(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 630 398**

(51) Int. CI.:

H03G 5/00 (2006.01)
H03G 5/16 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **17.03.2014 PCT/US2014/030663**

(87) Fecha y número de publicación internacional: **02.10.2014 WO14160548**

(96) Fecha de presentación y número de la solicitud europea: **17.03.2014 E 14724216 (8)**

(97) Fecha y número de publicación de la concesión europea: **03.05.2017 EP 2979359**

(54) Título: **Dispositivo de control y método de control del ecualizador**

(30) Prioridad:

26.03.2013 CN 201310100401
11.04.2013 US 201361811058 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
21.08.2017

(73) Titular/es:

DOLBY LABORATORIES LICENSING CORPORATION (100.0%)
100 Potrero Avenue
San Francisco, CA 94103-4813, US

(72) Inventor/es:

LU, LIE;
WANG, JUN;
SEEFELDT, ALAN y
HU, MINGQING

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

DESCRIPCIÓN

Dispositivo de control y método de control del ecualizador

Referencia cruzada a solicitudes relacionadas

Esta solicitud reivindica la prioridad para la solicitud de patente china nº 201310100401.X, presentada con fecha 26 de marzo de 2013 y la solicitud de patente provisional de Estados Unidos nº 61/811,058, presentada el 11 de abril de 2013.

CAMPO TÉCNICO

La presente solicitud de patente se refiere, en general, a un procesamiento de señales de audio. Más concretamente, las formas de realización de la presente invención se refieren a aparatos y métodos para la clasificación de audio y su procesamiento, en particular, el control de un dispositivo de mejora de diálogos, virtualizador de sonido envolvente, 1nivelador de volumen y ecualizador.

ANTECEDENTES DE LA INVENCIÓN

Algunos dispositivos de mejora del audio tienden a modificar las señales de audio, en un dominio temporal o en un dominio espectral, con el fin de mejorar la calidad global del audio y mejorar la experiencia de los usuarios, en correspondencia. Varios dispositivos de mejora de audio han sido desarrollados para varios fines. Algunos ejemplos típicos de dispositivos de mejora de audio incluyen:

Dispositivo de mejora del diálogo: El diálogo es la componente más importante en una película cinematográfica y programas de radio o TV para entender la narración. Se desarrollaron métodos para mejorar los diálogos con el fin de aumentar su claridad y su inteligibilidad, en particular, para las personas de edad avanzada con capacidad auditiva disminuida.

Virtualizador de sonido envolvente: Un virtualizador de sonido envolvente permite a una señal sonora envolvente (multi-canal) presentarse a través de los altavoces internos del PC o a través de auriculares. Es decir, con el dispositivo de estéreo (tales como altavoces y auriculares), crea un efecto virtualmente envolvente y proporciona una experiencia cinemática para los usuarios.

Nivelador de volumen: Un nivelador de volumen tiene como objetivo ajustar el volumen del contenido de audio en la reproducción y mantenerlo casi constante durante la escala temporal sobre la base de un valor de intensidad objetivo.

Ecualizador: Un ecualizador proporciona coherencia del equilibrio espectral, tal como se conoce como "tono" o "timbre", y permite a los usuarios configurar el perfil global (curva o forma) de la respuesta de frecuencia (ganancia) en cada banda de frecuencia individual, con el fin de resaltar algunos sonidos o eliminar sonidos indeseados. En un ecualizador tradicional, pueden proporcionarse diferentes preajustes del ecualizador para diferentes sonidos, tales como géneros musicales distintos. Una vez que se seleccione un preajuste, o se establezca un perfil de ecualización, las mismas ganancias de ecualización se aplicarán sobre la señal, hasta que el perfil de ecualización se modifique manualmente. Por el contrario, un ecualizador dinámico consigue la coherencia del equilibrio espectral controlando continuamente el equilibrio espectral del audio, comparándolo con un tono deseado y ajustando dinámicamente un filtro de ecualización para transformar el tono original de audio en el tono deseado.

