tiempo s y el comportamiento de frecuencias f de la losa, el valor $M_{s,f}$ representa la señal de contenido mixto en el mismo intervalo de tiempo y comportamiento de frecuencias de la losa, y la suma se encuentra por encima de todos los valores de s y f en todas las losas. Los parámetros $p_{n,b}$ pueden entregarse (como metadatos) con la propia señal de contenido mixto, para permitir a un receptor reconstruir el contenido de voz de cada segmento de la señal de contenido mixto.

Como se representa en la Figura 1, cada parámetro $p_{n,b}$ puede determinarse llevando a cabo una transformación de dominio temporal a dominio de la frecuencia en la señal de contenido mixto ("audio mixto") cuyo contenido de voz se mejorará, llevando a cabo una transformación de dominio temporal a dominio de la frecuencia en la señal de voz (el contenido de voz de la señal de contenido mixto), integrando la energía (de cada losa tiempo-frecuencia que tiene el índice temporal n y el índice de bandas de frecuencia b de la señal de voz) en todos los intervalos de tiempo y comportamientos de frecuencias en la losa, e integrando la energía de la losa tiempo-frecuencia correspondiente de la señal de contenido mixto en todos los intervalos de tiempo y comportamientos de frecuencias en la losa, y dividiendo el resultado de la primera integración por el resultado de la segunda integración para generar el parámetro $p_{n,b}$ para la losa.

Cuando cada losa tiempo-frecuencia de la señal de contenido mixto se multiplica por el parámetro $p_{n,b}$ para la losa, la señal resultante tiene envolventes espectrales y temporales similares que el contenido de voz de la señal de contenido mixto.

Los programas de audio típicos, p.ej., programas de audio de canal estéreo o 5.1, incluyen múltiples canales de altavoz. Normalmente, cada canal (o cada uno de un subconjunto de los canales) es indicativo de contenido de voz y no voz, y una señal de contenido mixto determina cada canal. El método de reconstrucción paramétrica de la voz descrito puede aplicarse de manera independiente a cada canal para reconstruir el componente de voz de todos los canales. Las señales de voz reconstruida (una para cada uno de los canales) pueden añadirse a las señales de canal de contenido mixto correspondientes, con una ganancia adecuada para cada canal, para lograr un impulso deseado del contenido de voz.

Las señales (canales) de contenido mixto de un programa multicanal pueden representarse como un conjunto de vectores de señal, donde cada elemento de vector es una colección de losas tiempo-frecuencia correspondientes a un conjunto específico de parámetros, a saber, todos los comportamientos de frecuencias (f) en la banda de parámetro (b) e intervalos de tiempo (s) en la trama (n). Un ejemplo de dicho conjunto de vectores, para una señal de contenido mixto de tres canales, es:

$$M_{n,b} = \begin{pmatrix} M_{c_1,n,b} \\ M_{c_2,n,b} \\ M_{c_3,n,b} \end{pmatrix} \tag{3}$$

donde c_i indica el canal. El ejemplo supone tres canales, pero el número de canales es una cantidad arbitraria.

De manera similar, el contenido de voz de un programa multicanal puede representarse como un conjunto de 1 × 1 matrices (donde el contenido de voz consiste en solamente un canal), $D_{n,b}$. La multiplicación de cada elemento de matriz de la señal de contenido mixto con un valor escalar resulta en una multiplicación de cada subelemento con el valor escalar. Un valor de voz reconstruida para cada losa se obtiene, por consiguiente, mediante el cálculo de

$$D_{r,n,b} = diag(P) \cdot M_{n,b} \tag{4}$$

para cada *n* y *b,* donde P es una matriz cuyos elementos son parámetros de predicción. La voz reconstruida (para todas las losas) puede también denotarse como:

$$D_r = diag(P) \cdot M \tag{5}$$

El contenido en los múltiples canales de una señal de contenido mixto multicanal hace que las correlaciones entre los canales que pueden emplearse realicen una mejor predicción de la señal de voz. Mediante el empleo de un predictor de Error Cuadrático Medio Mínimo (MMSE, por sus siglas en inglés) (p.ej., de un tipo convencional), los canales pueden combinarse con parámetros de predicción para reconstruir el contenido de voz con un error mínimo según el criterio de Error Cuadrático Medio (MSE, por sus siglas en inglés). Como se muestra en la Figura 2, suponiendo que una señal de entrada de contenido mixto de tres canales como, por ejemplo, un predictor MMSE (que funciona en el dominio de la frecuencia) genera, de manera iterativa, un conjunto de parámetros de predicción p_i (donde el índice *i* es 1, 2 o 3) en respuesta a la señal de entrada de contenido mixto y una sola señal de voz de entrada indicativa del contenido de voz de la señal de entrada de contenido mixto.

Un valor de voz reconstruido a partir de una losa de cada canal de la señal de entrada de contenido mixto (cada losa teniendo los mismos índices *n* y *b*) es una combinación lineal del contenido ($M_{ci,n,b}$) de cada canal (*i* = 1, 2 o 3) de la señal de contenido mixto controlada por un parámetro de ponderación para cada canal. Dichos parámetros de ponderación son los parámetros de predicción, p_i, para las losas que tienen los mismos índices *n* y *b.* Por consiguiente, la voz reconstruida a partir de todas las losas de todos los canales de la señal de contenido mixto es:

$$D_r = p_1 \cdot M_{c1} + p_2 \cdot M_{c2} + p_3 \cdot M_{c3} \tag{6}$$

o en forma de matriz de señal:

$$D_r = PM \qquad (7)$$

Por ejemplo, cuando la voz está presente, de manera coherente, en múltiples canales de la señal de contenido mixto mientras que los sonidos de fondo (no voz) son incoherentes entre los canales, una combinación de suma de los canales favorecerá la energía de la voz. Para dos canales, ello resulta en una mejor separación de la voz 3 dB en comparación con la reconstrucción independiente del canal. A modo de otro ejemplo, cuando la voz está presente en un canal y los sonidos de fondo están presentes, de manera coherente, en múltiples canales, una combinación de resta de canales eliminará (parcialmente) los sonidos de fondo mientras se preserva la voz.

En una clase de realizaciones, el método incluye las etapas de: (a) recibir un tren de bits indicativo de un programa de audio que incluye voz que tiene una forma de onda no mejorada y otro contenido de audio, en donde el tren de bits incluye: datos de audio no mejorados indicativos de la voz y del otro contenido de audio, datos de forma de onda indicativos de una versión de calidad reducida de la voz, la versión de calidad reducida de la voz tiene una segunda forma de onda similar (p.ej., al menos sustancialmente similar) a la forma de onda no mejorada, y la versión de calidad reducida tendrá una calidad objetable si se escucha de manera aislada, y datos paramétricos, en donde los datos paramétricos con los datos de audio no mejorados determinan la voz paramétricamente construida, y la voz paramétricamente construida es una versión paramétricamente reconstruida de la voz que concuerda al menos parcialmente con (p.ej., es una buena aproximación de) la voz; y (b) llevar a cabo la mejora de la voz en el tren de bits en respuesta a un indicador de mezcla y, de esta manera, generar datos indicativos de un programa de audio de voz mejorada, incluso mediante la combinación de los datos de audio no mejorados con una combinación de datos de voz de baja calidad determinados a partir de los datos de forma de onda, y datos de voz reconstruida, en donde la combinación se determina por el indicador de mezcla (p.ej., la combinación tiene una secuencia de estados determinados por una secuencia de valores actuales del indicador de mezcla), los datos de voz reconstruida se generan en respuesta a al menos algunos de los datos paramétricos y al menos algunos de los datos de audio no mejorados, y el programa de audio de voz mejorada tiene artefactos de codificación de mejora de la voz menos audibles (p.ej., artefactos de codificación de mejora de la voz que se enmascaran mejor) que un programa de audio de voz mejorada con forma de onda puramente codificada determinado mediante la combinación de solamente los datos de voz de baja calidad con los datos de audio no mejorados o un programa de audio de voz mejorada paramétrica codificada puramente determinado a partir de los datos paramétricos y datos de audio no mejorados.

En algunas realizaciones, el indicador de mezcla (que puede tener una secuencia de valores, p.ej., uno para cada uno de una secuencia de segmentos de tren de bits) se incluye en el tren de bits recibido en la etapa (a). En otras realizaciones, el indicador de mezcla se genera (p.ej., en un receptor que recibe y decodifica el tren de bits) en respuesta al tren de bits.

Debe comprenderse que la expresión "indicador de mezcla" no pretende denotar un solo parámetro o valor (o una secuencia de parámetros o valores únicos) para cada segmento del tren de bits. Más bien, se contempla que, en algunas realizaciones, un indicador de mezcla (para un segmento del tren de bits) puede ser un conjunto de dos o más parámetros o valores (p.ej., para cada segmento, un parámetro de control de mejora paramétrica codificada y un parámetro de control de mejora con forma de onda codificada). En algunas realizaciones, el indicador de mezcla para cada segmento puede ser una secuencia de valores que indican la mezcla por banda de frecuencia del segmento.

Los datos de forma de onda y datos paramétricos no necesitan proveerse para (p.ej., incluirse en) cada segmento del tren de bits, o usarse para llevar a cabo la mejora de la voz en cada segmento del tren de bits. Por ejemplo, en algunos casos, al menos un segmento puede incluir datos de forma de onda solamente (y la combinación determinada por el indicador de mezcla para cada segmento puede consistir en datos de forma de onda solamente) y al menos otro segmento puede incluir datos paramétricos solamente (y la combinación determinada por el indicador de mezcla para cada segmento puede consistir en datos de voz reconstruida solamente).

Se contempla que, en algunas realizaciones, un codificador genera el tren de bits que incluye la codificación (p.ej., compresión) de los datos de audio no mejorados, pero no los datos de forma de onda o datos paramétricos. Por consiguiente, cuando el tren de bits se entrega a un receptor, el receptor analizará el tren de bits para extraer los datos de audio no mejorados, los datos de forma de onda y los datos paramétricos (y el indicador de mezcla si se entrega en el tren de bits), pero decodificará solamente los datos de audio no mejorados. El receptor llevará a cabo la mejora de la voz en los datos de audio no mejorados decodificados (mediante el uso de los datos de forma de onda y/o datos paramétricos) sin aplicar a los datos de forma de onda o datos paramétricos el mismo proceso de decodificación que se aplica a los datos de audio.

Normalmente, la combinación (indicada por el indicador de mezcla) de los datos de forma de onda y datos de voz reconstruida cambia con el tiempo, con cada estado de la combinación perteneciendo al contenido de voz y otro contenido de audio de un segmento correspondiente del tren de bits. El indicador de mezcla se genera de modo que el estado actual de la combinación (de datos de forma de onda y datos de voz reconstruida) se determina por

propiedades de señal de la voz y otro contenido de audio (p.ej., una relación de la potencia del contenido de voz y la potencia de otro contenido de audio) en el segmento correspondiente del tren de bits.

La etapa (b) puede incluir una etapa de llevar a cabo la mejora de la voz con forma de onda codificada mediante la combinación (p.ej., mezcla) de al menos algunos de los datos de voz de baja calidad con los datos de audio no mejorados de al menos un segmento del tren de bits, y llevar a cabo la mejora de la voz paramétrica codificada mediante la combinación de datos de voz reconstruida con los datos de audio no mejorados de al menos un segmento del tren de bits. Una combinación de la mejora de la voz con forma de onda codificada y mejora de la voz paramétrica codificada se lleva a cabo en al menos un segmento del tren de bits mediante la mezcla tanto de los datos de voz de baja calidad como de los datos de voz reconstruida para el segmento con los datos de audio no mejorados del segmento. En algunas condiciones de señal, solo una (pero no ambas) de la mejora de la voz con forma de onda codificada y mejora de la voz paramétrica codificada se lleva a cabo (en respuesta al indicador de mezcla) en un segmento (o en cada uno de más de un segmento) del tren de bits.

4. Funciones de mejora de la voz

En la presente memoria, la expresión "SNR" (relación señal-ruido) se usa para denotar la relación (o nivel) de potencia del componente de voz (a saber, contenido de voz) de un segmento de un programa de audio (o de todo el programa) con respecto a la del componente de no voz (a saber, el contenido de no voz) del segmento o programa, o a la de todo el contenido (de voz y no voz) del segmento o programa. En algunas realizaciones, SNR se deriva de una señal de audio (para experimentar la mejora de la voz) y una señal separada indicativa del contenido de voz de la señal de audio (p.ej., una copia de baja calidad del contenido de voz que se ha generado para su uso en la mejora con forma de onda codificada). En algunas realizaciones, SNR se deriva de una señal de audio (para experimentar la mejora de la voz) y de datos paramétricos (que se han generado para su uso en la mejora paramétrica codificada de la señal de audio).

En una clase de realizaciones, el método implementa la conmutación basada en SNR temporal "a ciegas" entre la mejora paramétrica codificada y la mejora con forma de onda codificada de segmentos de un programa de audio. En el presente contexto, "a ciegas" denota que la conmutación no se guía perceptualmente por un modelo de enmascaramiento auditivo complejo (p.ej., de un tipo que se describirá en la presente memoria), sino que se guía por una secuencia de valores SRN (indicadores de mezcla) correspondientes a segmentos del programa. En una realización en la presente clase, la mejora de la voz híbrida codificada se logra por la conmutación temporal entre mejora paramétrica codificada y mejora con forma de onda codificada (en respuesta a un indicador de mezcla, p.ej., un indicador de mezcla generado en el subsistema 29 del codificador de la Figura 3, que indica que la mejora paramétrica codificada solamente o la mejora con forma de onda codificada deben llevarse a cabo en datos de audio correspondientes), de modo que la mejora paramétrica codificada o la mejora con forma de onda codificada (pero no la mejora paramétrica codificada y la mejora con forma de onda codificada) se lleva a cabo en cada segmento de un programa de audio en el cual la mejora de la voz se lleva a cabo. Mediante el reconocimiento de que la mejora con forma de onda codificada tiene un mejor rendimiento en la condición de SNR baja (en segmentos que tienen valores bajos de SNR) y la mejora paramétrica codificada tiene un mejor rendimiento en SNR favorables (en segmentos que tienen valores altos de SNR), la decisión de conmutación se basa normalmente en la relación de voz (diálogo) con respecto al audio restante en una mezcla de audio original.

Las realizaciones que implementan la conmutación basada en SNR temporal "a ciegas" normalmente incluyen las etapas de: segmentar la señal de audio no mejorada (mezcla de audio original) en segmentos de tiempo consecutivos, y determinar para cada segmento la SNR entre el contenido de voz y el otro contenido de audio (o entre el contenido de voz y el contenido de audio total) del segmento; y para cada segmento, comparar la SNR con un umbral y proveer un parámetro de control de mejora paramétrica codificada para el segmento (a saber, el indicador de mezcla para el segmento indica que la mejora paramétrica codificada debe llevarse a cabo) cuando la SNR es mayor que el umbral o proveer un parámetro de control de mejora con forma de onda codificada para el segmento (a saber, el indicador de mezcla para el segmento indica que la mejora con forma de onda codificada debe llevarse a cabo) cuando la SNR no es mayor que el umbral.

Cuando la señal de audio no mejorada se entrega (p.ej., se transmite) con los parámetros de control incluidos como metadatos a un receptor, el receptor puede llevar a cabo (en cada segmento) el tipo de mejora de la voz indicado por el parámetro de control para el segmento. Por consiguiente, el receptor lleva a cabo la mejora paramétrica codificada en cada segmento para el cual el parámetro de control es un parámetro de control de mejora paramétrica codificada, y la mejora con forma de onda codificada en cada segmento para el cual el parámetro de control es un parámetro de control de mejora con forma de onda codificada.

Si una persona desea incurrir en el coste de transmisión (con cada segmento de una mezcla de audio original) tanto los datos de forma de onda (para implementar la mejora de la voz con forma de onda codificada) como los parámetros de mejora paramétrica codificada con una mezcla original (no mejorada), un grado más alto de mejora de la voz puede lograrse mediante la aplicación tanto de la mejora con forma de onda codificada como de la mejora paramétrica codificada a segmentos individuales de la mezcla. Por consiguiente, en una clase de realizaciones, el

método implementa la mezcla basada en SNR temporal "a ciegas" entre la mejora paramétrica codificada y la mejora con forma de onda codificada de segmentos de un programa de audio. En el presente contexto, "a ciegas" denota que la conmutación no se guía perceptualmente por un modelo de enmascaramiento auditivo complejo (p.ej., de un tipo que se describirá en la presente memoria), sino que se guía por una secuencia de valores SRN correspondientes a segmentos del programa.

Las realizaciones que implementan la mezcla basada en SNR temporal "a ciegas" normalmente incluyen las etapas de: segmentar la señal de audio no mejorada (mezcla de audio original) en segmentos de tiempo consecutivos, y determinar para cada segmento la SNR entre el contenido de voz y el otro contenido de audio (o entre el contenido de voz y el contenido de audio total) del segmento; determinar (p.ej., recibir una solicitud de) una cantidad total ("T") de mejora de la voz; y para cada segmento, proveer un parámetro de control de mezcla, donde el valor del parámetro de control de mezcla se determina por (es una función de) la SNR para el segmento.

Por ejemplo, el indicador de mezcla para un segmento de un programa de audio puede ser un parámetro de indicador de mezcla (o conjunto de parámetros) generado en el subsistema 29 del codificador de la Figura 3 para el segmento.

El indicador de control de mezcla es un parámetro, α, para cada segmento de modo que T = α Pw + (1-α)Pp, donde Pw es la mejora con forma de onda codificada para el segmento que producirá la cantidad total predeterminada de mejora, T, si se aplica al contenido de audio no mejorado del segmento mediante el uso de datos de forma de onda provistos para el segmento (donde el contenido de voz del segmento tiene una forma de onda no mejorada, los datos de forma de onda para el segmento son indicativos de una versión de calidad reducida del contenido de voz del segmento, la versión de calidad reducida tiene una forma de onda similar (p.ej., al menos sustancialmente similar) a la forma de onda no mejorada, y la versión de calidad reducida del contenido de voz es de una calidad objetable cuando se reproduce y percibe de manera aislada), y Pp es la mejora paramétrica codificada que producirá la cantidad total predeterminada de mejora, T, si se aplica a contenido de audio no mejorado del segmento mediante el uso de datos paramétricos provistos para el segmento (donde los datos paramétricos para el segmento, con el contenido de audio no mejorado del segmento, determinan una versión paramétricamente reconstruida del contenido de voz del segmento).

Cuando la señal de audio no mejorada se entrega (p.ej., se transmite) con los parámetros de control como metadatos a un receptor, el receptor puede llevar a cabo (en cada segmento) la mejora de la voz híbrida indicada por los parámetros de control para el segmento. De manera alternativa, el receptor genera los parámetros de control a partir de la señal de audio no mejorada.

En algunas realizaciones, el receptor lleva a cabo (en cada segmento de la señal de audio no mejorada) una combinación de mejora paramétrica codificada Pp (escalada por el parámetro α para el segmento) y mejora con forma de onda codificada Pw (escalada por el valor (1 - α) para el segmento), de modo que la combinación de mejora paramétrica codificada escalada y mejora con forma de onda codificada escalada genera la cantidad total predeterminada de mejora, como en la expresión (1) (T = α Pw + (1-α)Pp).

Un ejemplo de la relación entre α y SNR para un segmento es la siguiente: α es una función no decreciente de SNR, el rango de α es 0 a 1, α tiene el valor 0 cuando la SNR para el segmento es menor que o igual a un valor umbral ("SNR_pobre"), y α tiene el valor 1 cuando la SNR es mayor que o igual a un valor umbral más grande ("SNR alta"). Cuando la SNR es favorable, α es alta, lo cual resulta en una gran proporción de mejora paramétrica codificada. Cuando la SNR es pobre, α es baja, lo cual resulta en una gran proporción de mejora con forma de onda codificada. La ubicación de los puntos de saturación (SNR_pobre y SNR alta) debe seleccionarse para admitir las implementaciones específicas de los algoritmos de mejora con forma de onda codificada y paramétrica codificada.

En otra clase de realizaciones, la combinación de mejora con forma de onda codificada y paramétrica codificada que se llevará a cabo en cada segmento de una señal de audio se determina por un modelo de enmascaramiento auditivo. En algunas realizaciones en la presente clase, la relación de mezcla óptima para una mezcla de mejora con forma de onda codificada y mejora paramétrica codificada que se llevará a cabo en un segmento de un programa de audio usa la cantidad más alta de mejora con forma de onda codificada que simplemente evita que el ruido de codificación se convierta en audible.

En las realizaciones de mezcla basadas en SNR a ciegas descritas más arriba, la relación de mezcla para un segmento se deriva de la SNR, y se supone que la SNR es indicativa de la capacidad de la mezcla de audio para enmascarar el ruido de codificación en la versión de calidad reducida (copia) de voz que se empleará para la mejora con forma de onda codificada. Las ventajas del enfoque basado en SNR a ciegas son la simplicidad de la implementación y la baja carga computacional en el codificador. Sin embargo, SNR es un predictor no fiable de cuán bien el ruido de codificación se enmascarará y un gran margen de seguridad debe aplicarse para asegurar que el ruido de codificación permanecerá enmascarado en todo momento. Ello significa que al menos parte del tiempo el nivel de la copia de voz de calidad reducida que se mezcla es menor de lo que podría ser, o, si el margen se establece de manera más agresiva, el ruido de codificación se convierte en audible parte del tiempo. El aporte de la mejora con forma de onda codificada en el esquema de codificación híbrido puede aumentarse mientras se asegura

que el ruido de codificación no se convierte en audible mediante el uso de un modelo de enmascaramiento auditivo para predecir de manera más exacta cómo el ruido de codificación en la copia de voz de calidad reducida se enmascara por la mezcla de audio del programa principal y para seleccionar la relación de mezcla de manera acorde.