En general, un dispositivo de mejora de audio tiene su propio escenario/contexto de aplicable. Es decir, un dispositivo de mejora de audio puede ser adecuado para solamente un determinado conjunto de contenidos pero no para todas las señales de audio posibles, puesto que diferentes contenidos pueden necesitar procesarse en formas distintas. A modo de ejemplo, un método de mejora de diálogos se suele aplicar en un contenido de película cinematográfica. Si se aplica sobre música en la que no existen diálogos, puede reforzar falsamente algunas sub-bandas de frecuencia e introducir un fuerte cambio de timbre y una incoherencia perceptual. De modo similar, si un método de supresión de ruido, se aplica sobre señales musicales, serán audibles fuertes artefactos acústicos.

Sin embargo, para un sistema de procesamiento de audio que suele comprender un conjunto de dispositivos de mejora de audio, su entrada podría ser, inevitablemente, la totalidad de los posibles tipos de señales de audio. A modo de ejemplo, un sistema de procesamiento de audio, integrado en un PC, recibirá contenido de audio procedente de una diversidad de fuentes, incluyendo películas cinematográficas, música, VoIP y juego. De este modo, la identificación o la diferenciación del contenido que se procesa adquiere mayor importancia, con el fin de aplicar los mejores algoritmos o mejores parámetros de cada algoritmo en el contenido correspondiente.

Con el fin de diferenciar el contenido de audio y aplicar mejores parámetros o mejores algoritmos de mejora del audio en correspondencia, los sistemas tradicionales suelen diseñar previamente un conjunto de preajustes, y se pide a los usuarios que elijan un preajuste para el contenido que se reproduce. Un preajuste suele codificar un conjunto de algoritmos de mejora de audio y/o sus mejores parámetros que se aplicarán, tal como un preajuste de 'película cinematográfica' y un preajuste de 'música' que está específicamente diseñado para reproducción de música o películas

cinematográficas.

Sin embargo, la selección manual es incómoda para los usuarios. Los usuarios no suelen conmutar entre los preajustes predefinidos sino que se limitan a mantener la utilización de un preajuste para todo el contenido. Además, incluso en algunas soluciones automáticas, los parámetros o algoritmos establecidos en los preajustes suelen ser discretos (tal como activación o desactivación para un algoritmo específico con respecto a un contenido específico), no pudiendo ajustar los parámetros en una manera continua basada en el contenido.

La Publicación de Patente Internacional número WO 2008/106036 A2 se refiere al procesamiento de señales de audio. Más concretamente, se refiere a la mejora de los sistemas de audio de entretenimiento tales como audio televisivo, para mejorar la calidad e inteligibilidad de la voz, tal como las señales de audio narrativo y de diálogos.
La Solicitud de Patente del Reino Unido publicada bajo el número GB 2 491 002 se refiere a un sistema de audio de consumo. El sistema de audio de consumo comprende un procesador de señal acoplado para la recepción de una señal de audio, en donde el contenido dinámico de la señal de audio controla la operación del procesador de señal. La señal de audio es muestreada en una pluralidad de tramas. Las tramas de audio muestreadas se separan en sub-tramas en función el tipo o contenido de frecuencia de la fuente generadora del sonido. Un procesador del dominio temporal genera parámetros de dominio temporal a partir de las sub-tramas separadas. Un procesador del dominio de la frecuencia genera parámetros del dominio de la frecuencia a partir de las sub-tramas separadas. El procesador del dominio temporal o el procesador del dominio de la frecuencia puede detectar la iniciación de una nota de la señal de audio. Una base de datos de signaturas tiene registros de signaturas que tienen, cada uno de ellos, parámetros del dominio temporal y parámetros del dominio de la frecuencia así como parámetros de control. Un detector de reconocimiento hace coincidir los parámetros del dominio temporal y los parámetros del dominio de la frecuencia de las sub-tramas separadas con respecto a un registro de signaturas de la base de datos de signaturas. Los parámetros de control del registro de signaturas de coincidencia controlan el funcionamiento del procesador de señal.
La Publicación de Patente Internacional número WO 2005/106843 A1 se refiere a un método y sistema para clasificar el contenido de audio en función del estilo o género musical. En un aspecto, se da a conocer un método y sistema para ajustar parámetros de un sistema de audio de conformidad con la clasificación del estilo o género musical.