Las realizaciones típicas que emplean un modelo de enmascaramiento auditivo incluyen las etapas de: segmentar la señal de audio no mejorada (mezcla de audio original) en segmentos de tiempo consecutivos, y proveer una copia de calidad reducida de la voz en cada segmento (para su uso en la mejora con forma de onda codificada) y parámetros de mejora paramétrica codificada (para su uso en la mejora paramétrica codificada) para cada segmento; para cada uno de los segmentos, usar el modelo de enmascaramiento auditivo para determinar una cantidad máxima de mejora con forma de onda codificada que puede aplicarse sin que los artefactos se conviertan en audibles; y generar un indicador de mezcla (para cada segmento de la señal de audio no mejorada) de una combinación de mejora con forma de onda codificada (en una cantidad que no supera la cantidad máxima de mejora con forma de onda codificada determinada mediante el uso del modelo de enmascaramiento auditivo para el segmento, y que preferiblemente concuerda al menos de forma sustancial con la cantidad máxima de mejora con forma de onda codificada determinada mediante el uso del modelo de enmascaramiento auditivo para el segmento) y mejora paramétrica codificada, de modo que la combinación de la mejora con forma de onda codificada y mejora paramétrica codificada genera una cantidad total predeterminada de mejora de la voz para el segmento.

En algunas realizaciones, cada indicador de mezcla se incluye (p.ej., por un codificador) en un tren de bits que también incluye datos de audio codificados indicativos de la señal de audio no mejorada. Por ejemplo, el subsistema 29 del codificador 20 de la Figura 3 puede configurarse para generar dichos indicadores de mezcla, y el subsistema 28 del codificador 20 puede configurarse para incluir los indicadores de mezcla en el tren de bits que se producirá desde el codificador 20. Para otro ejemplo, los indicadores de mezcla pueden generarse (p. ej., en el subsistema 13 del codificador de la Figura 7) a partir de los parámetros $g_{máx}(t)$ generados por el subsistema 14 del codificador de la Figura 7, y el subsistema 13 del codificador de la Figura 7 puede configurarse para incluir los indicadores de mezcla en el tren de bits que se producirá desde el codificador de la Figura 7 (o el subsistema 13 puede incluir, en el tren de bits que se producirá desde el codificador de la Figura 7, los parámetros $g_{máx}(t)$ generados por el subsistema 14, y un receptor que recibe y analiza el tren de bits puede configurarse para generar los indicadores de mezcla en respuesta a los parámetros $g_{máx}(t)$).

De manera opcional, el método también incluye una etapa de llevar a cabo (en cada segmento de la señal de audio no mejorada) en respuesta al indicador de mezcla para cada segmento, la combinación de mejora con forma de onda codificada y mejora paramétrica codificada determinada por el indicador de mezcla, de modo que la combinación de mejora con forma de onda codificada y mejora paramétrica codificada genera la cantidad total predeterminada de mejora de la voz para el segmento.

Un ejemplo del método que emplea un modelo de enmascaramiento auditivo se describirá con referencia a la Figura 7. En el presente ejemplo, una mezcla de voz y audio de fondo, A(t) (la mezcla de audio no mejorada) se determina (en el elemento 10 de la Figura 7) y pasa al modelo de enmascaramiento auditivo (implementado por el elemento 11 de la Figura 7) que predice un umbral de enmascaramiento $\Theta(f,t)$ para cada segmento de la mezcla de audio no mejorada. La mezcla de audio no mejorada A(t) también se provee al elemento 13 de codificación para la codificación para la transmisión.

El umbral de enmascaramiento generado por el modelo indica como una función de frecuencia y tiempo la excitación auditiva que cualquier señal debe superar para ser audible. Dichos modelos de enmascaramiento son conocidos en la técnica. El componente de voz, s(t), de cada segmento de la mezcla de audio no mejorada, A(t), se codifica (en el codificador de audio de velocidad binaria baja 15) para generar una copia de calidad reducida, s'(t), del contenido de voz del segmento. La copia de calidad reducida, s'(t) (que comprende menos bits que la voz original, s(t)), puede conceptualizarse como la suma de la voz original, s(t), y el ruido de codificación, n(t). Dicho ruido de codificación puede separarse de la copia de calidad reducida para el análisis a través de la resta (en el elemento 16) de la señal de voz alineada en el tiempo, s(t), de la copia de calidad reducida. De manera alternativa, el ruido de codificación puede estar disponible directamente del codificador de audio.

El ruido de codificación, n, se multiplica en el elemento 17 por un factor de escala, g(t), y el ruido de codificación escalado pasa a un modelo auditivo (implementado por el elemento 18) que predice la excitación auditiva, N(f,t), generada por el ruido de codificación escalado. Dichos modelos de excitación son conocidos en la técnica. En una etapa final, la excitación auditiva N(f,t) se compara con el umbral de enmascaramiento pronosticado $\Theta(f,t)$ y el factor de escala más grande, $g_{máx}(t)$, que asegura que el ruido de codificación se enmascara, a saber, el valor más grande de g(t) que asegura que $N(f,t) < \Theta(f,t)$, se encuentra (en el elemento 14). Si el modelo auditivo es no lineal, esto puede necesitar llevarse a cabo de manera iterativa (según se indica en la Figura 2) mediante la iteración del valor de g(t) aplicado al ruido de codificación, n(t) en el elemento 17; si el modelo auditivo es lineal, esto puede llevarse a cabo en una etapa hacia adelante de alimentación simple. El factor de escala resultante $g_{máx}(t)$ es el factor de escala más grande que puede aplicarse a la copia de voz de calidad reducida, s'(t), antes de añadirse al segmento correspondiente de la mezcla de audio no mejorada, A(t), sin que los artefactos de codificación en la copia de voz de

calidad reducida escalada se conviertan en audibles en la mezcla de la copia de voz de calidad reducida escalada, $g_{máx}(t)$* s'(t), y la mezcla de audio no mejorada, A(t).

El sistema de la Figura 7 también incluye el elemento 12, que se configura para generar (en respuesta a la mezcla de audio no mejorada, A(t) y la voz, s(t)) parámetros de mejora paramétrica codificada, p(t), para llevar a cabo la mejora de la voz paramétrica codificada en cada segmento de la mezcla de audio no mejorada.

Los parámetros de mejora paramétrica codificada, p(t), así como la copia de voz de calidad reducida, s'(t), generada en el codificador 15, y el factor, $g_{máx}(t)$, generado en el elemento 14, para cada segmento del programa de audio, también se aseveran al elemento 13 de codificación. El elemento 13 genera un tren de bits de audio codificado indicativo de la mezcla de audio no mejorado, A(t), parámetros de mejora paramétrica codificada, p(t), copia de voz de calidad reducida, s'(t), y el factor, $g_{máx}(t)$, para cada segmento del programa de audio, y dicho tren de bits de audio codificado puede transmitirse o de otra manera entregarse a un receptor.

En el ejemplo, la mejora de voz se lleva a cabo (p.ej., en un receptor al cual la salida codificada del elemento 13 se ha entregado) de la siguiente manera en cada segmento de la mezcla de audio no mejorada, A(t), para aplicar una cantidad total predeterminada (p.ej., solicitada) de mejora, T, mediante el uso del factor de escala $g_{máx}(t)$ para el segmento. El programa de audio codificado se decodifica para extraer la mezcla de audio no mejorada, A(t), los parámetros de mejora paramétrica codificada, p(t), la copia de voz de calidad reducida, s'(t), y el factor $g_{máx}(t)$ para cada segmento del programa de audio. Para cada segmento, se determina que la mejora con forma de onda codificada, Pw, es la mejora con forma de onda codificada que producirá la cantidad total predeterminada de mejora, T, si se aplica a contenido de audio no mejorado del segmento mediante el uso de la copia de voz de calidad reducida, s'(t), para el segmento, y se determina que la mejora paramétrica codificada, Pp, es la mejora paramétrica codificada que producirá la cantidad total predeterminada de mejora, T, si se aplica a contenido de audio no mejorado del segmento mediante el uso de datos paramétricos provistos para el segmento (donde los datos paramétricos para el segmento, con el contenido de audio no mejorado del segmento, determinan una versión paramétricamente reconstruida del contenido de voz del segmento). Para cada segmento, una combinación de mejora paramétrica codificada (en una cantidad escalada por un parámetro α_2 para el segmento) y mejora con forma de onda codificada (en una cantidad determinada por el valor α_1 para el segmento) se lleva a cabo, de modo que la combinación de mejora paramétrica codificada y mejora con forma de onda codificada genera la cantidad total predeterminada de mejora mediante el uso de la cantidad más grande de mejora con forma de onda codificada permitida por el modelo: $T = (\alpha_1(Pw) + \alpha_2(Pp))$, donde el factor α_1 es el valor máximo que no supera $g_{máx}(t)$ para el segmento y permite alcanzar la igualdad indicada ($T = (\alpha_1(Pw) + \alpha_2(Pp))$, y el parámetro α_2 es el valor mínimo no negativo que permite alcanzar la igualdad indicada ($T = (\alpha_1(Pw) + \alpha_2(Pp))$.

En un ejemplo alternativo, los artefactos de la mejora paramétrica codificada se incluyen en la evaluación (llevada a cabo por el modelo de enmascaramiento auditivo) para permitir que los artefactos de codificación (debido a la mejora con forma de onda codificada) se conviertan en audibles cuando esta es favorable por sobre los artefactos de la mejora paramétrica codificada.

En variaciones del ejemplo de la Figura 7 (y ejemplos similares al de la Figura 7 que emplean un modelo de enmascaramiento auditivo), al que a veces se hace referencia como realizaciones de división multibanda guiada del modelo auditivo, la relación entre el ruido de codificación de mejora con forma de onda codificada, N(f,t), en la copia de voz de calidad reducida y el umbral de enmascaramiento $\Theta(f,t)$ puede no ser uniforme a lo largo de todas las bandas de frecuencia. Por ejemplo, las características espectrales del ruido de codificación de mejora con forma de onda codificada pueden ser tales que en una primera región de frecuencia el ruido de enmascaramiento está cerca de superar el umbral de enmascaramiento mientras que en una segunda región de frecuencia el ruido de enmascaramiento se encuentra muy por debajo del umbral de enmascaramiento. En el ejemplo de la Figura 7, el aporte máximo de la mejora con forma de onda codificada se determinará por el ruido de codificación en la primera región de frecuencia y el factor de escalamiento máximo, *g*, que puede aplicarse a la copia de voz de calidad reducida se determina por el ruido de codificación y propiedades de enmascaramiento en la primera región de frecuencia. Este es más pequeño que el factor de escalamiento máximo, *g*, que podría aplicarse si la determinación del factor de escalamiento máximo se basara solamente en la segunda región de frecuencia. El rendimiento total puede mejorarse si los principios de mezcla temporal se aplican de forma separada en las dos regiones de frecuencia.

En una implementación de la división multibanda guida del modelo auditivo, la señal de audio no mejorada se divide en M bandas de frecuencia no superpuestas contiguas y los principios de mezcla temporal (a saber, mejora de la voz híbrida con una mezcla de mejora con forma de onda codificada y paramétrica codificada, según una realización de la invención) se aplican de forma independiente en cada una de las M bandas. Una implementación alternativa divide el espectro en una banda baja por debajo de una frecuencia de corte, fc, y una banda alta por encima de la frecuencia de corte, fc. La banda baja siempre se mejora con la mejora con forma de onda codificada y la banda superior siempre se mejora con la mejora paramétrica codificada. La frecuencia de corte varía con el tiempo y siempre se selecciona para que sea tan alta como sea posible bajo la limitación de que el ruido de codificación de

mejora con forma de onda codificada en una cantidad total predeterminada de mejora de la voz, T, se encuentra por debajo del umbral de enmascaramiento. En otras palabras, la frecuencia de corte máxima en cualquier momento es:

$$\max(fc \mid T*N(f<fc,t) < \Theta(f,t)) \quad (8)$$

Las realizaciones descritas más arriba han supuesto que el medio disponible para evitar que los artefactos de codificación de mejora con forma de onda codificada se conviertan en audibles es ajustar la relación de mezcla (de la mejora con forma de onda codificada frente a la mejora paramétrica codificada) o escalar otra vez la cantidad total de mejora. Una alternativa es controlar la cantidad de ruido de codificación de mejora con forma de onda codificada a través de una asignación variable de la velocidad binaria para generar la copia de voz de calidad reducida. En un ejemplo de la presente realización alternativa, se aplica una cantidad básica constante de mejora paramétrica codificada y la mejora con forma de onda codificada adicional se aplica para alcanzar la cantidad deseada (predeterminada) de mejora total. La copia de voz de calidad reducida se codifica con una velocidad binaria variable y dicha velocidad binaria se selecciona como la velocidad binaria más baja que mantiene el ruido de codificación de mejora con forma de onda codificada por debajo del umbral enmascarado del audio principal mejorado paramétrico codificado.

En algunas realizaciones, el programa de audio cuyo contenido de voz se mejorará según las realizaciones de la invención incluye canales de altavoz, pero no canales de objeto. En otras realizaciones, el programa de audio cuyo contenido de voz se mejorará según las realizaciones de la invención es un programa de audio basado en objeto (normalmente, un programa de audio basado en objeto multicanal) que comprende al menos un canal de objeto y opcionalmente también al menos un canal de altavoz.

Otros aspectos de la presente descripción incluyen un codificador configurado para llevar a cabo cualquier ejemplo del método de codificación descrito para generar una señal de audio codificada en respuesta a una señal de entrada de audio (p.ej., en respuesta a datos de audio indicativos de una señal de entrada de audio multicanal), un decodificador configurado para decodificar dicha señal codificada y llevar a cabo la mejora de voz en el contenido de audio decodificado, y un sistema que incluye dicho codificador y dicho decodificador. El sistema de la Figura 3 es un ejemplo de dicho sistema.

El sistema de la Figura 3 incluye el codificador 20, que se configura (p.ej., se programa) para llevar a cabo un ejemplo del método de codificación para generar una señal de audio codificada en respuesta a datos de audio indicativos de un programa de audio. Normalmente, el programa es un programa de audio multicanal. En algunas realizaciones, el programa de audio multicanal comprende solamente canales de altavoz. En otras realizaciones, el programa de audio multicanal es un programa de audio basado en objeto que comprende al menos un canal de objeto y, opcionalmente, también al menos un canal de altavoz.

Los datos de audio incluyen datos (identificados como datos de "audio mixto" en la Figura 3) indicativos de contenido de audio mixto (una mezcla de contenido de voz y no voz) y datos (identificados como datos "de voz" en la Figura 3) indicativos del contenido de voz del contenido de audio mixto.

Los datos de voz experimentan una transformación de dominio temporal a dominio de la frecuencia (QMF, por sus siglas en inglés) en la etapa 21, y los componentes QMF resultantes se aseveran al elemento de generación de parámetros de mejora 23. Los datos de voz mixto experimentan una transformación de dominio temporal a dominio de la frecuencia (QMF) en la etapa 22, y los componentes QMF resultantes se aseveran al elemento 23 y al subsistema de codificación 27.

Los datos de voz también se aseveran al subsistema 25 que se configura para generar datos de forma de onda (a los que a veces se hace referencia en la presente memoria como una copia de voz de "calidad reducida" o de " baja calidad ") indicativos de una copia de baja calidad de los datos de voz, para su uso en la mejora de la voz con forma de onda codificada del contenido mixto (voz y no voz) determinado por los datos de audio mixto. La copia de voz de baja calidad comprende menos bits que los datos de voz original, es de calidad objetable cuando se reproduce y percibe de manera aislada, y cuando se reproduce es indicativa de voz que tiene una forma de onda similar (p.ej., al menos sustancialmente similar) a la forma de onda de la voz indicada por los datos de voz original. Los métodos de implementación del sistema 25 son conocidos en la técnica. Ejemplos son los codificadores de voz de predicción lineal excitada por código (CELP, por sus siglas en inglés) como, por ejemplo, AMR y G729.1 o codificadores mixtos modernos como, por ejemplo, la Codificación Unificada de Voz y Audio (USAC, por sus siglas en inglés) MPEG, que normalmente funciona a una velocidad binaria baja (p.ej., 20 kbps). De manera alternativa, pueden usarse codificadores de dominio de la frecuencia, los ejemplos incluyen Siren (G722.1), MPEG 2 Capa II/III, MPEG AAC.

La mejora de la voz híbrida llevada a cabo (p.ej., en el subsistema 43 del decodificador 40) según las realizaciones típicas de la invención incluye la etapa de llevar a cabo (en los datos de forma de onda) lo inverso de la codificación llevada a cabo (p.ej., en el subsistema 25 del codificador 20) para generar los datos de forma de onda, para recuperar una copia de baja calidad del contenido de voz de la señal de audio mixto que se mejorará. La copia de

baja calidad recuperada de la voz se usa entonces (con datos paramétricos y datos indicativos de la señal de audio mixto) para llevar a cabo las restantes etapas de la mejora de la voz.

El elemento 23 se configura para generar datos paramétricos en respuesta a datos producidos desde las etapas 21 y 22. Los datos paramétricos, con los datos de audio mixto original, determinan la voz paramétricamente construida que es una versión paramétricamente reconstruida de la voz indicada por los datos de voz original (a saber, el contenido de voz de los datos de audio mixto). La versión paramétricamente reconstruida de la voz concuerda al menos sustancialmente con (p.ej., es una buena aproximación de) la voz indicada por los datos de voz original. Los datos paramétricos determinan un conjunto de parámetros de mejora paramétrica codificada, *p(t),* para llevar a cabo la mejora de la voz paramétrica codificada en cada segmento del contenido mixto no mejorado determinado por los datos de audio mixto.

El elemento de generación de indicador de mezcla 29 se configura para generar un indicador de mezcla ("BI", por sus siglas en inglés) en respuesta a los datos producidos desde las etapas 21 y 22. Se contempla que el programa de audio indicado por el tren de bits producido desde el codificador 20 experimentará una mejora de la voz híbrida (p.ej., en el decodificador 40) para determinar un programa de audio de voz mejorada, incluso mediante la combinación de los datos de audio no mejorados del programa original con una combinación de datos de voz de baja calidad (determinados a partir de los datos de forma de onda) y los datos paramétricos. El indicador de mezcla determina dicha combinación (p.ej., la combinación tiene una secuencia de estados determinados por una secuencia de valores actuales del indicador de mezcla), de modo que el programa de audio de voz mejorada tiene artefactos de codificación de mejora de la voz menos audibles (p.ej., artefactos de codificación de mejora de la voz que están mejor enmascarados) que un programa de audio de voz mejorada con forma de onda puramente codificada determinado mediante la combinación solamente de los datos de voz de baja calidad con los datos de audio no mejorados o un programa de audio de voz mejorada paramétrica codificada puramente determinado mediante la combinación solamente de la voz paramétricamente construida con los datos de audio no mejorados.

En variaciones de la realización de la Figura 3, el indicador de mezcla empleado para la mejora de la voz híbrida no se genera en el codificador (y no se incluye en el tren de bits producido desde el codificador), sino que se genera, en su lugar, (p.ej., en una variación del receptor 40) en respuesta al tren de bits producido desde el codificador (cuyo tren de bits incluye datos de forma de onda y datos paramétricos).

Debe comprenderse que la expresión "indicador de mezcla" no pretende denotar un solo parámetro o valor (o una secuencia de parámetros o valores únicos) para cada segmento del tren de bits. Más bien, se contempla que, en algunas realizaciones, un indicador de mezcla (para un segmento del tren de bits) puede ser un conjunto de dos o más parámetros o valores (p.ej., para cada segmento, un parámetro de control de mejora paramétrica codificada, y un parámetro de control de mejora con forma de onda codificada).

El subsistema de codificación 27 genera datos de audio codificados indicativos del contenido de audio de los datos de audio mixto (normalmente, una versión comprimida de los datos de audio mixto). El subsistema de codificación 27 normalmente implementa un inverso de la transformación llevada a cabo en la etapa 22 así como otras funciones de codificación.

La etapa de formateo 28 se configura para ensamblar los datos paramétricos producidos desde el elemento 23, los datos de forma de onda producidos desde el elemento 25, el indicador de mezcla generado en el elemento 29, y los datos de audio codificados producidos desde el subsistema 27 en un tren de bits codificado indicativo del programa de audio. El tren de bits (que puede tener formato E-AC-3 o AC-3, en algunas implementaciones) incluye los datos paramétricos no codificados, datos de forma de onda e indicador de mezcla.

El tren de bits de audio codificado (una señal de audio codificada) producido desde el codificador 20 se provee al subsistema de entrega 30. El subsistema de entrega 30 se configura para almacenar la señal de audio codificada (p.ej., para almacenar datos indicativos de la señal de audio codificada) generada por el codificador 20 y/o para transmitir la señal de audio codificada.