SUMARIO DE LA INVENCIÓN

El primer aspecto de la presente solicitud de patente es configurar automáticamente dispositivos de mejora de audio de una manera continua sobre la base del contenido del audio en la reproducción. Con este modo “automático”, los usuarios pueden simplemente disfrutar su contenido sin molestarse en seleccionar diferentes preajustes. Por otro lado, la sintonía continua es más importante con el fin de evitar la presencia de artefactos audibles en los puntos de transición. Varias formas de realización dan a conocer un controlador de ecualizador, un método de control de ecualizador y un soporte legible por ordenador según se estipula en las reivindicaciones 1, 2, 15, 16 y 17.
El segundo aspecto de la presente solicitud es desarrollar una componente de identificación del contenido para identificar múltiples tipos de audio y los resultados detectados pueden utilizarse para dirigir/guiar los comportamientos de varios dispositivos de mejora de audio, buscando los mejores parámetros de una manera continua.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

La presente solicitud de patente se ilustra a modo de ejemplo, y no a modo de limitación, en las figuras de los dibujos adjuntos, en donde las referencias numéricas similares se refieren a elementos similares y en donde:

La Figura 1 es un diagrama que ilustra un aparato de procesamiento de audio en conformidad con una forma de realización de la solicitud de patente;

Las Figuras 2 y 3 son diagramas que ilustran variantes de la forma de realización según se representa en la Figura 1;

Las Figuras 4 a 6 son diagramas que ilustran una posible arquitectura de clasificadores para identificar múltiples tipos de audio y el cálculo del valor de confianza;

Las Figuras 7 a 9 son diagramas que ilustran más formas de realización del aparato de procesamiento de audio de la presente solicitud de patente;

La Figura 10 es un diagrama que ilustra un retardo de transición entre diferentes tipos de audio;

Las Figuras 11 a 14 son diagramas de flujo que ilustran un método de procesamiento de audio en conformidad con formas de realización de la presente solicitud de patente;

La Figura 15 es un diagrama que ilustra un controlador de mejora de diálogos en conformidad con una forma de realización de la presente solicitud de patente;

Las Figuras 16 y 17 son diagramas de flujo que ilustran el uso del método de procesamiento de audio en conformidad con la presente solicitud de patente en el control de un dispositivo de mejora de diálogos;

La Figura 18 es un diagrama que ilustra un controlador de virtualizador de sonido envolvente en conformidad con una forma de realización de la presente solicitud de patente;

La Figura 19 es un diagrama de flujo que ilustra el uso del método de procesamiento de audio en conformidad con la presente solicitud de patente en la función de control de un virtualizador de sonido envolvente;

La Figura 20 es un diagrama que ilustra un controlador de nivelador de volumen en conformidad con una forma de realización de la presente solicitud de patente;

La Figura 21 es un diagrama que ilustra el efecto del controlador de nivelador de volumen en conformidad con la presente solicitud de patente;

La Figura 22 es un diagrama que ilustra un controlador de ecualizador en conformidad con una forma de realización de la presente solicitud de patente;

La Figura 23 ilustra varios ejemplos de ajustes de equilibrio espectral deseados;

La Figura 24 es un diagrama que ilustra un clasificador de audio en conformidad con una forma de realización de la presente solicitud de patente,

Las Figuras 25 y 26 son diagramas que ilustran algunas características a utilizarse por el clasificador de audio de la presente solicitud de patente;