El decodificador 40 se acopla y se configura (p.ej., se programa) para recibir la señal de audio codificada del subsistema 30 (p.ej., mediante la lectura o recuperación de datos indicativos de la señal de audio codificada del almacenamiento en el subsistema 30, o la recepción de la señal de audio codificada que se ha transmitido por el subsistema 30) y para decodificar datos indicativos de contenido de audio mixto (voz y no voz) de la señal de audio codificada, y para llevar a cabo la mejora de la voz híbrida en el contenido de audio mixto decodificado. El decodificador 40 se configura normalmente para generar y producir (p.ej., para un sistema de reproducción, no se muestra en la Figura 3) una señal de audio de voz mejorada decodificada indicativa de una versión de voz mejorada del contenido de audio mixto ingresado al codificador 20. De manera alternativa, ello incluye dicho sistema de reproducción que se acopla para recibir la salida del subsistema 43.

La memoria intermedia 44 del decodificador 40 almacena (p.ej., en una manera no transitoria) al menos un segmento (p.ej., trama) de la señal de audio codificada (tren de bits) recibida por el decodificador 40. En el

funcionamiento típico, una secuencia de los segmentos del tren de bits de audio codificado se provee a la memoria intermedia 44 y se asevera de la memoria intermedia 44 a la etapa de desformateo 41.

La etapa de desformateo (análisis) 41 del decodificador 40 se configura para analizar el tren de bits codificado del subsistema de entrega 30, para extraer de allí los datos paramétricos (generados por el elemento 23 del codificador 20), los datos de forma de onda (generados por el elemento 25 del codificador 20), el indicador de mezcla (generado en el elemento 29 del codificador 20) y los datos de audio mixto (voz y no voz) codificados (generados en el subsistema de codificación 27 del codificador 20).

Los datos de audio mixto codificados se decodifican en el subsistema de decodificación 42 del decodificador 40 y los datos de audio mixto (voz y no voz) decodificados resultantes se aseveran al subsistema de mejora de la voz híbrida 43 (y se producen, opcionalmente, desde el decodificador 40 sin experimentar la mejora de la voz).

En respuesta a los datos de control (incluido el indicador de mezcla) extraídos por la etapa 41 del tren de bits (o generados en la etapa 41 en respuesta a los metadatos incluidos en el tren de bits), y en respuesta a los datos paramétricos y datos de forma de onda extraídos por la etapa 41, el subsistema de mejora de la voz 43 lleva a cabo la mejora de la voz híbrida en los datos de audio mixto (voz y no voz) decodificados del subsistema de decodificación 42 según una realización de la invención. La señal de audio de voz mejorada producida desde el subsistema 43 es indicativa de una versión de voz mejorada del contenido de audio mixto ingresado al codificador 20.

En varias implementaciones del codificador 20 de la Figura 3, el subsistema 23 puede generar cualquiera de los ejemplos descritos de parámetros de predicción, p_i, para cada losa de cada canal de la señal de entrada de audio mixto, para su uso (p.ej., en el decodificador 40) para la reconstrucción del componente de voz de una señal de audio mixto decodificada.

Con una señal de voz indicativa del contenido de voz de la señal de audio mixto decodificada (p.ej., la copia de baja calidad de la voz generada por el subsistema 25 del codificador 20, o una reconstrucción del contenido de voz generado mediante el uso de parámetros de predicción, p_i, generados por el subsistemas 23 del codificador 20), la mejora de la voz puede llevarse a cabo (p.ej., en el subsistema 43 del decodificador 40 de la Figura 3) mediante la mezcla de la señal de voz con la señal de audio mixto decodificada. Mediante la aplicación de una ganancia a la voz que se añadirá (mezclará), es posible controlar la cantidad de mejora de la voz. Para una mejora 6 dB, la voz puede añadirse con una ganancia 0 dB (siempre que la voz en la mezcla de voz mejorada tenga el mismo nivel que la señal de voz transmitida o reconstruida). La señal de voz mejorada es:

$$M_e = M + g \cdot D_r \tag{9}$$

En algunas realizaciones, para lograr una ganancia de mejora de la voz, G, se aplica la siguiente ganancia de mezcla:

$$g = 10^{G/20} - 1 \tag{10}$$

En el caso de la reconstrucción de voz independiente de canal, la mezcla de voz mejorada, M_e, se obtiene de la siguiente manera:

$$M_e = M \cdot (1 + diag(P) \cdot g) \tag{11}$$

En el ejemplo descrito más arriba, el aporte de voz en cada canal de la señal de audio mixto se reconstruye con la misma energía. Cuando la voz se ha transmitido como una señal lateral (p.ej., como una copia de baja calidad del contenido de voz de una señal de audio mixto) o cuando la voz se reconstruye mediante el uso de múltiples canales (como, por ejemplo, con un predictor MMSE), la mezcla de mejora de la voz requiere información de reproducción de voz con el fin de mezclar la voz con la misma distribución en los diferentes canales que el componente de voz ya presente en la señal de audio mixto que se mejorará.

La presente información de reproducción puede proveerse por un parámetro de reproducción r_i para cada canal, que puede representarse como un vector de reproducción R que tiene la forma

$$R = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} \tag{12}$$

cuando hay tres canales. La mezcla de mejora de la voz es:

$$M_e = M + R \cdot g \cdot D_r \tag{13}$$

En el caso de que existan múltiples canales, y la voz (que se mezclará con cada canal de una señal de audio mixto) se reconstruya mediante el uso de parámetros de predicción p_i, la ecuación previa puede escribirse como:

$$M_e = M + R \cdot g \cdot P \cdot M = (I + R \cdot g \cdot P) \cdot M \tag{14}$$

donde *I* es la matriz de identidad.

5. Reproducción de la voz

La Figura 4 es un diagrama de bloques de un sistema de reproducción de la voz que implementa la mezcla de mejora de la voz convencional de forma:

$$M_e = M + R \cdot g \cdot D_r \tag{15}$$

En la Figura 4, la señal de audio mixto de tres canales que se mejorará se encuentra (o se transforma en) el dominio de la frecuencia. Los componentes de frecuencia del canal izquierdo se aseveran a una entrada del elemento de mezcla 52, los componentes de frecuencia del canal central se aseveran a una entrada del elemento de mezcla 53 y los componentes de frecuencia del canal derecho se aseveran a una entrada del elemento de mezcla 54.

La señal de voz que se mezclará con la señal de audio mixto (para mejorar la última señal) puede haberse transmitido como una señal lateral (p.ej., como una copia de baja calidad del contenido de voz de la señal de audio mixto) o puede haberse reconstruido a partir de parámetros de predicción, p_i, transmitidos con la señal de audio mixto. La señal de voz se indica por datos de dominio de la frecuencia (p.ej., comprende componentes de frecuencia generados mediante la transformación de una señal de dominio temporal en el dominio de la frecuencia) y dichos componentes de frecuencia se aseveran a una entrada del elemento de mezcla 51, en el cual se multiplican por el parámetro de ganancia, *g*.

La salida del elemento 51 se asevera al subsistema de reproducción 50. También se aseveran al subsistema de reproducción 50 los parámetros CLD (diferencia de nivel de canal), CLD_1 y CLD_2, que se han transmitido con la señal de audio mixto. Los parámetros CLD (para cada segmento de la señal de audio mixto) describen cómo la señal de voz se mezcla para los canales de dicho segmento del contenido de señal de audio mixto. CLD_1 indica un coeficiente de panoramización para un par de canales de altavoz (p.ej., que define la panoramización de la voz entre los canales izquierdo y central), y CLD_2 indica un coeficiente de panoramización para otro par de los canales de altavoz (p.ej., que define la panoramización de la voz entre los canales central y derecho). Por consiguiente, el subsistema de reproducción 50 asevera (al elemento 52) datos indicativos de $R \cdot g \cdot D_r$ para el canal izquierdo (el contenido de voz, escalado por el parámetro de ganancia y el parámetro de reproducción para el canal izquierdo) y dichos datos se suman con el canal izquierdo de la señal de audio mixto en el elemento 52. El subsistema de reproducción 50 asevera (al elemento 53) datos indicativos de $R \cdot g \cdot D_r$ para el canal central (el contenido de voz, escalado por el parámetro de ganancia y el parámetro de reproducción para el canal central) y dichos datos se suman con el canal central de la señal de audio mixto en el elemento 53. El subsistema de reproducción 50 asevera (al elemento 54) datos indicativos de $R \cdot g \cdot D_r$ para el canal derecho (el contenido de voz, escalado por el parámetro de ganancia y el parámetro de reproducción para el canal derecho) y dichos datos se suman con el canal derecho de la señal de audio mixto en el elemento 54.

Las salidas de los elementos 52, 53 y 54 se emplean, respectivamente, para dirigir el altavoz izquierdo I, altavoz central C y altavoz derecho "Derecho".

La Figura 5 es un diagrama de bloques de un sistema de reproducción de la voz que implementa la mezcla de mejora de la voz convencional de forma:

$$M_e = M + R \cdot g \cdot P \cdot M = (I + R \cdot g \cdot P) \cdot M \quad (16)$$

En la Figura 5, la señal de audio mixto de tres canales que se mejorará se encuentra (o se transforma) en el dominio de la frecuencia. Los componentes de frecuencia del canal izquierdo se aseveran a una entrada del elemento de mezcla 52, los componentes de frecuencia del canal central se aseveran a una entrada del elemento de mezcla 53 y los componentes de frecuencia del canal derecho se aseveran a una entrada del elemento de mezcla 54.

La señal de voz que se mezclará con la señal de audio mixto se reconstruye (según se indica) a partir de los parámetros de predicción, p_i, transmitidos con la señal de audio mixto. El parámetro de predicción p_1 se emplea para reconstruir voz del primer canal (izquierdo) de la señal de audio mixto, el parámetro de predicción p_2 se emplea para reconstruir voz del segundo canal (central) de la señal de audio mixto y el parámetro de predicción p_3 se emplea para reconstruir voz del tercer canal (derecho) de la señal de audio mixto. La señal de voz se indica por datos de dominio de la frecuencia y dichos componentes de frecuencia se aseveran a una entrada del elemento de mezcla 51, en el cual se multiplican por el parámetro de ganancia, *g*.

La salida del elemento 51 se asevera al subsistema de reproducción 55. También se aseveran al subsistema de reproducción los parámetros CLD (diferencia de nivel de canal), CLD_1 y CLD_2, que se han transmitido con la señal de audio mixto. Los parámetros CLD (para cada segmento de la señal de audio mixto) describen cómo la señal de voz se mezcla para los canales de dicho segmento del contenido de señal de audio mixto. CLD_1 indica un coeficiente de panoramización para un par de canales de altavoz (p.ej., que define la panoramización de la voz entre los canales izquierdo y central), y CLD_2 indica un coeficiente de panoramización para otro par de los canales de altavoz (p.ej., que define la panoramización de la voz entre los canales central y derecho). Por consiguiente, el subsistema de reproducción 55 asevera (al elemento 52) datos indicativos de $R \cdot g \cdot P \cdot M$ para el canal izquierdo (el contenido de voz reconstruida mezclado con el canal izquierdo del contenido de audio mixto, escalado por el parámetro de ganancia y el parámetro de reproducción para el canal izquierdo, mezclado con el canal izquierdo del contenido de audio mixto) y dichos datos se suman con el canal izquierdo de la señal de audio mixto en el elemento 52. El subsistema de reproducción 55 asevera (al elemento 53) datos indicativos de $R \cdot g \cdot P \cdot M$ para el canal central (el contenido de voz reconstruida mezclado con el canal central del contenido de audio mixto, escalado por el parámetro de ganancia y el parámetro de reproducción para el canal central) y dichos datos se suman con el canal central de la señal de audio mixto en el elemento 53. El subsistema de reproducción 55 asevera (al elemento 54) datos indicativos de $R \cdot g \cdot P \cdot M$ para el canal derecho (el contenido de voz reconstruida mezclado con el canal derecho del contenido de audio mixto, escalado por el parámetro de ganancia y el parámetro de reproducción para el canal derecho) y dichos datos se suman con el canal derecho de la señal de audio mixto en el elemento 54.

Las salidas de los elementos 52, 53 y 54 se emplean, respectivamente, para dirigir el altavoz izquierdo I, altavoz central C y altavoz derecho "Derecho".

Los parámetros CLD (diferencia de nivel canal) se transmiten convencionalmente con señales de canal de altavoz (p.ej., para determinar relaciones entre los niveles en los cuales diferentes canales deben reproducirse). Estos se usan en una manera innovadora en algunas realizaciones de la invención (p.ej., para panoramizar la voz mejorada, entre canales de altavoz de un programa de audio de voz mejorada).

En realizaciones típicas, los parámetros de reproducción r_i son (o son indicativos de) coeficientes de mezcla ascendente de la voz y, por consiguiente, describen cómo la señal de voz se mezcla para los canales de la señal de audio mixto que se mejorará. Dichos coeficientes pueden transmitirse de manera eficaz al mejorador de voz mediante el uso de parámetros de diferencia de nivel de canal (CLD, por sus siglas en inglés). Una CLD indica coeficientes de panoramización para dos altavoces. Por ejemplo,

$$\beta_1 = \sqrt{\frac{1}{1 + 10^{\frac{CLD}{10}}}} \quad (17)$$

$$\beta_2 = \sqrt{\frac{10^{\frac{CLD}{10}}}{1 + 10^{\frac{CLD}{10}}}} \quad (18)$$

donde β_1 indica la ganancia para la alimentación de altavoz para el primer altavoz y β_2 indica la ganancia para la alimentación de altavoz para el segundo altavoz en un instante durante la panoramización. Con CLD = 0, la panoramización ocurre totalmente en el primer altavoz, mientras que con la infinidad de enfoque de CLD, la

panoramización ocurre totalmente hacia el segundo altavoz. Con CLD definidas en el dominio dB, un número limitado de niveles de cuantificación puede ser suficiente para describir la panoramización.

Con dos CLD, puede definirse la panoramización en tres altavoces. Las CLD pueden derivarse de la siguiente manera de los coeficientes de reproducción:

$$CLD_1 = 10 \cdot \log_{10}\left(\frac{\bar{r}_2^2}{\bar{r}_1^2}\right) \tag{19}$$

$$CLD_2 = 10 \cdot \log_{10}\left(\frac{\bar{r}_3^2}{\bar{r}_1^2+\bar{r}_2^2}\right) \tag{20}$$

donde $\bar{r}_x^2 = \frac{r_x^2}{\sum_i r_i^2}$ son los coeficientes de reproducción normalizados de modo que

$$\bar{r}_1^2 + \bar{r}_2^2 + \bar{r}_3^2 = 1 \tag{21}$$

Los coeficientes de reproducción pueden entonces reconstruirse a partir de las CLD por:

$$R = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{1}{\left(1+10^{\frac{CLD_1}{10}}\right)\left(1+10^{\frac{CLD_2}{10}}\right)}} \\ \sqrt{\frac{10^{\frac{CLD_1}{10}}}{\left(1+10^{\frac{CLD_1}{10}}\right)\cdot\left(1+10^{\frac{CLD_2}{10}}\right)}} \\ \sqrt{\frac{10^{\frac{CLD_2}{10}}}{1+10^{\frac{CLD_2}{10}}}} \end{pmatrix} \tag{22}$$

Según se advierte en otra parte en la presente memoria, la mejora de la voz con forma de onda codificada usa una copia de baja calidad del contenido de voz de la señal de contenido mixto que se mejorará. La copia de baja calidad se codifica normalmente a una velocidad binaria baja y se transmite como una señal lateral con la señal de contenido mixto y, por lo tanto, la copia de baja calidad normalmente contiene artefactos de codificación significativos. Por consiguiente, la mejora de la voz con forma de onda codificada provee un buen rendimiento de mejora de la voz en situaciones con una SNR baja (a saber, baja relación entre voz y todos los otros sonidos indicados por la señal de contenido mixto) y normalmente provee un rendimiento pobre (a saber, resulta en artefactos de codificación audibles indeseables) en situaciones con SNR alta.

Por el contrario, cuando el contenido de voz (de una señal de contenido mixto que se mejorará) se selecciona (p.ej., se provee como el único contenido de un canal central de una señal de contenido mixto multicanal) o la señal de contenido mixto de otra manera tiene una SNR alta, la mejora de la voz paramétrica codificada provee un buen rendimiento de mejora de la voz.

Por lo tanto, la mejora de la voz con forma de onda codificada y mejora de la voz paramétrica codificada tienen un rendimiento complementario. Según las propiedades de la señal cuyo contenido de voz se mejorará, una clase de realizaciones de la invención mezcla los dos métodos para apalancar sus rendimientos.

La Figura 6 es un diagrama de bloques de un sistema de reproducción de voz en la presente clase de realizaciones que se configura para llevar a cabo la mejora de la voz híbrida. En una implementación, el subsistema 43 del decodificador 40 de la Figura 3 realiza el sistema de la Figura 6 (excepto por los tres altavoces que se muestran en la Figura 6). La mejora de la voz híbrida (mezcla) puede describirse por

$$M_e = R \cdot g_1 \cdot D_r + (I + R \cdot g_2 \cdot P) \cdot M \quad (23)$$

donde $R \cdot g_1 \cdot D_r$ es una mejora de la voz con forma de onda codificada del tipo implementado por el sistema convencional de la Figura 4, $R \cdot g_2 \cdot P \cdot M$ es una mejora de la voz paramétrica codificada del tipo implementado por el sistema convencional de la Figura 5, y los parámetros g_1 y g_2 controlan la ganancia de mejora total y la transacción entre los dos métodos de mejora de la voz. Un ejemplo de una definición de los parámetros g_1 y g_2 es:

$$g_1 = \alpha_c \cdot (10^{G/20} - 1) \quad (24)$$

$$g_2 = (1 - \alpha_c) \cdot (10^{G/20} - 1) \quad (25)$$

donde el parámetro α_c define la transacción entre la mejora de la voz paramétrica codificada y métodos de mejora de la voz paramétrica codificada. Con un valor de $\alpha_c = 1$, solo la copia de baja calidad de la voz se usa para la mejora de la voz con forma de onda codificada. El modo de mejora paramétrica codificada es aportar totalmente a la mejora cuando $\alpha_c = 0$. Los valores de α_c entre 0 y 1 mezclan los dos métodos. En algunas implementaciones, α_c es un parámetro de banda ancha (que se aplica a todas las bandas de frecuencia de los datos de audio). Los mismos principios pueden aplicarse dentro de bandas de frecuencia individuales, de modo que la mezcla se optimiza en una manera dependiente de la frecuencia mediante el uso de un valor diferente del parámetro α_c para cada banda de frecuencia.

En la Figura 6, la señal de audio mixto de tres canales que se mejorará se encuentra (o se transforma) en el dominio de la frecuencia. Los componentes de frecuencia del canal izquierdo se aseveran a una entrada del elemento de mezcla 65, los componentes de frecuencia del canal central se aseveran a una entrada del elemento de mezcla 66 y los componentes de frecuencia del canal derecho se aseveran a una entrada del elemento de mezcla 67.

La señal de voz que se mezclará con la señal de audio mixto (para mejorar la última señal) incluye una copia de baja calidad (identificada como "Voz" en la Figura 6) del contenido de voz de la señal de audio mixto que se ha generado a partir de datos de forma de onda transmitidos (según la mejora de la voz con forma de onda codificada) con la señal de audio mixto (p.ej., como una señal lateral) y una señal de voz reconstruida (producida desde el elemento de reconstrucción de la voz paramétrica codificada 68 de la Figura 6) que se reconstruye a partir de la señal de audio mixto y parámetros de predicción, p_i, transmitidos (según la mejora de la voz paramétrica codificada) con la señal de audio mixto. La señal de voz se indica por datos de dominio de la frecuencia (p.ej., comprende componentes de frecuencia generados mediante la transformación de una señal de dominio temporal en el dominio de la frecuencia). Los componentes de frecuencia de la copia de voz de baja calidad se aseveran a una entrada al elemento de mezcla 61, en el cual se multiplican por el parámetro de ganancia, g_2. Los componentes de frecuencia de la señal de voz paramétricamente reconstruida se aseveran de la salida del elemento 68 a una entrada del elemento de mezcla 62, en el cual se multiplican por el parámetro de ganancia, g_1. En realizaciones alternativas, la mezcla llevada a cabo para implementar la mejora de la voz se lleva a cabo en el dominio temporal, antes que en el dominio de la frecuencia como en la realización de la Figura 6.

La salida de elementos 61 y 62 se suma por el elemento de suma 63 para generar la señal de voz que se mezclará con la señal de audio mixto, y dicha señal de voz se asevera de la salida del elemento 63 al subsistema de reproducción 64. También se aseveran al subsistema de reproducción 64 los parámetros CLD (diferencia de nivel de canal), CLD_1 y CLD_2, que se han transmitido con la señal de audio mixto. Los parámetros CLD (para cada segmento de la señal de audio mixto) describen cómo la señal de voz se mezcla para los canales de dicho segmento del contenido de señal de audio mixto. CLD_1 indica un coeficiente de panoramización para un par de canales de altavoz (p.ej., que define la panoramización de la voz entre los canales izquierdo y central), y CLD_2 indica un coeficiente de panoramización para otro par de los canales de altavoz (p.ej., que define la panoramización de la voz entre los canales central y derecho). Por consiguiente, el subsistema de reproducción 64 asevera (al elemento 52) datos indicativos de $R \cdot g_1 \cdot D_r + (R \cdot g_2 \cdot P) \cdot M$ para el canal izquierdo (el contenido de voz reconstruida mezclado con el canal izquierdo del contenido de audio mixto, escalado por el parámetro de ganancia y el parámetro de reproducción para el canal izquierdo, mezclado con el canal izquierdo del contenido de audio mixto) y dichos datos se suman con el canal izquierdo de la señal de audio mixto en el elemento 52. El subsistema de reproducción 64 asevera (al elemento 53) datos indicativos de $R \cdot g_1 \cdot D_r + (R \cdot g_2 \cdot P) \cdot M$ para el canal central (el contenido de voz reconstruida mezclado con el canal central del contenido de audio mixto, escalado por el parámetro de ganancia y el parámetro de reproducción para el canal central) y dichos datos se suman con el canal central de la señal de audio mixto en el elemento 53. El subsistema de reproducción 64 asevera (al elemento 54) datos indicativos de $R \cdot g1 \cdot D_r + (R \cdot g_2 \cdot P) \cdot M$ para el canal derecho (el contenido de voz reconstruida mezclado con el canal derecho del contenido de audio mixto, escalado por el parámetro de ganancia y el parámetro de reproducción para el canal derecho) y dichos datos se suman con el canal derecho de la señal de audio mixto en el elemento 54.

Las salidas de los elementos 52, 53 y 54 se emplean, respectivamente, para dirigir el altavoz izquierdo I, altavoz central C y altavoz derecho "Derecho".

El sistema de la Figura 6 puede implementar la conmutación basada en SNR temporal cuando el parámetro α_c se limita para que tenga el valor $\alpha_c = 0$ o el valor $\alpha_c = 1$. Dicha implementación es especialmente útil en situaciones de velocidad binaria fuertemente limitada en las cuales los datos de copia de voz de baja calidad pueden enviarse o los datos paramétricos pueden enviarse, pero no ambos. Por ejemplo, en una implementación, la copia de voz de baja calidad se transmite con la señal de audio mixto (p.ej., como una señal lateral) solamente en segmentos para los cuales $\alpha_c = 1$, y los parámetros de predicción, p_i, se transmiten con la señal de audio mixto (p.ej., como una señal lateral) solamente en segmentos para los cuales $\alpha_c = 0$.

La conmutación (implementada por los elementos 61 y 62 de la presente implementación de la Figura 6) determina si la mejora con forma de onda codificada o la mejora paramétrica codificada se llevará a cabo en cada segmento, según la relación (SNR) entre voz y todo el otro contenido de audio en el segmento (la presente relación determina, a su vez, el valor de α_c). Dicha implementación puede usar un valor umbral de la SNR para decidir qué método elegir:

$$\alpha_c = \begin{cases} 0 & si \;\; SNR > \tau \\ 1 & si \;\; SNR \leq \tau \end{cases} \tag{26}$$

donde τ es un valor umbral (p.ej., τ puede ser igual a 0).

Algunas implementaciones de la Figura 6 emplean la histéresis para evitar la conmutación alterna rápida entre los modos de mejora con forma de onda codificada y mejora paramétrica codificada cuando la SNR es de alrededor del valor umbral para varias tramas.

El sistema de la Figura 6 puede implementar la mezcla basada en SNR temporal cuando se permite que el parámetro α_c tenga cualquier valor real en el rango de 0 a 1, inclusive.

Una implementación del sistema de la Figura 6 usa dos valores objetivo, τ_1 y τ_2 (de la SNR de un segmento de la señal de audio mixto que se mejorará) más allá de los cuales siempre se considera que un método (la mejora con forma de onda codificada o mejora paramétrica codificada) provee el mejor rendimiento. Entre dichos objetivos, la interpolación se emplea para determinar el valor del parámetro α_c para el segmento. Por ejemplo, la interpolación lineal puede emplearse para determinar el valor del parámetro α_c para el segmento:

$$\alpha_c = \begin{cases} 0 & si \;\; SNR > \tau_2 \\ 1 - \dfrac{SNR - \tau_1}{\tau_2 - \tau_1} & si \;\; \tau_1 < SNR \leq \tau_2 \\ 1 & si \;\; SNR \leq \tau_1 \end{cases} \tag{27}$$

De manera alternativa, pueden usarse otros esquemas de interpolación apropiados. Cuando la SNR no se encuentra disponible, los parámetros de predicción en muchas implementaciones pueden usarse para proveer una aproximación de la SNR.

En otra clase de realizaciones, la combinación de mejora con forma de onda codificada y paramétrica codificada que se llevará a cabo en cada segmento de una señal de audio se determina por un modelo de enmascaramiento auditivo. En realizaciones típicas en la presente clase, la relación de mezcla óptima para una mezcla de mejora con forma de onda codificada y mejora paramétrica codificada que se llevará a cabo en un segmento de un programa de audio usa la cantidad más alta de mejora con forma de onda codificada que simplemente evita que el ruido de codificación se convierta en audible. Un ejemplo de una realización del método que emplea un modelo de enmascaramiento auditivo se describe en la presente memoria con referencia a la Figura 7.

De manera más general, las siguientes consideraciones pertenecen a realizaciones en las cuales un modelo de enmascaramiento auditivo se usa para determinar una combinación (p.ej., mezcla) de mejora con forma de onda codificada y paramétrica codificada que se llevará a cabo en cada segmento de una señal de audio. En dichas realizaciones, los datos indicativos de una mezcla de voz y audio de fondo, A(t), al que se hará referencia como una mezcla de audio no mejorada, se provee y procesa según el modelo de enmascaramiento auditivo (p.ej., el modelo implementado por el elemento 11 de la Figura 7). El modelo predice un umbral de enmascaramiento $\Theta(f,t)$ para cada segmento de la mezcla de audio no mejorada. El umbral de enmascaramiento de cada losa tiempo-frecuencia de la

mezcla de audio no mejorada, que tiene un índice temporal *n* y un índice de bandas de frecuencias *b,* puede denotarse como $\Theta_{n,b}$.

El umbral de enmascaramiento $\Theta_{n,b}$ indica para la trama *n* y banda *b* cuánta distorsión puede añadirse sin ser audible. Es preciso dejar que $\varepsilon_{D,n,b}$ sea el error de codificación (a saber, ruido de cuantificación) de la copia de voz de baja calidad (que se empleará para la mejora con forma de onda codificada), y $\varepsilon_{P,n,b}$ sea el error de predicción paramétrica.

Algunas realizaciones en la presente clase implementan una conmutación dura al método (mejora con forma de onda codificada o paramétrica codificada) que se enmascare mejor por el contenido mixto de audio no mejorado:

$$\alpha_c = \begin{cases} 0 & si \quad \sum_{n,b} \Theta_{n,b} - \varepsilon_{P,n,b} > \sum_{n,b} \Theta_{n,b} - \varepsilon_{D,n,b} \\ 1 & si \quad \sum_{n,b} \Theta_{n,b} - \varepsilon_{P,n,b} \leq \sum_{n,b} \Theta_{n,b} - \varepsilon_{D,n,b} \end{cases} \tag{28}$$

En muchas situaciones prácticas, el error de predicción paramétrica exacto $\varepsilon_{P,n,b}$ puede no estar disponible al momento de generación de los parámetros de mejora de la voz, dado que estos pueden generarse antes de que la mezcla mixta no mejorada se codifique. Esquemas de codificación especialmente paramétricos pueden tener un efecto significativo sobre el error de una reconstrucción paramétrica de la voz desde los canales de contenido mixto.

Por lo tanto, algunas realizaciones alternativas mezclan la mejora de la voz paramétrica codificada (con mejora con forma de onda codificada) cuando los artefactos de codificación en la copia de voz de baja calidad (que se empleará para la mejora con forma de onda codificada) no se enmascaran por el contenido mixto:

$$\alpha_c = \begin{cases} 1 & si \quad \sum_{n,b} \Theta_{n,b} - \varepsilon_{D,n,b} \geq 0 \\ 1 - \frac{\sum_{n,b} \Theta_{n,b} - \varepsilon_{D,n,b}}{\tau_a} & si \quad -\tau_a \leq \sum_{n,b} \Theta_{n,b} - \varepsilon_{D,n,b} < 0 \\ 0 & si \quad \sum_{n,b} \Theta_{n,b} - \varepsilon_{D,n,b} < -\tau_a \end{cases} \tag{29}$$

en el cual τ_a es un umbral de distorsión más allá del cual solo la mejora paramétrica codificada se aplica. La presente solución inicia la mezcla de la mejora con forma de onda codificada y paramétrica codificada cuando la distorsión total es más grande que el potencial de enmascaramiento total. En la práctica, ello significa que las distorsiones ya eran audibles. Por lo tanto, un segundo umbral puede usarse con un valor más alto que 0. De manera alternativa, uno puede usar condiciones que se centran más bien en las losas tiempo-frecuencia no enmascaradas en lugar del comportamiento promedio.

De manera similar, el presente enfoque puede combinarse con una regla de mezcla guiada por SNR cuando las distorsiones (artefactos de codificación) en la copia de voz de baja calidad (que se empleará para la mejora con forma de onda codificada) son demasiado altas. Una ventaja del presente enfoque es que en casos de SNR muy baja, el modo de mejora paramétrica codificada no se usa dado que produce ruido más audible que las distorsiones de la copia de voz de baja calidad.

En otra realización, el tipo de mejora de la voz llevado a cabo para algunas losas tiempo-frecuencia se desvía de aquella determinada por los esquemas a modo de ejemplo descritos más arriba (o esquemas similares) cuando un agujero espectral se detecta en cada losa tiempo-frecuencia. Los agujeros espectrales pueden detectarse, por ejemplo, mediante la evaluación de la energía en la losa correspondiente en la reconstrucción paramétrica mientras que la energía es 0 en la copia de voz de baja calidad (que se empleará para la mejora con forma de onda codificada). Si dicha energía supera un umbral, puede considerarse como audio relevante. En dichos casos, el parámetro α_c para la losa puede establecerse en 0 (o, dependiendo de la SNR, el parámetro α_c para la losa puede inclinarse hacia 0).

En algunos ejemplos que son útiles para comprender la invención, el codificador es utilizable en cualquier modo seleccionado de los siguientes modos:

1. Paramétrico independiente de canal - En el presente modo, un conjunto de parámetros se transmite para cada canal que contiene voz. Mediante el uso de dichos parámetros, un decodificador que recibe el programa de audio codificado puede llevar a cabo la mejora de la voz paramétrica codificada en el programa para impulsar la voz en dichos canales en una cantidad arbitraria. Una velocidad binaria a modo de ejemplo para la transmisión del conjunto de parámetros es de 0,75 - 2,25 kbps.

2. Predicción de voz multicanal - En el presente modo, múltiples canales del contenido mixto se combinan en una combinación lineal para predecir la señal de voz. Un conjunto de parámetros se transmite para cada canal. Mediante el uso de dichos parámetros, un decodificador que recibe el programa de audio codificado puede llevar a cabo la mejora de la voz paramétrica codificada en el programa. Datos posicionales adicionales se transmiten con el programa de audio codificado para permitir la reproducción de la voz impulsada otra vez hacia la mezcla. Una velocidad binaria a modo de ejemplo para la transmisión del conjunto de parámetros y datos posicionales es de 1,5 - 6,75 kbps por diálogo.

3. Voz con forma de onda codificada - En el presente modo, una copia de baja calidad del contenido de voz del programa de audio se transmite de forma separada, por cualquier medio adecuado, en paralelo con el contenido de audio regular (p.ej., como una subcorriente separada). Un decodificador que recibe el programa de audio codificado puede llevar a cabo la mejora de la voz con forma de onda codificada en el programa mediante la mezcla en la copia de baja calidad separada del contenido de voz con la mezcla principal. La mezcla de la copia de baja calidad de la voz con una ganancia de 0 dB impulsará, normalmente, la voz en 6 dB, dado que la amplitud se duplica. Para el presente modo, también se transmiten datos posicionales de modo que la señal de voz se distribuye correctamente en los canales relevantes. Una velocidad binaria a modo de ejemplo para la transmisión de la copia de baja calidad de la voz y datos posicionales es de más de 20 kbps por diálogo.

4. Híbrido con forma de onda paramétrica - En el presente modo, tanto una copia de baja calidad del contenido de voz del programa de audio (para su uso al llevar a cabo la mejora de la voz con forma de onda codificada en el programa) como un conjunto de parámetros para cada canal que contiene voz (para su uso al llevar a cabo la mejora de la voz paramétrica codificada en el programa) se transmiten en paralelo con el contenido de audio mixto (voz y no voz) no mejorado del programa. Cuando la velocidad binaria para la copia de baja calidad de la voz se reduce, más artefactos de codificación se convierten en audibles en la presente señal y el ancho de banda requerido para la transmisión se reduce. También se transmite un indicador de mezcla que determina una combinación de mejora de la voz con forma de onda codificada y mejora de la voz paramétrica codificada que se llevará a cabo en cada segmento del programa mediante el uso de la copia de baja calidad de la voz y el conjunto de parámetros. En un receptor, la mejora de la voz híbrida se lleva a cabo en el programa, incluso llevando a cabo una combinación de mejora de la voz con forma de onda codificada y mejora de la voz paramétrica codificada determinadas por el indicador de mezcla y, de esta manera, se generan datos indicativos de un programa de audio de voz mejorada. Nuevamente, los datos posicionales también se transmiten con el contenido de audio mixto no mejorado del programa para indicar dónde reproducir la señal de voz. Una ventaja del presente enfoque es que la complejidad receptor/decodificador requerida puede reducirse si el receptor/decodificador descarta la copia de baja calidad de la voz y aplica solamente el conjunto de parámetros para llevar a cabo la mejora paramétrica codificada. Una velocidad binaria a modo de ejemplo para la transmisión de la copia de baja calidad de la voz, conjunto de parámetros, indicador de mezcla y datos posicionales es de 8 - 24 kbps por diálogo.

Por motivos prácticos, la ganancia de mejora de la voz puede limitarse al rango de 0 - 12 dB. Un codificador puede implementarse para poder además reducir el límite superior del presente rango por medio de un campo de tren de bits. En algunas realizaciones, la sintaxis del programa codificado (salida del codificador) soportará múltiples diálogos mejorables simultáneos (además del contenido de no voz del programa), de modo que cada diálogo puede reconstruirse y reproducirse de manera separada. En dichas realizaciones, en los últimos modos, las mejoras de la voz para diálogos simultáneos (de múltiples fuentes en diferentes posiciones espaciales) se reproducirán en una sola posición.

En algunas realizaciones en las cuales el programa de audio codificado es un programa de audio basado en objeto, uno o más (del número total máximo) de clústeres de objetos pueden seleccionarse para la mejora de la voz. Pares de valores CLD pueden incluirse en el programa codificado para su uso por el sistema de mejora de la voz y reproducción para panoramizar la voz mejorada entre los clústeres de objetos. De manera similar, en algunas realizaciones en las cuales el programa de audio codificado incluye canales de altavoz en un formato 5.1 convencional, uno o más de los canales de altavoz frontales pueden seleccionarse para la mejora de la voz.

Otro aspecto de la presente descripción es un método (p.ej., un método llevado a cabo por el decodificador 40 de la Figura 3) para decodificar y llevar a cabo la mejora de la voz híbrida en una señal de audio codificada que se ha generado según una realización del método de codificación descrito.

La invención puede implementarse en hardware, firmware o software, o una combinación de ellos (p.ej., como una matriz lógica programable). Salvo que se especifique lo contrario, los algoritmos o procesos incluidos como parte de la invención no se relacionan de forma inherente con un ordenador particular u otro aparato. En particular, varias máquinas de propósito general pueden usarse con programas escritos según las enseñanzas de la presente memoria, o puede ser más conveniente construir aparatos más especializados (p.ej., circuitos integrados) para llevar a cabo las etapas requeridas del método. Por consiguiente, la invención puede implementarse en uno o más programas de ordenador que se ejecutan en uno o más sistemas informáticos programables (p.ej., un sistema informático que implementa el decodificador 40 de la Figura 3), cada uno comprendiendo al menos un procesador, al menos un sistema de almacenamiento de datos (incluida una memoria no permanente y permanente y/o elementos de almacenamiento), al menos un dispositivo o puerto de entrada, y al menos un dispositivo o puerto de salida. El código de programa se aplica a datos de entrada para llevar a cabo las funciones descritas en la presente memoria y generar información de salida. La información de salida se aplica a uno o más dispositivos de salida, de manera conocida.

Cada programa puede implementarse en cualquier lenguaje informático deseado (incluidos lenguajes de programación de máquina, conjunto o de procedimiento de alto nivel, lógico u orientado al objeto) para comunicarse con un sistema informático. En cualquier caso, el lenguaje puede ser un lenguaje compilado o interpretado.

Por ejemplo, cuando se implementan por secuencias de instrucciones de software de ordenador, varias funciones y etapas de las realizaciones de la invención pueden implementarse por secuencias de instrucciones de software multihilo que se ejecutan en hardware de procesamiento de señales digitales apropiado, en cuyo caso los diferentes dispositivos, etapas y funciones de las realizaciones pueden corresponder a porciones de las instrucciones de software.

Cada programa de ordenador se almacena, preferiblemente, en o se descarga a un medio o dispositivo de almacenamiento (p.ej., memoria o medios de estado sólido o medios magnéticos u ópticos) legible por un ordenador programable de propósito general o especial, para configurar y hacer funcionar el ordenador cuando el medio o dispositivo de almacenamiento se lee por el sistema informático para llevar a cabo los procedimientos descritos en la presente memoria. El sistema puede también implementarse como un medio de almacenamiento legible por ordenador, configurado con (a saber, que almacena) un programa de ordenador, donde el medio de almacenamiento así configurado hace que un sistema informático funcione en una manera específica y predefinida para llevar a cabo las funciones descritas en la presente memoria.

Se ha descrito un número de realizaciones de la invención. Sin embargo, se comprenderá que varias modificaciones pueden llevarse a cabo sin apartarse del alcance de la invención, según se define por las reivindicaciones anexas. Numerosas modificaciones y variaciones de la presente invención son posibles a la luz de las enseñanzas de más arriba. Se comprenderá que, dentro del alcance de las reivindicaciones anexas, la invención puede practicarse de manera diferente a como se describe específicamente en la presente memoria.

6. Representación media/lateral

Las funciones de mejora de la voz según se describe en la presente memoria pueden llevarse a cabo por un decodificador de audio según al menos en parte datos de control, parámetros de control, etc., en la representación M/S. Los datos de control, parámetros de control, etc., en la representación M/S pueden generarse por un codificador de audio corriente arriba y extraerse por el decodificador de audio de la señal de audio codificada generada por el codificador de audio corriente arriba.

En un modo de mejora paramétrica codificada en el cual el contenido de voz (p.ej., uno o más diálogos, etc.) se predice a partir de contenido mixto, las funciones de mejora de la voz pueden representarse, en general, con una sola matriz, *H*, como se muestra en la siguiente expresión:

$$\begin{pmatrix} M_{e,c_1} \\ M_{e,c_2} \end{pmatrix} = H \cdot \begin{pmatrix} M_{c_1} \\ M_{c_2} \end{pmatrix} \tag{30}$$

donde el lado izquierdo (LHS, por sus siglas en inglés) representa una señal de contenido mixto de voz mejorada generada por las funciones de mejora de la voz según se representa por la matriz *H* que funciona en una señal de contenido mixto original en el lado derecho (RHS, por sus siglas en inglés).

A los fines de la ilustración, cada una de la señal de contenido mixto de voz mejorada (p.ej., el LHS de la expresión (30), etc.) y la señal de contenido mixto original (p.ej., la señal de contenido mixto original operada por *H* en la expresión (30), etc.) comprende dos señales de componentes que tienen contenido de voz mejorada y mixto original en dos canales, c_1 y c_2, respectivamente. Los dos canales c_1 y c_2 pueden ser canales de audio no M/S (p.ej., canal frontal izquierdo, canal frontal derecho, etc.) según una representación no M/S. Debe notarse que, en varias realizaciones, cada una de la señal de contenido mixto de voz mejorada y la señal de contenido mixto original puede

además comprender señales de componentes que tienen contenido de no voz en los canales (p.ej., canales envolventes, un canal con efecto de frecuencia baja, etc.) diferentes de los dos canales no M/S c_1 y c_2. Debe notarse además que, en varias realizaciones, cada una de la señal de contenido mixto de voz mejorada y la señal de contenido mixto original puede posiblemente comprender señales de componentes que tienen contenido de voz en uno, dos, según se ilustra en la expresión (30), o más de dos canales. El contenido de voz según se describe en la presente memoria puede comprender uno, dos o más diálogos.

En algunas realizaciones, las funciones de mejora de la voz según se representa por H en la expresión (30) pueden usarse (p.ej., según lo ordenado por una regla de mezcla guiada por SNR, etc.) para segmentos de tiempo del contenido mixto con valores SNR relativamente altos entre el contenido de voz y otro contenido (p.ej., no voz, etc.) en el contenido mixto.