Las Figuras 27 a 29 son diagramas que ilustran algunas formas de realización del clasificador de audio en conformidad con la presente solicitud de patente;

Las Figuras 30 a 33 son diagramas de flujo que ilustran un método de clasificación de audio en conformidad con las formas de realización de la presente solicitud de patente;

La Figura 34 es un diagrama que ilustra un clasificador de audio en conformidad con otra forma de realización de la presente solicitud de patente;

La Figura 35 es un diagrama que ilustra un clasificador de audio en conformidad con otra forma de realización de la presente solicitud de patente;

La Figura 36 es un diagrama que ilustra reglas heurísticas utilizadas en el clasificador de audio de la presente solicitud de patente;

Las Figuras 37 y 38 son diagramas que ilustran más formas de realización del clasificador de audio en conformidad con la presente solicitud de patente;

Las Figuras 39 y 40 son diagramas de flujo que ilustran un método de clasificación de audio en conformidad con formas de realización de la presente solicitud de patente;

La Figura 41 es un diagrama de bloques que ilustra un sistema a modo de ejemplo para poner en práctica las formas de realización de la presente solicitud de patente.

DESCRIPCIÓN DETALLADA DE LAS FORMAS DE REALIZACIÓN

Las formas de realización de la presente solicitud de patente se describen a continuación haciendo referencia a los dibujos adjuntos. Conviene señalar que, para fines de mayor claridad, las representaciones y descripciones sobre dichos componentes y procesos conocidos por los expertos en esta técnica, pero no necesarias para entender la presente solicitud de patente, se omiten en los dibujos y en la descripción.

Como se apreciará por un experto en esta técnica, aspectos de la presente idea inventiva pueden materializarse como un sistema, un dispositivo (p.ej., un teléfono celular, un reproductor multimedia portátil, un ordenador personal, un servidor, una caja decodificadora de televisión o una grabadora de vídeo digital o cualquier otro reproductor multimedia), un método o un producto de programa informático. En consecuencia, aspectos de la presente idea inventiva pueden adoptar la forma de un modo de realización de hardware, un modo de realización de software (incluyendo firmware, software residente, microcódigos, etc.) o una forma de realización que combine aspectos de software y de hardware que pueden referirse, en general, en esta descripción como un "circuito", "módulo" o "sistema". Además, aspectos de la presente idea inventiva pueden adoptar la forma de un producto de programa informático materializado en uno o más soportes legibles por ordenador que tienen incorporado un código de programa legible por ordenador.

Cualquier combinación de uno o más soportes legibles por ordenador pueden utilizarse. El soporte legible por ordenador

puede ser un soporte de señal legible por ordenador o un soporte de memorización legible por ordenador. Un soporte de memorización legible por ordenador puede ser, a modo de ejemplo, pero sin limitación, un sistema, aparato o dispositivo electrónico, magnético, óptico, electromagnético, de infrarrojos o de semiconductores o cualquier combinación adecuada de lo que antecede. Más ejemplos específicos (una lista no exhaustiva) de los soportes de memorización legibles por ordenador incluirán lo que sigue: una conexión eléctrica que tenga uno o más hilos de conexión, un disquete informático portátil, un disco duro, una memoria de acceso directorio (RAM), una memoria de solamente lectura (ROM), una memoria de solamente lectura programable y borrable (EPROM o memoria instantánea), una fibra óptica, una memoria de solamente lectura de disco compacto portátil (CD-ROM), un dispositivo de almacenamiento óptico, un dispositivo de almacenamiento magnético o cualquier combinación adecuada de lo que antecede. En el contexto de este documento, un soporte de memorización legible por ordenador puede ser cualquier soporte tangible que pueda contener o memorizar un programa para uso por o en relación con un sistema, aparato o dispositivo de ejecución de instrucciones.