La matriz H puede reescribirse/expandirse como un producto de una matriz, H_{MS} que representa funciones de mejora en la representación M/S, multiplicarse en la derecha con una matriz de transformación hacia adelante de la representación no M/S a la representación M/S y multiplicarse en la izquierda con un inverso (que comprende un factor de 1/2) de la matriz de transformación hacia adelante, como se muestra en la siguiente expresión:

$$\begin{pmatrix} M_{e,c_1} \\ M_{e,c_2} \end{pmatrix} = \tfrac{1}{2} \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot H_{MS} \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} M_{c_1} \\ M_{c_2} \end{pmatrix} \tag{31}$$

donde la matriz de transformación a modo de ejemplo a la derecha de la matriz H_{MS} define la señal de contenido mixto de canal medio en la representación M/S como la suma de las dos señales de contenido mixto en los dos canales c_1 y c_2, y define la señal de contenido mixto de canal lateral en la representación M/S como la diferencia de las dos señales de contenido mixto en los dos canales c_1 y c_2, según la matriz de transformación hacia adelante. Debe notarse que, en varias realizaciones, otras matrices de transformación (p.ej., asignar diferentes ponderaciones a diferentes canales no M/S, etc.) diferentes de las matrices de transformación a modo de ejemplo que se muestran en la expresión (31) pueden también usarse para transformar las señales de contenido mixto de una representación a una representación diferente. Por ejemplo, para la mejora del diálogo con el diálogo reproducido no en el centro imaginario sino panoramizado entre las dos señales con ponderaciones desiguales λ_1 y λ_2. Las matrices de transformación M/S pueden modificarse para minimizar la energía del componente de diálogo en la señal lateral, como se muestra en la siguiente expresión:

$$\begin{pmatrix} M_{e,c_1} \\ M_{e,c_2} \end{pmatrix} = \tfrac{1}{2} \cdot \lambda_1 \cdot \lambda_2 \cdot \begin{pmatrix} \frac{1}{\lambda_2} & \frac{1}{\lambda_2} \\ \frac{1}{\lambda_1} & -\frac{1}{\lambda_1} \end{pmatrix} \cdot H_{MS} \cdot \begin{pmatrix} \frac{1}{\lambda_1} & \frac{1}{\lambda_2} \\ \frac{1}{\lambda_1} & -\frac{1}{\lambda_2} \end{pmatrix} \cdot \begin{pmatrix} M_{c_1} \\ M_{c_2} \end{pmatrix} \tag{31A}$$

En una realización a modo de ejemplo, la matriz H_{MS} que representa funciones de mejora en la representación M/S puede definirse como una matriz diagonalizada (p.ej., Hermitiana, etc.) como se muestra en la siguiente expresión:

$$H_{MS} = \begin{pmatrix} g \cdot p_1 + 1 & 0 \\ 0 & g \cdot p_2 + 1 \end{pmatrix} \tag{32}$$

donde p_1 y p_2 representan parámetros de predicción de canal medio y canal lateral, respectivamente. Cada uno de los parámetros de predicción p_1 y p_2 puede comprender un conjunto de parámetros de predicción de tiempo variable para las losas tiempo-frecuencia de una señal de contenido mixto correspondiente en la representación M/S que se usará para reconstruir contenido de voz desde la señal de contenido mixto. El parámetro de ganancia g corresponde a una ganancia de mejora de la voz, G, por ejemplo, como se muestra en la expresión (10).

En algunas realizaciones, las funciones de mejora de la voz en la representación M/S se llevan a cabo en el modo de mejora paramétrica independiente de canal. En algunas realizaciones, las funciones de mejora de la voz en la representación M/S se llevan a cabo con el contenido de voz pronosticado en la señal de canal medio y señal de canal lateral, o con el contenido de voz pronosticado en la señal de canal medio solamente. A los fines de la ilustración, las funciones de mejora de la voz en la representación M/S se llevan a cabo con la señal de contenido mixto en el canal medio solamente, como se muestra en la siguiente expresión:

$$H_{MS} = \begin{pmatrix} g \cdot p_1 + 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{33}$$

donde el parámetro de predicción p_1 comprende un solo conjunto de parámetros de predicción para losas tiempo-frecuencia de la señal de contenido mixto en el canal medio de la representación M/S que se usará para reconstruir el contenido de voz de la señal de contenido mixto en el canal medio solamente.

Según la matriz diagonalizada H_{MS} dada en la expresión (33), las funciones de mejora de la voz en el modo de mejora paramétrica, según se representa por la expresión (31), pueden además reducirse a la siguiente expresión, que provee un ejemplo explícito de la matriz H en la expresión (30):

$$\begin{pmatrix} M_{e,c_1} \\ M_{e,c_2} \end{pmatrix} = \frac{1}{2} \cdot \begin{pmatrix} 2 + g \cdot p_1 & g \cdot p_1 \\ g \cdot p_1 & 2 + g \cdot p_1 \end{pmatrix} \cdot \begin{pmatrix} M_{c_1} \\ M_{c_2} \end{pmatrix} \quad (34)$$

En un modo de mejora híbrido con forma de onda paramétrica, las funciones de mejora de la voz pueden representarse en la representación M/S con las siguientes expresiones a modo de ejemplo:

$$M_e = g_1 \cdot \begin{pmatrix} d_{c,1} \\ 0 \end{pmatrix} + \begin{pmatrix} g_2 \cdot p_1 + 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \quad (35)$$

$$= H_d \cdot D_c + H_p \cdot M$$

donde m_1 y m_2 denotan la señal de contenido mixto de canal medio (p.ej., la suma de las señales de contenido mixto en los canales no M/S como, por ejemplo, canales frontales izquierdo y derecho) y la señal de contenido mixto de canal lateral (p.ej., la diferencia de las señales de contenido mixto en los canales no M/S como, por ejemplo, canales frontales izquierdo y derecho, etc.), respectivamente, en un vector de señal de contenido mixto M. A señal, $d_{c,1}$ denota la señal de forma de onda de diálogo de canal medio (p.ej., formas de onda codificadas que representan una versión reducida de un diálogo en el contenido mixto, etc.) en un vector de señal de diálogo D_c de la representación M/S. Una matriz, H_d, representa funciones de mejora de la voz en la representación M/S según la señal de diálogo $d_{c,1}$ en el canal medio de la representación M/S, y puede comprender solamente un elemento de matriz en la fila 1 y columna 1 (1x1). Una matriz, H_p, representa funciones de mejora de la voz en la representación M/S según un diálogo reconstruido mediante el uso del parámetro de predicción p_1 para el canal medio de la representación M/S. En algunas realizaciones, los parámetros de ganancia g_1 y g_2 conjuntamente (p.ej., después de aplicarse respectivamente a la señal de forma de onda de diálogo y al diálogo reconstruido, etc.) corresponden a una ganancia de mejora de la voz, G, por ejemplo, según se representa en las expresiones (23) y (24). De manera específica, el parámetro g_1 se aplica en las funciones de mejora de la voz con forma de onda codificada relacionadas con la señal de diálogo $d_{c,1}$ en el canal medio de la representación M/S, mientras que el parámetro g_2 se aplica en las funciones de mejora de la voz paramétrica codificada relacionadas con las señales de contenido mixto m_1 y $m2$ en el canal medio y canal lateral de la representación M/S. Los parámetros g_1 y g_2 controlan la ganancia de mejora total y la transacción entre los dos métodos de mejora de la voz.

En la representación no M/S, las funciones de mejora de la voz correspondientes a aquellas representadas con la expresión (35) pueden representarse con las siguientes expresiones:

$$\begin{pmatrix} M_{e,c_1} \\ M_{e,c_2} \end{pmatrix} = \frac{1}{2} \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot H_d \cdot D_c + \frac{1}{2} \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot H_p \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} M_{c_1} \\ M_{c_2} \end{pmatrix} \quad (36)$$

$$= \frac{1}{2} \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \left(H_d \cdot D_c + H_p \cdot \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} M_{c_1} \\ M_{c_2} \end{pmatrix} \right)$$

donde las señales de contenido mixto m_1 y m_2 en la representación M/S según se muestra en la expresión (35) se reemplazan por las señales de contenido mixto M_{c1} y M_{c2} en los canales no M/S multiplicados con la matriz de transformación hacia adelante entre la representación no M/S y la representación M/S. La matriz de transformación inversa (con un factor de ½) en la expresión (36) convierte las señales de contenido mixto de voz mejorada en la representación M/S, como se muestra en la expresión (35), otra vez en las señales de contenido mixto de voz mejorada en la representación no M/S (p.ej., canales frontales izquierdo y derecho, etc.).

De manera adicional, opcional o alternativa, en algunas realizaciones en las cuales no se realiza un procesamiento adicional basado en QMF después de las funciones de mejora de la voz, algunas o todas las funciones de mejora de la voz (p.ej., según se representa por H_d, H_p, transformaciones, etc.) que combinan contenido de voz mejorada según la señal de diálogo $d_{c,1}$ y contenido mixto de voz mejorada según el diálogo reconstruido a través de la predicción pueden llevarse a cabo después de un banco de filtros de síntesis QMF en el dominio temporal por motivos de eficacia.

Un parámetro de predicción usado para construir/predecir contenido de voz de una señal de contenido mixto en uno o ambos del canal medio y canal lateral de la representación M/S puede generarse según uno de uno o más

métodos de generación de parámetros de predicción incluidos, pero sin limitación a ello, cualquiera de: métodos de predicción de diálogo independiente de canal según se representa en la Figura 1, métodos de predicción de diálogo multicanal según se representa en la Figura 2, etc. En algunas realizaciones, al menos uno de los métodos de generación de parámetros de predicción puede basarse en MMSE, descenso del gradiente, uno o más de otros métodos de optimización, etc.

En algunas realizaciones, un método de conmutación basado en SNR temporal "a ciegas" según se describe previamente puede usarse entre datos de mejora paramétrica codificada (p.ej., relacionados con contenido de voz mejorada según la señal de diálogo $d_{c,1}$, etc.) y mejora con forma de onda codificada (p.ej., relacionada con contenido mixto de voz mejorada en el diálogo reconstruido a través de la predicción, etc.) de segmentos de un programa de audio en la representación M/S.

En algunas realizaciones, una combinación (p.ej., indicada por un indicador de mezcla previamente descrito, una combinación de g_1 y g_2 en la expresión (35), etc.) de los datos de forma de onda (p.ej., relacionados con contenido de voz mejorada según la señal de diálogo $d_{c,1}$, etc.) y los datos de voz reconstruida (p.ej., relacionados con contenido mixto de voz mejorada según el diálogo reconstruido a través de la predicción, etc.) en la representación M/S cambia con el tiempo, con cada estado de la combinación perteneciendo al contenido de voz y otro contenido de audio de un segmento correspondiente del tren de bits que lleva los datos de forma de onda y el contenido mixto usado en la reconstrucción de datos de voz. El indicador de mezcla se genera de modo que el estado actual de la combinación (de datos de forma de onda y datos de voz reconstruida) se determina por propiedades de señal de la voz y otro contenido de audio (p.ej., una relación de la potencia del contenido de voz y la potencia de otro contenido de audio, una SNR, etc.) en el segmento correspondiente del programa. El indicador de mezcla para un segmento de un programa de audio puede ser un parámetro (o conjunto de parámetros) de indicador de mezcla generado en el subsistema 29 del codificador de la Figura 3 para el segmento. Un modelo de enmascaramiento auditivo según se describe previamente puede usarse para predecir de manera más exacta cómo los ruidos de codificación en la copia de voz de calidad reducida en el vector de señal de diálogo *Dc* se enmascaran por la mezcla de audio del programa principal y para seleccionar la relación de mezcla de manera acorde.

El subsistema 28 del codificador 20 de la Figura 3 puede configurarse para incluir indicadores de mezcla relacionados con funciones de mejora de la voz M/S en el tren de bits como parte de los metadatos de mejora de la voz M/S que se producirán desde el codificador 20. Los indicadores de mezcla relacionados con las funciones de mejora de la voz M/S pueden generarse (p.ej., en el subsistema 13 del codificador de la Figura 7) a partir de factores de escalamiento $g_{máx}(t)$ relacionados con artefactos de codificación en la señal de diálogo *Dc,* etc. Los factores de escalamiento $g_{máx}(t)$ pueden generarse por el subsistema 14 del codificador de la Figura 7. El subsistema 13 del codificador de la Figura 7 puede configurarse para incluir los indicadores de mezcla en el tren de bits que se producirá desde el codificador de la Figura 7. De manera adicional, opcional o alternativa, el subsistema 13 puede incluir, en el tren de bits que se producirá desde el codificador de la Figura 7, los factores de escalamiento $g_{máx}(t)$ generados por el subsistema 14.

En algunas realizaciones, la mezcla de audio no mejorada, A(t), generada por la función 10 de la Figura 7 representa (p.ej., segmentos de tiempo de, etc.) un vector de señal de contenido mixto en la configuración de canal de audio de referencia. Los parámetros de mejora paramétrica codificada, p(t), generados por el elemento 12 de la Figura 7 representan al menos una parte de los metadatos de mejora de la voz M/S para llevar a cabo la mejora de la voz paramétrica codificada en la representación M/S con respecto a cada segmento del vector de señal de contenido mixto. En algunas realizaciones, la copia de voz de calidad reducida, s'(t), generada por el codificador 15 de la Figura 7 representa un vector de señal de diálogo en la representación M/S (p.ej., con la señal de diálogo de canal medio, la señal de diálogo de canal lateral, etc.).

En algunas realizaciones, el elemento 14 de la Figura 7 genera los factores de escalamiento, $g_{máx}(t)$, y los provee al elemento de codificación 13. En algunas realizaciones el elemento 13 genera un tren de bits de audio codificado indicativo del vector de señal de contenido mixto (p.ej., no mejorado, etc.) en la configuración de canal de audio de referencia, los metadatos de mejora de la voz M/S, el vector de señal de diálogo en la representación M/S si fuera aplicable, y los factores de escalamiento $g_{máx}(t)$ si fuera aplicable, para cada segmento del programa de audio, y el presente tren de bits de audio codificado pueden transmitirse o de otra manera entregarse a un receptor.

Cuando la señal de audio no mejorada en una representación no M/S se entrega (p.ej., se transmite) con metadatos de mejora de la voz M/S a un receptor, el receptor puede transformar cada segmento de la señal de audio no mejorada en la representación M/S y llevar a cabo funciones de mejora de la voz M/S indicadas por los metadatos de mejora de la voz M/S para el segmento. El vector de señal de diálogo en la representación M/S para un segmento del programa puede proveerse con el vector de señal de contenido mixto no mejorado en la representación no M/S si las funciones de mejora de la voz para el segmento se llevan a cabo en el modo de mejora de la voz híbrido, o en el modo de mejora con forma de onda codificada. Si fuera aplicable, un receptor que recibe y analiza el tren de bits puede configurarse para generar los indicadores de mezcla en respuesta a los factores de escalamiento $g_{máx}(t)$ y determinar los parámetros de ganancia g_1 y g_2 en la expresión (35).

En algunas realizaciones, las funciones de mejora de la voz se llevan a cabo al menos parcialmente en la representación M/S en un receptor al cual la salida codificada del elemento 13 se ha entregado. En un ejemplo, en cada segmento de la señal de contenido mixto no mejorada, los parámetros de ganancia g_1 y g_2 en la expresión (35) correspondientes a una cantidad total predeterminada (p.ej., solicitada) de mejora pueden aplicarse según al menos en parte indicadores de mezcla analizados desde el tren de bits recibido por el receptor. En otro ejemplo, en cada segmento de la señal de contenido mixto no mejorada, los parámetros de ganancia g_1 y g_2 en la expresión (35) correspondientes a una cantidad total predeterminada (p.ej., solicitada) de mejora pueden aplicarse según al menos en parte indicadores de mezcla determinados a partir de factores de escala $g_{máx}(t)$ para el segmento analizado desde el tren de bits recibido por el receptor.

En algunas realizaciones, el elemento 23 del codificador 20 de la Figura 3 se configura para generar datos paramétricos que incluyen metadatos de mejora de la voz M/S (p.ej., parámetros de predicción para reconstruir contenido diálogo/voz a partir de contenido mixto en el canal medio y/o en el canal lateral, etc.) en respuesta a datos producidos desde las etapas 21 y 22. En algunas realizaciones, el elemento de generación de indicador de mezcla 29 del codificador 20 de la Figura 3 se configura para generar un indicador de mezcla ("BI") para determinar una combinación de contenido de voz paramétricamente mejorada (p.ej., con el parámetro de ganancia g_1, etc.) y contenido de voz mejorada basado en forma de onda (p.ej., con el parámetro de ganancia g_1, etc.) en respuesta a los datos producidos desde las etapas 21 y 22.

En variaciones de la realización de la Figura 3, el indicador de mezcla empleado para la mejora de la voz híbrida M/S no se genera en el codificador (y no se incluye en el tren de bits producido desde el codificador), sino que se genera, en su lugar, (p.ej., en una variación del receptor 40) en respuesta al tren de bits producido desde el codificador (cuyo tren de bits incluye datos de forma de onda en los canales M/S y metadatos de mejora de la voz M/S).

El decodificador 40 se acopla y configura (p.ej., se programa) para recibir la señal de audio codificada del subsistema 30 (p.ej., mediante la lectura o recuperación de datos indicativos de la señal de audio codificada del almacenamiento en el subsistema 30, o la recepción de la señal de audio codificada que se ha transmitido por el subsistema 30) y para decodificar datos indicativos de vector de señal de contenido mixto (voz y no voz) en la configuración de canal de audio de referencia de la señal de audio codificada, y para llevar a cabo funciones de mejora de la voz al menos en parte en la representación M/S en el contenido mixto decodificado en la configuración de canal de audio de referencia. El decodificador 40 puede configurarse para generar y producir (p.ej., a un sistema de reproducción, etc.) una señal de audio decodificada de voz mejorada indicativa de contenido mixto de voz mejorada.

En algunas realizaciones, algunos o todos los sistemas de reproducción ilustrados en la Figura 4 a la Figura 6 pueden configurarse para reproducir contenido mixto de voz mejorada generado por funciones de mejora de la voz M/S al menos algunos de las cuales son funciones llevadas a cabo en la representación M/S. La Figura 6A ilustra un sistema de reproducción a modo de ejemplo configurado para llevar a cabo las funciones de mejora de la voz según se representa en la expresión (35).

El sistema de reproducción de la Figura 6A puede configurarse para llevar a cabo funciones de mejora de la voz paramétrica en respuesta a la determinación de que al menos un parámetro de ganancia (p.ej., g_2 en la expresión (35), etc.) usado en las funciones de mejora de la voz paramétrica es no cero (p.ej., en modo de mejora híbrido, en modo de mejora paramétrico, etc.). Por ejemplo, después de dicha determinación, el subsistema 68A de la Figura 6A puede configurarse para llevar a cabo una transformación en un vector de señal de contenido mixto ("audio mixto (T/F)") que se distribuye en canales no M/S para generar un vector de señal de contenido mixto correspondiente que se distribuye en canales M/S. La presente transformación puede usar una matriz de transformación hacia adelante según corresponda. Los parámetros de predicción (p.ej.., p_1, p_2, etc.), parámetros de ganancia (p.ej., g_2 en la expresión (35), etc.) para funciones de mejora paramétrica pueden aplicarse para predecir contenido de voz del vector de señal de contenido mixto de los canales M/S y mejorar el contenido de voz pronosticado.

El sistema de reproducción de la Figura 6A puede configurarse para llevar a cabo funciones de mejora de la voz con forma de onda codificada en respuesta a la determinación de que al menos un parámetro de ganancia (p.ej., g_1 en la expresión (35), etc.) usado en las funciones de mejora de la voz con forma de onda codificada es no cero (p.ej., en modo de mejora híbrido, en modo de mejora con forma de onda codificada, etc.). Por ejemplo, después de dicha determinación, el sistema de reproducción de la Figura 6A puede configurarse para recibir/extraer, de la señal de audio codificada recibida, un vector de señal de diálogo (p.ej., con una versión reducida de contenido de voz presente en el vector de señal de contenido mixto) que se distribuye en canales M/S. Los parámetros de ganancia (p.ej., g_1 en la expresión (35), etc.) para funciones de mejora con forma de onda codificada pueden aplicarse para mejorar contenido de voz representado por el vector de señal de diálogo de los canales M/S. Una ganancia (G) de mejora definible por el usuario puede usarse para derivar parámetros de ganancia g1 y g2 mediante el uso de un parámetro de mezcla, que puede o no estar presente en el tren de bits. En algunas realizaciones, el parámetro de mezcla que se usará con la ganancia (G) de mejora definible por el usuario para derivar parámetros de ganancia g1 y g2 puede extraerse de metadatos en la señal de audio codificada recibida. En algunas otras realizaciones, dicho

parámetro de mezcla no puede extraerse de datos en la señal de audio codificada recibida, sino que, más bien, puede derivarse por un codificador receptor según el contenido de audio en la señal de audio codificada recibida.

En algunas realizaciones, una combinación del contenido de voz mejorada paramétrico y el contenido de voz mejorada con forma de onda codificada en la representación M/S se asevera o ingresa al subsistema 64A de la Figura 6A. El subsistema 64A de la Figura 6 puede configurarse para llevar a cabo una transformación en la combinación de contenido de voz mejorada que se distribuye en canales M/S para generar un vector de señal de contenido de voz mejorada que se distribuye en canales no M/S. La presente transformación puede usar una matriz de transformación inversa según corresponda. El vector de señal de contenido de voz mejorada de los canales no M/S puede combinarse con el vector de señal de contenido mixto ("audio mixto (T/F)") que se distribuye en los canales no M/S para generar un vector de señal de contenido mixto de voz mejorada.

En algunas realizaciones, la sintaxis de la señal de audio codificada (p.ej., producida desde el codificador 20 de la Figura 3, etc.) soporta una transmisión de una bandera M/S de un codificador de audio corriente arriba (p.ej., codificador 20 de la Figura 3, etc.) a decodificadores de audio corriente abajo (p.ej., decodificador 40 de la Figura 3, etc.). La bandera M/S está presente/se establece por el codificador de audio (p. ej., elemento 23 en el codificador 20 de la Figura 3, etc.) cuando funciones de mejora de la voz se llevan a cabo por un decodificador de audio receptor (p.ej., decodificador 40 de la Figura 3, etc.) al menos en parte con datos de control M/S, parámetros de control, etc., que se transmiten con la bandera M/S. Por ejemplo, cuando la bandera M/S se establece, una señal estéreo (p.ej., de canales izquierdo y derecho, etc.) en canales no M/S puede transformarse primero por el decodificador de audio receptor (p.ej., decodificador 40 de la Figura 3, etc.) en el canal medio y el canal lateral de la representación M/S antes de aplicar funciones de mejora de la voz M/S con los datos de control M/S, parámetros de control, etc., como recibidos con la bandera M/S, según uno o más de los algoritmos de mejora de la voz (p.ej., predicción de diálogo independiente de canal, predicción de diálogo multicanal, basados en forma de onda, híbridos de forma de onda-paramétrico, etc.). En el decodificador de audio receptor (p.ej., decodificador 40 de la Figura 3, etc.), después de llevar a cabo las funciones de mejora de la voz M/S, las señales de voz mejoradas en la representación M/S pueden transformarse otra vez en los canales no M/S.

En algunas realizaciones, los metadatos de mejora de la voz generados por un codificador de audio (p.ej., codificador 20 de la Figura 3, elemento 23 del codificador 20 de la Figura 3, etc.) según se describe en la presente memoria pueden llevar una o más banderas específicas para indicar la presencia de uno o más conjuntos de datos de control de mejora de la voz, parámetros de control, etc., para uno o más tipos diferentes de funciones de mejora de la voz. El único o más conjuntos de datos de control de mejora de la voz, parámetros de control, etc., para el único o más tipos diferentes de funciones de mejora de la voz pueden, pero sin limitación a ello solamente, incluir un conjunto de datos de control M/S, parámetros de control, etc., como metadatos de mejora de la voz M/S. Los metadatos de mejora de la voz pueden también incluir una bandera de preferencia para indicar qué tipo de funciones de mejora de la voz (p.ej., funciones de mejora de la voz M/S, funciones de mejora de la voz no M/S, etc.) se prefieren para que el contenido de audio sea de voz mejorada. Los metadatos de mejora de la voz pueden entregarse a un decodificador corriente abajo (p.ej., decodificador 40 de la Figura 3, etc.) como parte de los metadatos entregados en una señal de audio codificada que incluye contenido de audio mixto codificado para una configuración de canal de audio de referencia no M/S. En algunas realizaciones, solo los metadatos de mejora de la voz M/S pero no los metadatos de mejora de la voz no M/S se incluyen en la señal de audio codificada.

De manera adicional, opcional o alternativa, un decodificador de audio (p.ej., 40 de la Figura 3, etc.) puede configurarse para determinar y llevar a cabo un tipo específico (p.ej., mejora de la voz M/S, mejora de la voz no M/S, etc.) de funciones de mejora de la voz según uno o más factores. Dichos factores pueden incluir, pero sin limitación a ello solamente: una o más de una entrada de usuario que especifica una preferencia para un tipo seleccionado por el usuario específico de función de mejora de la voz, entrada de usuario que especifica una preferencia para un tipo de sistema seleccionado de funciones de mejora de la voz, capacidades de la configuración de canal de audio específica que funcionan por el decodificador de audio, disponibilidad de metadatos de mejora de la voz para el tipo específico de función de mejora de la voz, cualquier bandera de preferencia generada por codificador para un tipo de función de mejora de la voz, etc. En algunas realizaciones, el decodificador de audio puede implementar una o más reglas de precedencia, puede solicitar una entrada de usuario adicional, etc., para determinar un tipo específico de función de mejora de la voz si dichos factores entran en conflicto los unos con los otros.

7. Flujos de proceso a modo de ejemplo

La Figura 8A y Figura 8B ilustran flujos de proceso a modo de ejemplo. En algunas realizaciones, uno o más dispositivos o unidades informáticas en un sistema de procesamiento de medios pueden llevar a cabo el presente flujo de proceso.

La Figura 8A ilustra un proceso a modo de ejemplo que es útil para comprender la invención. El flujo de proceso a modo de ejemplo puede implementarse por un codificador de audio (p.ej., codificador 20 de la Figura 3) según se describe en la presente memoria. En el bloque 802 de la Figura 8A, el codificador de audio recibe contenido de audio mixto, con una mezcla de contenido de voz y contenido de audio de no voz, en una representación de canal

de audio de referencia, que se distribuye en múltiples canales de audio de la representación de canal de audio de referencia.

En el bloque 804, el codificador de audio transforma una o más porciones del contenido de audio mixto que se distribuyen en uno o más canales no Medio/Lateral (M/S) en los múltiples canales de audio de la representación de canal de audio de referencia en una o más porciones de contenido de audio mixto transformado en una representación de canal de audio M/S que se distribuyen en uno o más canales M/S de la representación de canal de audio M/S.

En el bloque 806, el codificador de audio determina los metadatos de mejora de la voz M/S para la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S.

En el bloque 808, el codificador de audio genera una señal de audio que comprende el contenido de audio mixto en la representación de canal de audio de referencia y los metadatos de mejora de la voz M/S para la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S.

En un ejemplo, el codificador de audio se configura además para llevar a cabo: la generación de una versión del contenido de voz, en la representación de canal de audio M/S, de manera independiente del contenido de audio mixto; y la producción de la señal de audio codificada con la versión del contenido de voz en la representación de canal de audio M/S.

En un ejemplo, el codificador de audio se configura además para llevar a cabo: la generación de datos que indican la mezcla que permiten que un decodificador de audio receptor aplique la mejora de la voz al contenido de audio mixto con una combinación cuantitativa específica de mejora de la voz con forma de onda codificada según la versión del contenido de voz en la representación de canal de audio M/S y mejora de la voz paramétrica según una versión reconstruida del contenido de voz en la representación de canal de audio M/S; y la emisión de la señal de audio codificada con los datos que indican la mezcla.

En un ejemplo, el codificador de audio se configura además para evitar la codificación de la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S como parte de la señal de audio.

La Figura 8B ilustra un flujo de proceso a modo de ejemplo que puede implementarse por un decodificador de audio (p.ej., decodificador 40 de la Figura 3) según se describe en la presente memoria. En el bloque 822 de la Figura 8B, el decodificador de audio recibe una señal de audio que comprende contenido de audio mixto en una representación de canal de audio de referencia y metadatos de mejora de la voz Media/Lateral (M/S).

En el bloque 824 de la Figura 8B, el decodificador de audio transforma una o más porciones del contenido de audio mixto que se distribuyen en uno, dos o más canales no M/S en múltiples canales de audio de la representación de canal de audio de referencia en una o más porciones de contenido de audio mixto transformado en una representación de canal de audio M/S que se distribuyen en uno o más canales M/S de la representación de canal de audio M/S.

En el bloque 826 de la Figura 8B, el decodificador de audio lleva a cabo una o más funciones de mejora de la voz M/S, según los metadatos de mejora de la voz M/S, en la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S para generar una o más porciones de contenido de voz mejorada en la representación M/S.

En el bloque 828 de la Figura 8B, el decodificador de audio combina la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S con el único o más de contenido de voz mejorada en la representación M/S para generar una o más porciones de contenido de audio mixto de voz mejorad en la representación M/S.

En una realización, el decodificador de audio se configura además para transformar, de manera inversa, la única o más porciones de contenido de audio mixto de voz mejorada en la representación M/S en una o más porciones de contenido de audio mixto de voz mejorada en la representación de canal de audio de referencia.

En una realización, el decodificador de audio se configura además para llevar a cabo: la extracción de una versión del contenido de voz, en la representación de canal de audio M/S, de manera independiente del contenido de audio mixto de la señal de audio; y llevar a cabo una o más funciones de mejora de la voz, según los metadatos de mejora de la voz M/S, en una o más porciones de la versión del contenido de voz en la representación de canal de audio M/S para generar una o más segundas porciones de contenido de voz mejorada en la representación de canal de audio M/S.

En una realización, el decodificador de audio se configura además para llevar a cabo: la determinación de datos que indican mezcla para la mejora de la voz; y la generación, según los datos que indican la mezcla para la mejora de la voz, de una combinación cuantitativa específica de mejora de la voz con forma de onda codificada según la versión

del contenido de voz en la representación de canal de audio M/S y mejora de la voz paramétrica según una versión reconstruida del contenido de voz en la representación de canal de audio M/S.

En una realización, los datos que indican la mezcla se generan según al menos en parte uno o más valores SNR para la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S. El único o más valores SNR representan una o más de las relaciones de potencia de contenido de voz y contenido de audio de no voz de la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S, o relaciones de potencia de contenido de voz y contenido de audio total de la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S.

En una realización, la combinación cuantitativa específica de la mejora de la voz con forma de onda codificada según la versión del contenido de voz en la representación de canal de audio M/S y mejora de la voz paramétrica según una versión reconstruida del contenido de voz en la representación de canal de audio M/S se determina con un modelo de enmascaramiento auditivo en el cual la mejora de la voz con forma de onda codificada según la versión del contenido de voz en la representación de canal de audio M/S representa una cantidad relativa más grande de mejora de la voz en múltiples combinaciones de mejoras de la voz con forma de onda codificada y mejora de la voz paramétrica que asegura que el ruido de codificación en un programa de audio de voz mejorada de salida no sea audible de manera objetable.

En una realización, al menos una porción de los metadatos de mejora de la voz M/S permite que un decodificador de audio receptor reconstruya una versión del contenido de voz en la representación M/S del contenido de audio mixto en la representación de canal de audio de referencia.

En una realización, los metadatos de mejora de la voz M/S comprenden metadatos relacionados con una o más de funciones de mejora de la voz con forma de onda codificada en la representación de canal de audio M/S, o funciones de mejora de la voz paramétrica en el canal de audio M/S.

En una realización, la representación de canal de audio de referencia comprende canales de audio relacionados con altavoces envolventes. En una realización, el único o más canales no M/S de la representación de canal de audio de referencia comprenden uno o más de un canal central, un canal izquierdo, o un canal derecho, mientras que el único o más canales M/S de la representación de canal de audio M/S comprenden uno o más de un canal medio o un canal lateral.

En una realización, los metadatos de mejora de la voz M/S comprenden un solo conjunto de metadatos de mejora de la voz relacionados con un canal medio de la representación de canal de audio M/S. En una realización, los metadatos de mejora de la voz M/S representan una parte de los metadatos de audio totales codificados en la señal de audio. En una realización, los metadatos de audio codificados en la señal de audio comprenden un campo de datos para indicar una presencia de los metadatos de mejora de la voz M/S. En una realización, la señal de audio es una parte de una señal audiovisual.

En una realización, un aparato que comprende un procesador se configura para llevar a cabo el método de decodificación según se describe en la presente memoria.

En una realización, un medio de almacenamiento legible por ordenador no transitorio comprende instrucciones de software que, cuando se ejecutan por uno o más procesadores, provocan la realización de cualquiera de los métodos de decodificación según se describe en la presente memoria. Es preciso notar que, aunque realizaciones separadas se describen en la presente memoria, cualquier combinación de realizaciones y/o realizaciones parciales descritas en la presente memoria pueden combinarse para formar realizaciones adicionales.

8. Mecanismos de implementación - Resumen de hardware

Según una realización, las técnicas descritas en la presente memoria se implementan por uno o más dispositivos informáticos de propósito especial. Los dispositivos informáticos de propósito especial pueden cablearse para llevar a cabo las técnicas, o pueden incluir dispositivos electrónicos digitales como, por ejemplo, uno o más circuitos integrados para aplicaciones específicas (ASIC, por sus siglas en inglés) o matrices de puertas programables por campo (FPGA, por sus siglas en inglés) que se programan de forma persistente para llevar a cabo las técnicas, o pueden incluir uno o más procesadores de hardware de propósito general programados para llevar a cabo las técnicas según instrucciones de programa en firmware, memoria, otro almacenamiento o una combinación. Dichos dispositivos informáticos de propósito especial pueden también combinar lógica cableada personalizada, ASIC o FPGA con programación personalizada para lograr las técnicas. Los dispositivos informáticos de propósito especial pueden ser sistemas de ordenador de sobremesa, sistemas de ordenador portátil, dispositivos portátiles, dispositivos conectados en red o cualquier otro dispositivo que incorpora lógica cableada y/o de programa para implementar las técnicas.

Por ejemplo, la Figura 9 es un diagrama de bloques que ilustra un sistema informático 900 en el cual una realización de la invención puede implementarse. El sistema informático 900 incluye un bus 902 u otro mecanismo de

comunicación para comunicar información, y un procesador de hardware 904 acoplado con el bus 902 para procesar información. El procesador de hardware 904 puede ser, por ejemplo, un microprocesador de propósito general.

El sistema informático 900 también incluye una memoria principal 906 como, por ejemplo, una memoria de acceso aleatorio (RAM, por sus siglas en inglés) u otro dispositivo de almacenamiento dinámico, acoplado al bus 902 para almacenar información e instrucciones que se ejecutarán por el procesador 904. La memoria principal 906 también puede usarse para almacenar variables temporales u otra información intermedia durante la ejecución de las instrucciones que se ejecutarán por el procesador 904. Dichas instrucciones, cuando se almacenan en un medio de almacenamiento no transitorio accesible para el procesador 904, reproducen el sistema informático 900 en una máquina de propósito especial que es específica para el dispositivo para llevar a cabo las funciones especificadas en las instrucciones.

El sistema informático 900 además incluye una memoria de solo lectura (ROM, por sus siglas en inglés) 908 u otro dispositivo de almacenamiento estático acoplado al bus 902 para almacenar información estática e instrucciones para el procesador 904. Un dispositivo de almacenamiento 910 como, por ejemplo, un disco magnético o disco óptico, se provee y acopla al bus 902 para almacenar información e instrucciones.

El sistema informático 900 puede acoplarse mediante el bus 902 a una visualización 912 como, por ejemplo, una pantalla de cristal líquido (LCD, por sus siglas en inglés), para mostrar información a un usuario de ordenador. Un dispositivo de entrada 914, incluidas claves alfanuméricas y otras, se acopla al bus 902 para comunicar información y selecciones de comando al procesador 904. Otro tipo de dispositivo de entrada de usuario es el control de cursor 916 como, por ejemplo, un ratón, una bola de mando, o teclas de dirección de cursor para comunicar información de dirección y selecciones de comando al procesador 904 y para controlar el movimiento del cursor en la visualización 912. El presente dispositivo de entrada normalmente tiene dos grados de libertad en dos ejes, un primer eje (p.ej., x) y un segundo eje (p.ej., y), que permite al dispositivo especificar posiciones en un plano.

El sistema informático 900 puede implementar las técnicas descritas en la presente memoria mediante el uso de lógica cableada específica para el dispositivo, uno o más ASIC o FPGA, firmware y/o lógica de programa que, en combinación con el sistema informático, programa o hace que el sistema informático 900 sea una máquina de propósito especial. Según una realización, las técnicas en la presente memoria se llevan a cabo por el sistema informático 900 en respuesta a que al procesador 904 ejecuta una o más secuencias de una o más instrucciones contenidas en la memoria principal 906. Dichas instrucciones pueden leerse en la memoria principal 906 desde otro medio de almacenamiento como, por ejemplo, el dispositivo de almacenamiento 910. La ejecución de las secuencias de instrucciones contenidas en la memoria principal 906 hace que el procesador 904 lleve a cabo las etapas del proceso descritas en la presente memoria. En realizaciones alternativas, los circuitos cableados pueden usarse en lugar de o en combinación con instrucciones de software.

El término "medios de almacenamiento" según su uso en la presente memoria se refiere a cualquier medio no transitorio que almacena datos y/o instrucciones que hacen que una máquina funcione en una manera específica. Dichos medios de almacenamiento pueden comprender medios permanentes y/o medios no permanentes. Los medios permanentes incluyen, por ejemplo, discos ópticos o magnéticos como, por ejemplo, el dispositivo de almacenamiento 910. Los medios no permanentes incluyen memoria dinámica como, por ejemplo, la memoria principal 906. Formas comunes de medios de almacenamiento incluyen, por ejemplo, un disco flexible, disco duro, unidad de estado sólido, cinta magnética, o cualquier otro medio de almacenamiento de datos magnético, un CD-ROM, cualquier otro medio de almacenamiento de datos óptico, cualquier medio físico con patrones de orificios, una RAM, una PROM, y EPROM, una FLASH-EPROM, NVRAM, cualquier otro chip o cartucho de memoria.

Los medios de almacenamiento son distintos de, pero pueden usarse en conjunto con medios de transmisión. Los medios de transmisión participan en la transferencia de información entre los medios de almacenamiento. Por ejemplo, los medios de transmisión incluyen cables coaxiales, alambre de cobre y fibras ópticas, incluidos los cables que comprenden el bus 902. Los medios de transmisión pueden también tomar la forma de ondas acústicas o de luz como, por ejemplo, aquellas generadas durante comunicaciones de datos infrarrojos y ondas radioeléctricas.

Varias formas de medios pueden estar implicadas al llevar una o más secuencias de una o más instrucciones al procesador 904 para la ejecución. Por ejemplo, las instrucciones pueden llevarse inicialmente en un disco magnético o unidad de estado sólido de un ordenador remoto. El ordenador remoto puede cargar las instrucciones en su memoria dinámica y enviar las instrucciones en una línea telefónica mediante el uso de un módem. Un módem local para el sistema informático 900 puede recibir los datos en la línea telefónica y usar un transmisor infrarrojo para convertir los datos en una señal infrarroja. Un detector infrarrojo puede recibir los datos transportados en la señal infrarroja y circuitos apropiados pueden colocar los datos en el bus 902. El bus 902 lleva los datos a la memoria principal 906, desde la cual el procesador 904 recupera y ejecuta las instrucciones. Las instrucciones recibidas por la memoria principal 906 pueden almacenarse, de manera opcional, en el dispositivo de almacenamiento 910 antes o después de la ejecución por el procesador 904.

El sistema informático 900 también incluye una interfaz de comunicación 918 acoplada al bus 902. La interfaz de comunicación 918 provee un acoplamiento de comunicación de datos de dos vías a un enlace de red 920 que se

conecta a una red local 922. Por ejemplo, la interfaz de comunicación 918 puede ser una tarjeta de red digital de servicios integrados (ISDN, por sus siglas en inglés), módem de cable, módem de satélite, o un módem para proveer una conexión de comunicación de datos a un tipo correspondiente de línea telefónica. Como otro ejemplo, la interfaz de comunicación 918 puede ser una tarjeta de red de área local (LAN, por sus siglas en inglés) para proveer una conexión de comunicación de datos a una LAN compatible. Enlaces inalámbricos también pueden implementarse. En cualquiera de dichas implementaciones, la interfaz de comunicación 918 envía y recibe señales eléctricas, electromagnéticas u ópticas que llevan trenes de datos digitales que representan varios tipos de información.

El enlace de red 920 normalmente provee una comunicación de datos a través de una o más redes a otros dispositivos de datos. Por ejemplo, el enlace de red 920 puede proveer una conexión a través de la red local 922 a un ordenador anfitrión 924 o a un equipo de datos operado por un Proveedor de Servicios de Internet (ISP, por sus siglas en inglés) 926. El ISP 926, a su vez, provee servicios de comunicación de datos a través de la red de comunicación de datos de paquete mundial a la que ahora se hace referencia comúnmente como "Internet" 928. La red local 922 e Internet 928 usan, ambas, señales eléctricas, electromagnéticas u ópticas que llevan trenes de datos digitales. Las señales a través de las varias redes y las señales en el enlace de red 920 y a través de la interfaz de comunicación 918, que llevan los datos digitales a y desde el sistema informático 900, son formas a modo de ejemplo de medios de transmisión.

El sistema informático 900 puede enviar mensajes y recibir datos, incluido el código de programa, a través de la red, enlace de red 920 e interfaz de comunicación 918. En el ejemplo de Internet, un servidor 930 puede transmitir un código solicitado para un programa de aplicación a través de Internet 928, ISP 926, red local 922 e interfaz de comunicación 918.

El código recibido puede ejecutarse por el procesador 904 cuando se recibe, y/o almacenarse en el dispositivo de almacenamiento 910, u otro almacenamiento permanente para su posterior ejecución.

9. Equivalentes, extensiones, alternativas y varios

En la anterior memoria descriptiva, las realizaciones de la invención se han descrito con referencia a numerosos detalles específicos que pueden variar de implementación a implementación. Por consiguiente, el único indicador exclusivo de qué es la invención, y qué se pretende por los solicitantes que sea la invención, es el conjunto de reivindicaciones que se emiten a partir de la presente solicitud, en la forma específica en la cual dichas reivindicaciones se emiten, incluida cualquier corrección subsiguiente. Cualquier definición expresamente establecida en la presente memoria para términos contenidos en dichas reivindicaciones regirá el significado de dichos términos según su uso en las reivindicaciones. Por lo tanto, ninguna limitación, elemento, característica, ventaja o atributo que no se incluya expresamente en una reivindicación debe limitar el alance de dicha reivindicación de manera alguna. La memoria descriptiva y los dibujos se considerarán, por consiguiente, en un sentido ilustrativo antes que en uno restrictivo.

REIVINDICACIONES

1. Un método en un decodificador de audio, que comprende:

recibir una señal de audio que comprende contenido de audio mixto en una representación de canal de audio de referencia y metadatos para la mejora de la voz de uno o más canales de audio en una representación de canal de audio Medio/Lateral, M/S, el contenido de audio mixto teniendo una mezcla de contenido de voz y contenido de audio de no voz;

transformar una o más porciones del contenido de audio mixto que se esparcen en dos o más canales no M/S en múltiples canales de audio de la representación de canal de audio de referencia en una o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S que se esparcen en uno o más canales M/S de la representación de canal de audio M/S, en donde la representación de canal de audio M/S comprende al menos un canal medio y un canal lateral, en donde el canal medio representa una suma ponderada o no ponderada de dos canales no M/S de la representación de canal de audio de referencia, y en donde el canal lateral representa una diferencia ponderada o no ponderada de los mismos dos canales no M/S de la representación de canal de audio de referencia;

llevar a cabo una o más funciones de mejora de la voz, según los metadatos para la mejora de la voz, en la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S para generar una o más porciones de contenido de voz mejorada en la representación M/S;

combinar la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S con la única o más porciones de contenido de voz mejorada en la representación M/S para generar una o más porciones de contenido de audio mixto de voz mejorada en la representación M/S;

en donde el método se lleva a cabo por uno o más dispositivos informáticos.

2. El método de la reivindicación 1, que además comprende:

extraer una versión del contenido de voz, en la representación de canal de audio M/S, de manera independiente del contenido de audio mixto de la señal de audio; llevar a cabo una o más funciones de mejora de la voz, según los metadatos para la mejora de la voz, en una o más porciones de la versión del contenido de voz en la representación de canal de audio M/S para generar una o más segundas porciones de contenido de voz mejorada en la representación de canal de audio M/S;

determinar datos que indican mezcla para la mejora de la voz; y

generar, según los datos que indican mezcla para la mejora de la voz, una combinación cuantitativa específica de mejora de la voz con forma de onda codificada según la versión del contenido de voz en la representación de canal de audio M/S y mejora de la voz paramétrica según una versión reconstruida del contenido de voz en la representación de canal de audio M/S,

en donde la mejora de la voz con forma de onda codificada según la versión del contenido de voz en la representación de canal de audio M/S comprende añadir la versión del contenido de voz, en la representación de canal de audio M/S, a la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S, y

en donde la mejora de la voz paramétrica según una versión reconstruida del contenido de voz en la representación de canal de audio M/S comprende añadir la versión reconstruida del contenido de voz en la representación de canal de audio M/S a la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S.

3. El método de la reivindicación 2, en donde los datos que indican mezcla se generan, por uno de un codificador de audio corriente arriba que genera la señal de audio o dicho decodificador de audio que recibe la señal de audio, según al menos en parte uno o más valores SNR para la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S, en donde el único o más valores SNR representan una o más de relaciones de potencia de contenido de voz y contenido de audio de no voz de la única o más porciones de contenido de audio mixto transformado en la representación de canal de audio M/S, o relaciones de potencia de contenido de voz y contenido de audio total de la única o más porciones de uno del contenido de audio mixto transformado en la representación de canal de audio M/S o contenido de audio mixto en una representación de canal de audio de referencia.

4. El método de cualquiera de las reivindicaciones 2-3, en donde la combinación cuantitativa específica de la mejora de la voz con forma de onda codificada según la versión del contenido de voz en la representación de canal de audio M/S y mejora de la voz paramétrica según una versión reconstruida del contenido de voz en la representación de canal de audio M/S se determina con un modelo de enmascaramiento auditivo, según se construye por uno de un

codificador de audio corriente arriba que genera la señal de audio o dicho decodificador de audio que recibe la señal de audio, en la cual la mejora de la voz con forma de onda codificada según la versión del contenido de voz en la representación de canal de audio M/S representa una cantidad relativa más grande de mejora de la voz en múltiples combinaciones de mejora de la voz con forma de onda codificada y la mejora de la voz paramétrica que asegura que el ruido de codificación en un programa de audio de voz mejorada de salida no sea audible de manera objetable.

5. El método de cualquiera de las reivindicaciones 2-4, en donde los metadatos para la mejora de la voz comprenden metadatos relacionados con una o más de las funciones de mejora de la voz con forma de onda codificada en la representación de canal de audio M/S, o las funciones de mejora de la voz paramétrica en la representación de canal de audio M/S.

6. El método de cualquiera de las reivindicaciones 1-5, en donde los metadatos para la mejora de la voz comprenden un solo conjunto de metadatos de mejora de la voz relacionados con el canal medio de la representación de canal de audio M/S.

7. Un aparato que comprende un procesador y que se configura para llevar a cabo cualquiera de los métodos incluidos en las reivindicaciones 1-6.

8. Un medio de almacenamiento legible por ordenador no transitorio, que comprende instrucciones de software que, cuando se ejecutan por uno o más procesadores, hacen que los procesadores lleven a cabo cualquiera de los métodos incluidos en las reivindicaciones 1-6.

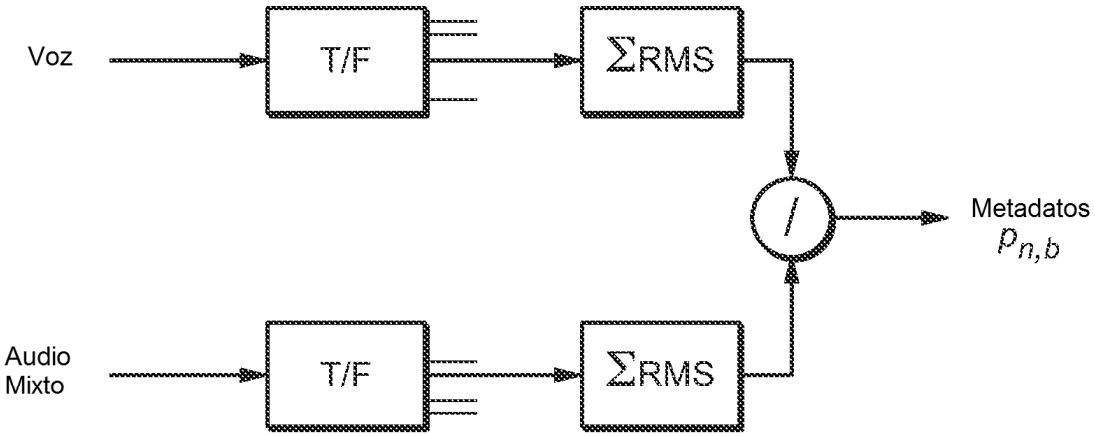
Voz
T/F
ΣRMS
/
Metadatos
$p_{n,b}$
Audio
Mixto
T/F
ΣRMS

FIG. 1

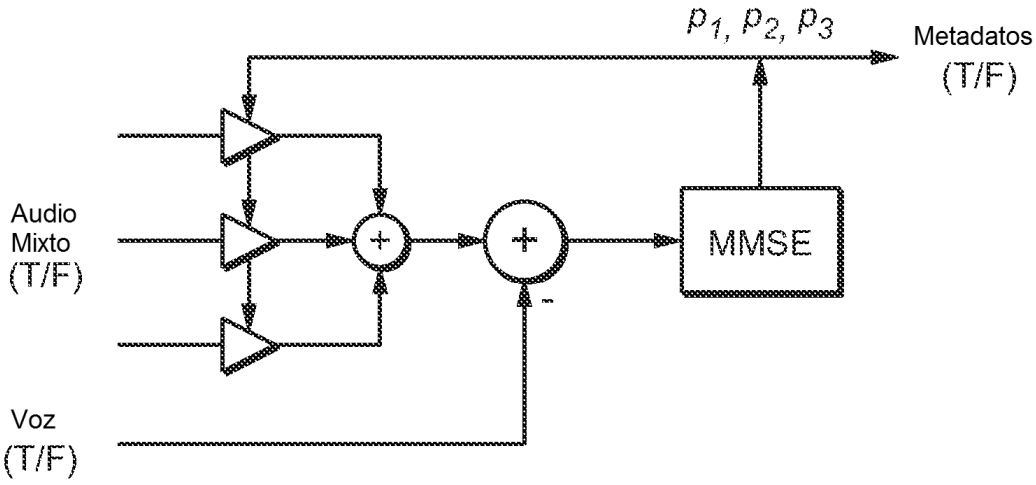
p_1, p_2, p_3
Metadatos
(T/F)
Audio
Mixto
(T/F)
+
+
-
MMSE
Voz
(T/F)

FIG. 2

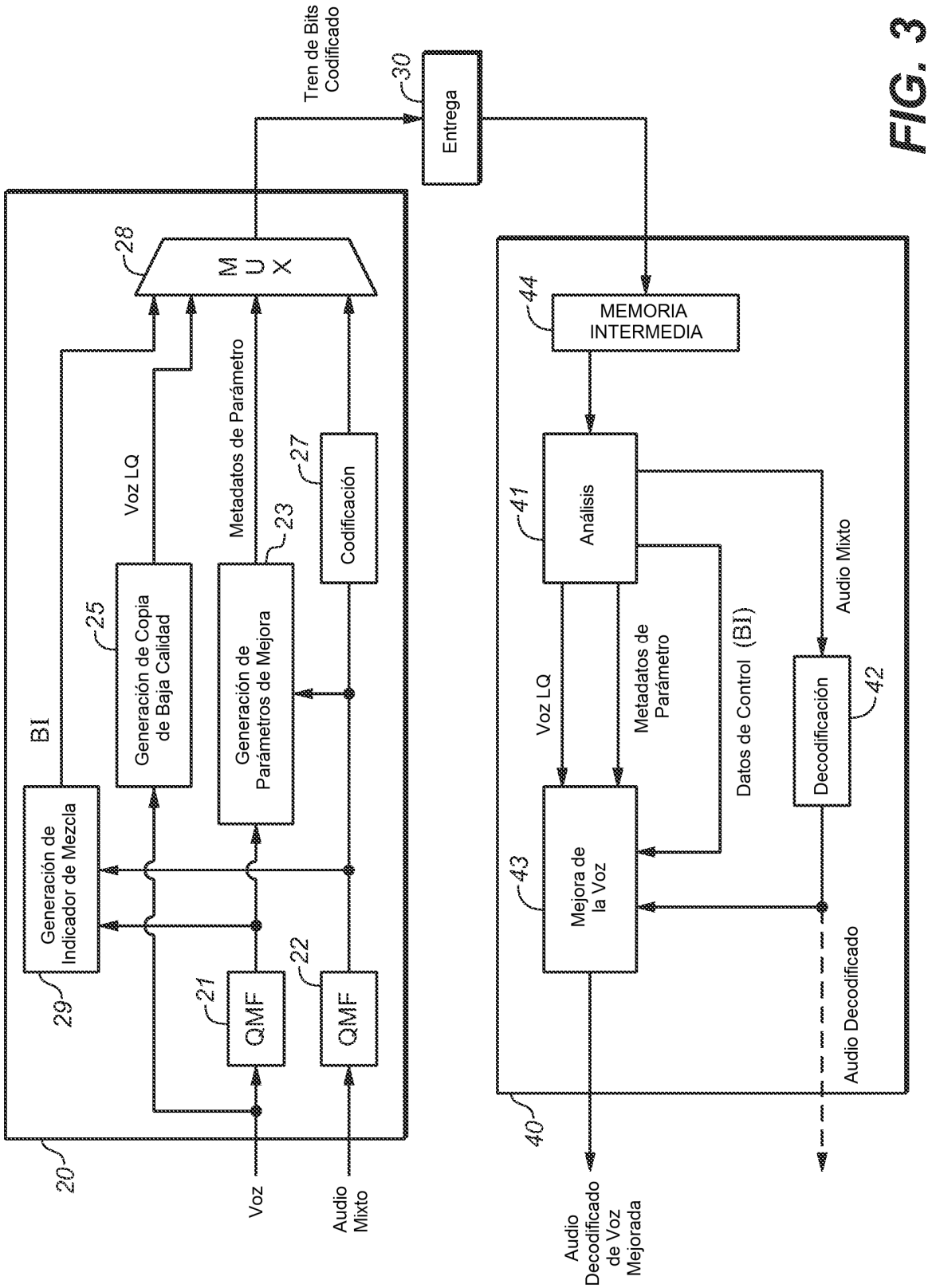
Tren de Bits Codificado
30
Entrega
28
MUX
Voz LQ
Metadatos de Parámetro
27
Codificación
23
Generación de Parámetros de Mejora
25
Generación de Copia de Baja Calidad
BI
Generación de Indicador de Mezcla
29
21
QMF
22
QMF
20
Voz
Audio Mixto
44
MEMORIA INTERMEDIA
41
Análisis
Voz LQ
Metadatos de Parámetro
Datos de Control (BI)
Audio Mixto
42
Decodificación
43
Mejora de la Voz
40
Audio Decodificado
Audio Decodificado de Voz Mejorada

FIG. 3

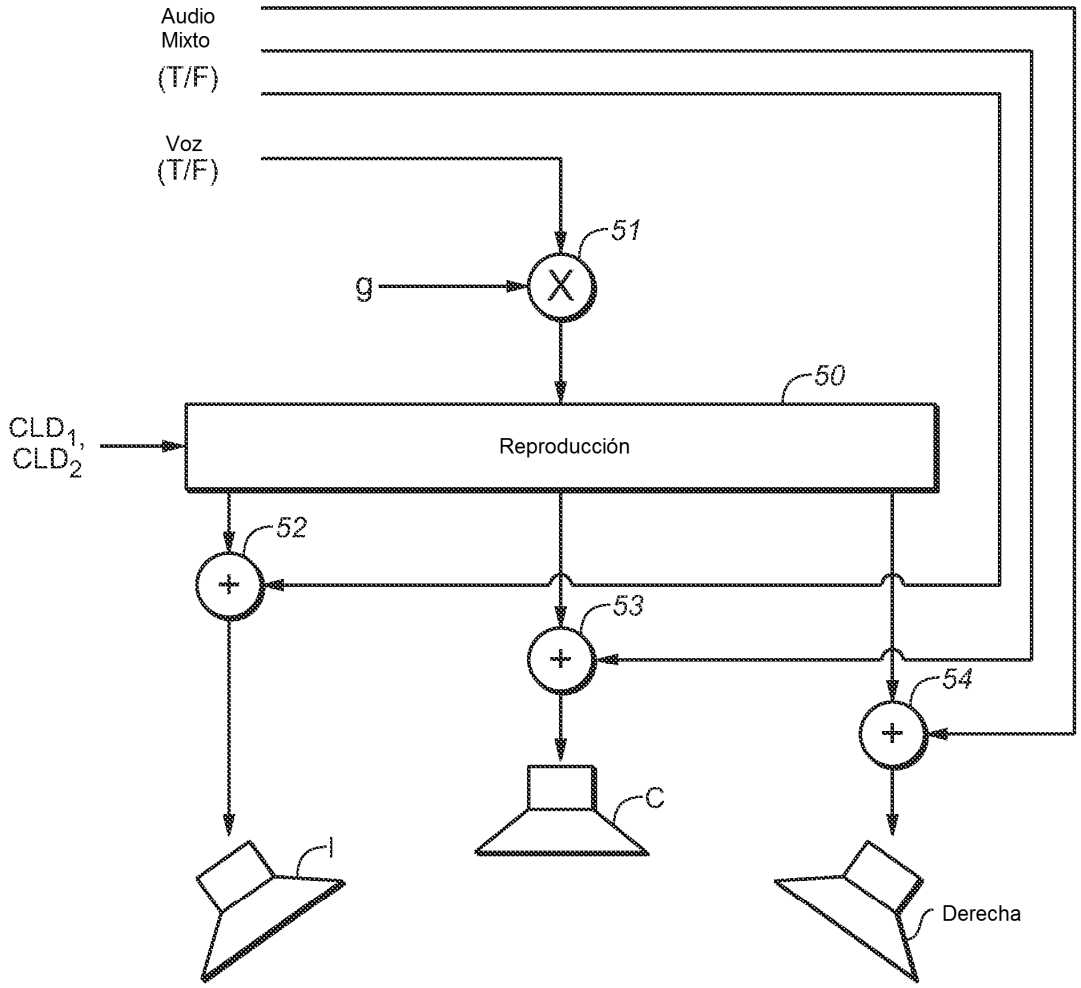
Audio
Mixto
(T/F)
Voz
(T/F)
51
g
X
50
CLD1,
CLD2
Reproducción
52
53
54
C
I
Derecha

FIG. 4

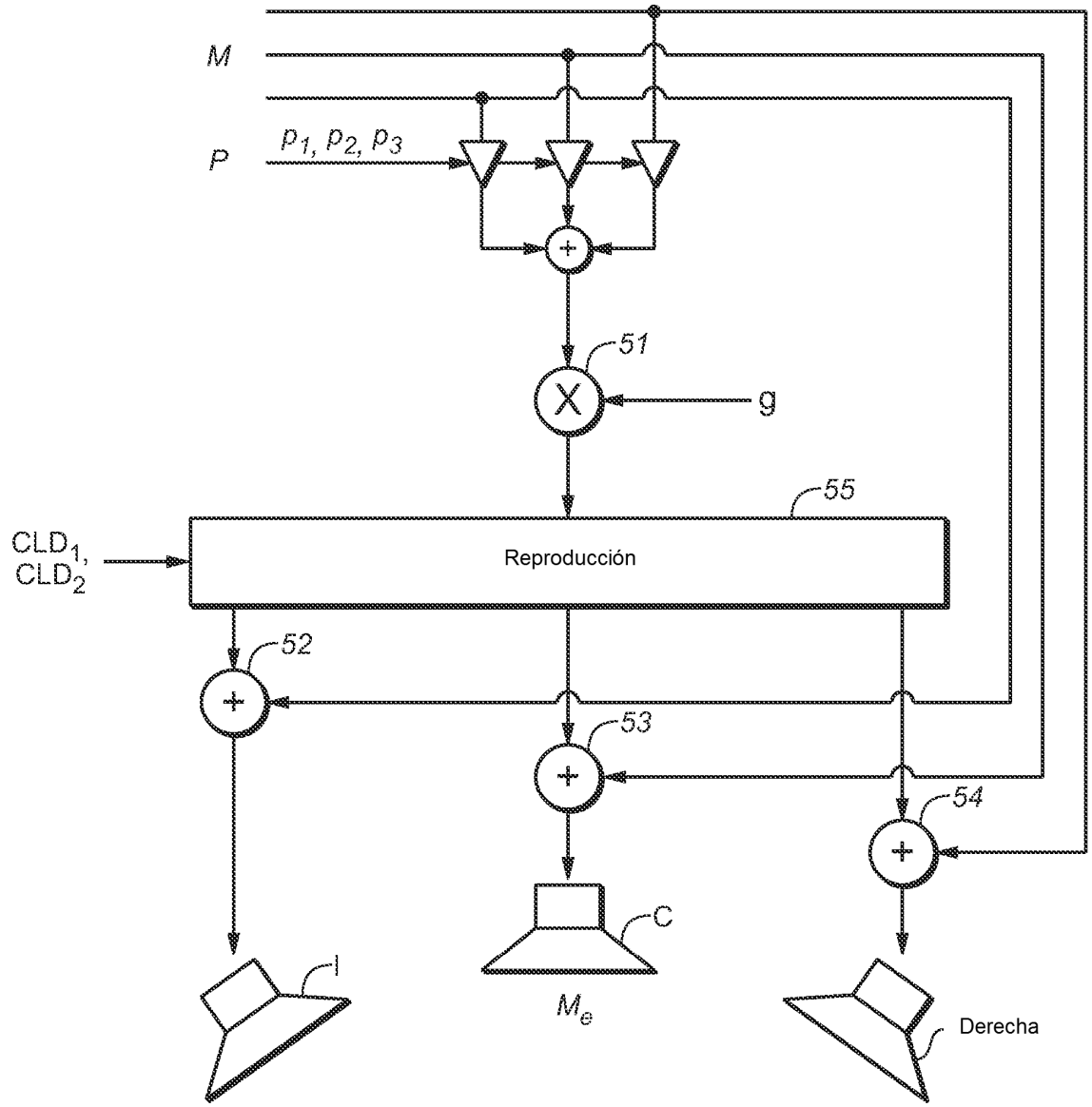
M
P
p_1, p_2, p_3
51
g
55
CLD_1, CLD_2
Reproducción
52
53
54
C
M_e
I
Derecha

FIG. 5

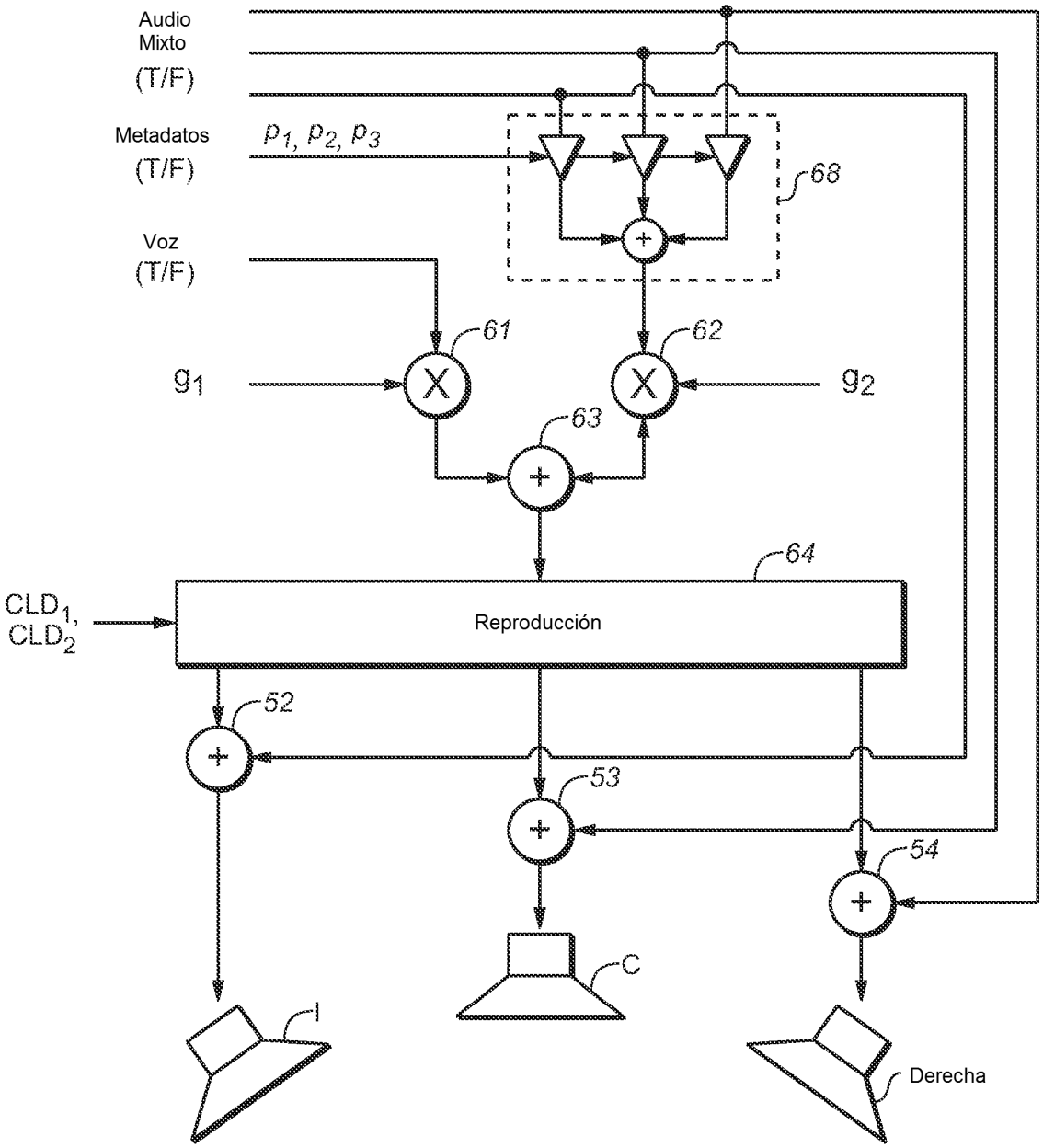
Audio
Mixto
(T/F)
Metadatos
(T/F)
p_1, p_2, p_3
68
Voz
(T/F)
61
62
g_1
g_2
63
64
CLD_1,
CLD_2
Reproducción
52
53
54
C
I
Derecha

FIG. 6

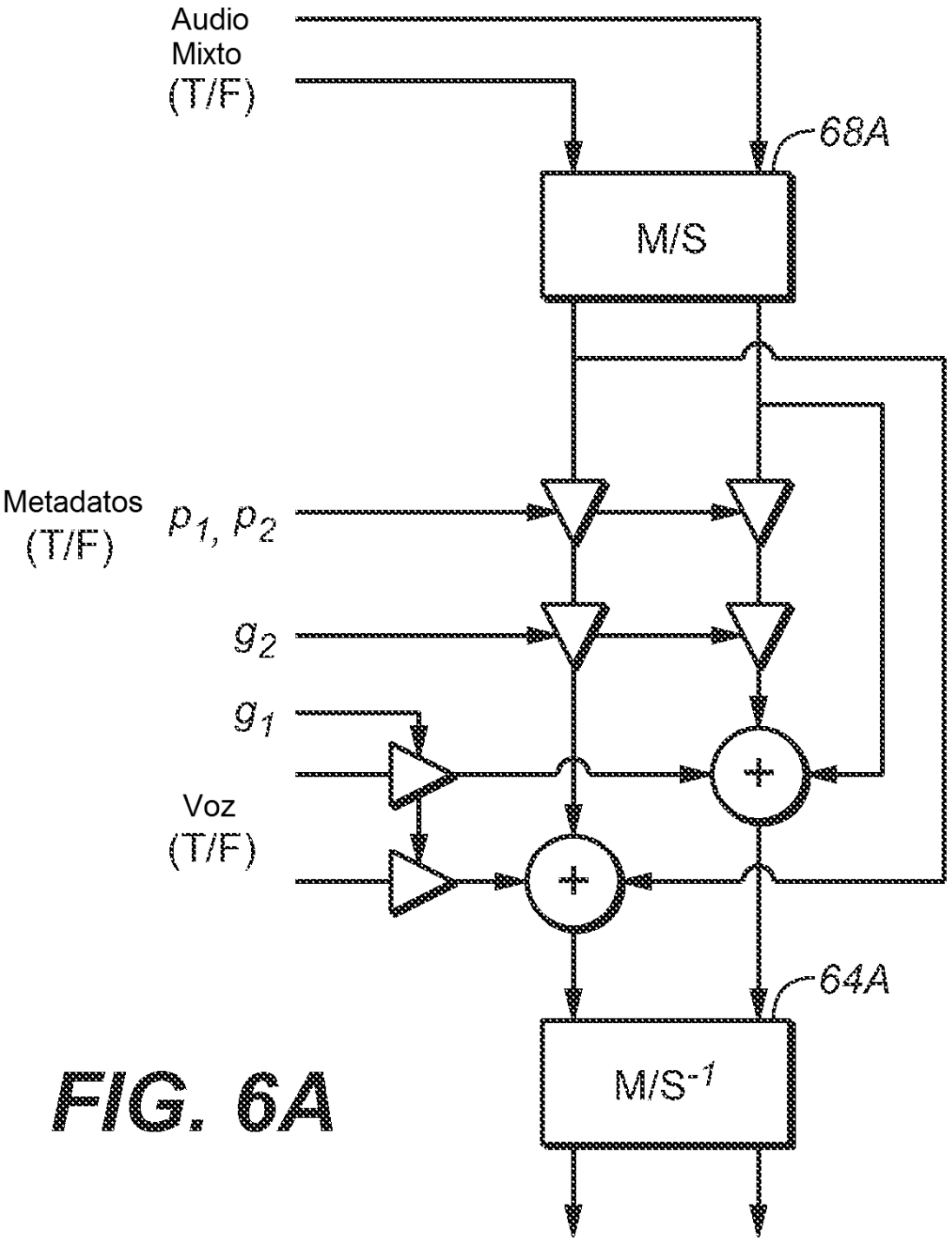

FIG. 6A

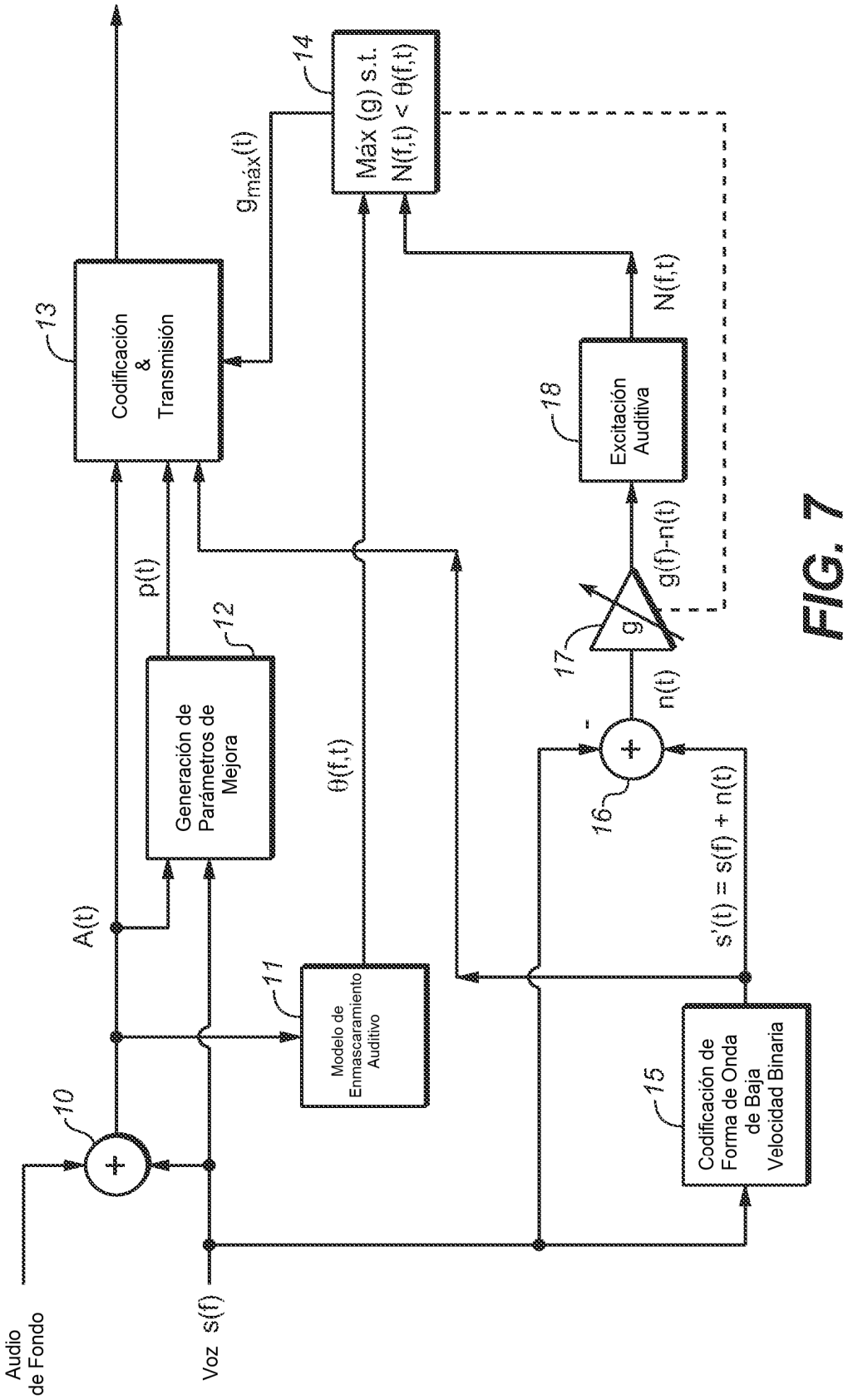
Audio de Fondo
Voz s(f)
10
A(t)
p(t)
13
Codificación & Transmisión
12
Generación de Parámetros de Mejora
11
Modelo de Enmascaramiento Auditivo
θ(f,t)
g_máx(t)
14
Máx (g) s.t. N(f,t) < θ(f,t)
s'(t) = s(f) + n(t)
15
Codificación de Forma de Onda de Baja Velocidad Binaria
16
n(t)
17
g
g(f)·n(t)
18
Excitación Auditiva
N(f,t)

FIG. 7

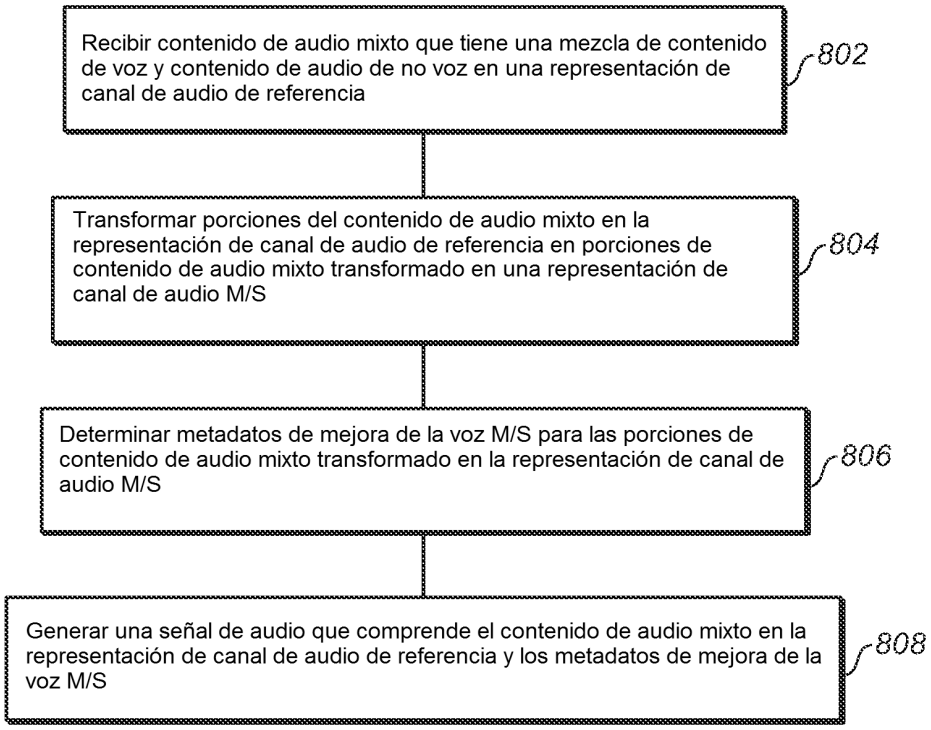
Recibir contenido de audio mixto que tiene una mezcla de contenido de voz y contenido de audio de no voz en una representación de canal de audio de referencia
802
Transformar porciones del contenido de audio mixto en la representación de canal de audio de referencia en porciones de contenido de audio mixto transformado en una representación de canal de audio M/S
804
Determinar metadatos de mejora de la voz M/S para las porciones de contenido de audio mixto transformado en la representación de canal de audio M/S
806
Generar una señal de audio que comprende el contenido de audio mixto en la representación de canal de audio de referencia y los metadatos de mejora de la voz M/S
808

FIG. 8A

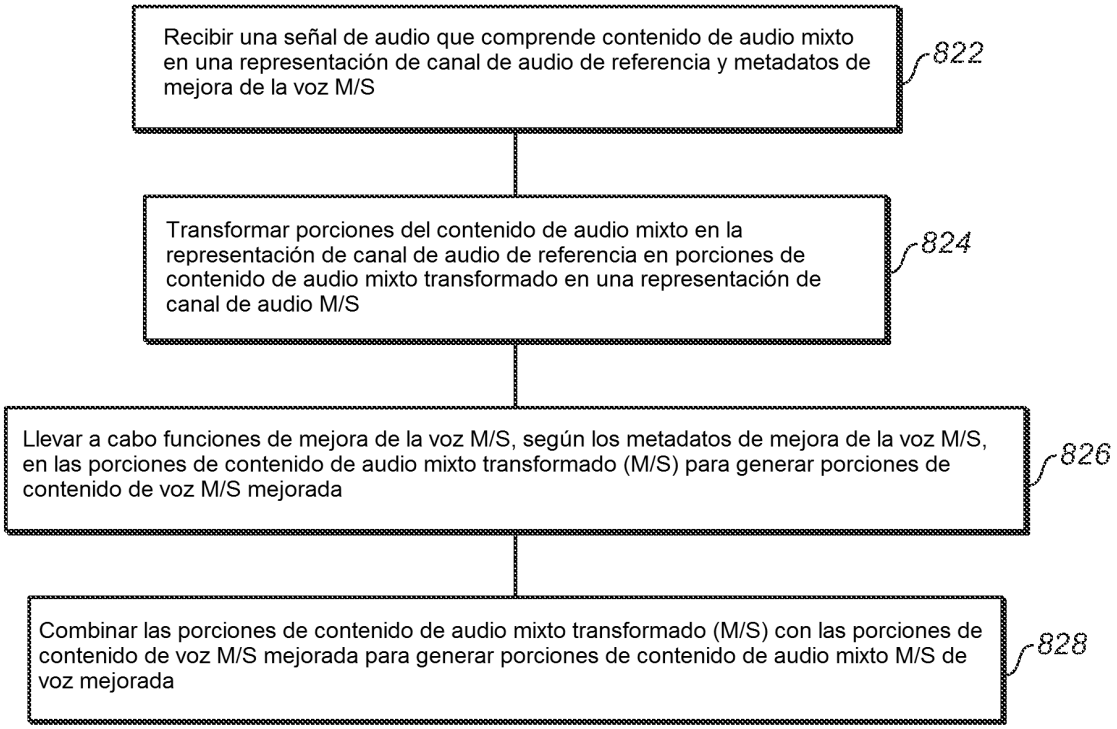
Recibir una señal de audio que comprende contenido de audio mixto en una representación de canal de audio de referencia y metadatos de mejora de la voz M/S
822
Transformar porciones del contenido de audio mixto en la representación de canal de audio de referencia en porciones de contenido de audio mixto transformado en una representación de canal de audio M/S
824
Llevar a cabo funciones de mejora de la voz M/S, según los metadatos de mejora de la voz M/S, en las porciones de contenido de audio mixto transformado (M/S) para generar porciones de contenido de voz M/S mejorada
826
Combinar las porciones de contenido de audio mixto transformado (M/S) con las porciones de contenido de voz M/S mejorada para generar porciones de contenido de audio mixto M/S de voz mejorada
828

FIG. 8B

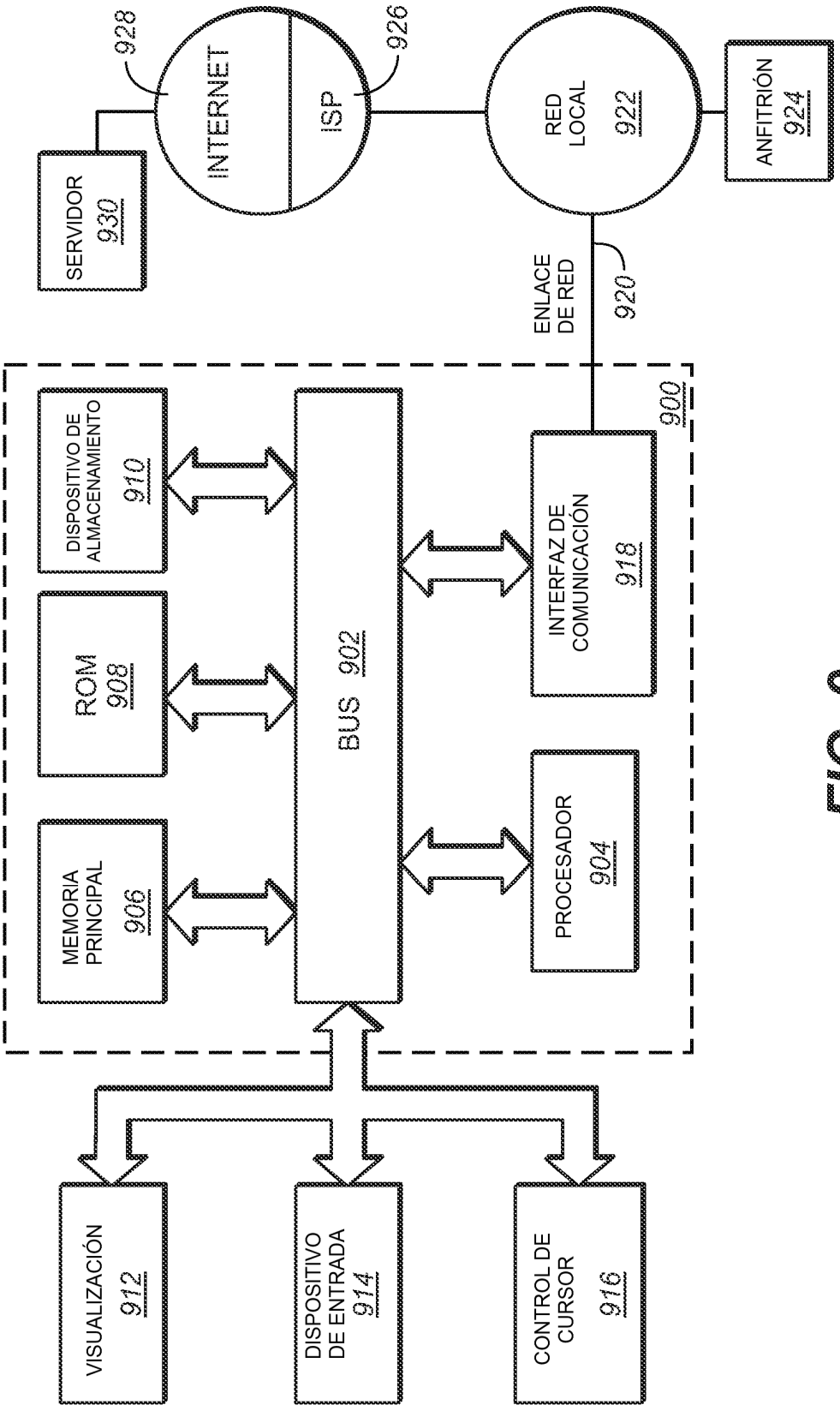
928
INTERNET
ISP
926
RED LOCAL 922
ANFITRIÓN 924
SERVIDOR 930
ENLACE DE RED
920
900
DISPOSITIVO DE ALMACENAMIENTO 910
ROM 908
MEMORIA PRINCIPAL 906
BUS 902
INTERFAZ DE COMUNICACIÓN 918
PROCESADOR 904
VISUALIZACIÓN 912
DISPOSITIVO DE ENTRADA 914
CONTROL DE CURSOR 916

FIG. 9