Un soporte de señal legible por ordenador puede incluir una señal de datos propagada con un código de programa legible por ordenador incorporado, a modo de ejemplo, en la banda base o como parte de una onda portadora. Dicha señal propagada puede adoptar cualquiera de una diversidad de formas, incluyendo, sin limitación, la forma de una señal electromagnética u óptica, o cualquiera de sus combinaciones adecuadas.

Un soporte de señal legible por ordenador puede ser cualquier soporte legible por ordenador que no sea un soporte de memorización legible por ordenador y que pueda comunicar, propagar o transportar un programa para su uso por o en relación con un sistema, aparato o dispositivo de ejecución de instrucciones.

Un código de programa incorporado en un soporte legible por ordenador puede transmitirse utilizando cualquier soporte adecuado incluyendo, sin limitación a un soporte inalámbrico, de línea cableada, de cable de fibra óptica, de RF, etc. o cualquier combinación adecuada de lo que antecede.

El código de programa informático para realizar operaciones para aspectos de la presente idea inventiva puede ser objeto de escritura en cualquier combinación de uno o más lenguajes de programación, incluyendo un lenguaje de programación orientado al objeto tal como Java, Smalltalk, C++ o similar y lenguajes de programación de procedimiento convencionales, tales como el lenguaje de programación "C" o lenguajes de programación similares. El código de programa puede ejecutarse completamente en el ordenador del usuario como un paquete de software autónomo, o en parte en el ordenador del usuario y en parte en un ordenador distante o completamente en el ordenador distante o servidor. En este último escenario operativo, el ordenador distante puede estar conectado al ordenador del usuario por intermedio de cualquier tipo de red, incluyendo una red de área local (LAN) o una red de área amplia (WAN), o la conexión puede realizarse a un ordenador externo (a modo de ejemplo, por intermedio de Internet utilizando un denominado Proveedor de Servicios de Internet).

Aspecto de la presente idea inventiva se describen a continuación haciendo referencia a ilustraciones de diagramas de flujo y/o diagramas de bloques de métodos, aparatos (sistemas) y productos de programas informáticos de conformidad con las formas de realización de la presente solicitud de patente. Se entenderá que cada bloque de las ilustraciones del diagrama de flujo y/o diagramas de bloques y combinaciones de bloques en las ilustraciones de diagramas de flujos y/o diagramas de bloques, pueden realizarse mediante instrucciones de programas informáticos. Estas instrucciones de programas informáticos pueden proporcionarse a un procesador de un ordenador de uso general, ordenador de uso especial u otro aparato de procesamiento de datos programable para obtener una máquina, de modo que las instrucciones, que se ejecutan mediante el procesador del ordenador u otro aparato de procesamiento de datos programable, puedan crear medios para realizar las funciones/actos especificados en el bloque de diagramas de flujo y/o diagrama de bloques o bloques:

Estas instrucciones de programas informáticos pueden memorizarse también en un soporte legible por ordenador que puede controlar un ordenador, otro aparato de procesamiento de datos programable, u otros dispositivos para funcionar en una manera particular, tal como las instrucciones memorizadas en el soporte legible por ordenador que presentan un artículo de fabricación incluyendo instrucciones que realizan la función/acto especificado en el bloque de diagramas de flujo y/o diagrama de bloques o bloques.

Las instrucciones de programa informático pueden cargarse también en un ordenador, otro aparato de procesamiento de datos programable, u otros dispositivos para hacer que una serie de operaciones funcionales se realicen en el ordenador, otro aparato programable u otros dispositivos para proporcionar un proceso realizado por ordenador tal como las instrucciones que se ejecutan en el ordenador u otro aparato programable que proporcionan procesos para poner en práctica las funciones/actos especificados en el diagrama de flujo y/o bloques o bloque de diagramas de bloques.

A continuación se describirán en detalle las formas de realización de la presente idea inventiva. Para mayor claridad, la descripción se organiza en la arquitectura siguiente:

Parte 1: Aparatos y Métodos de procesamiento de audio

Sección 1.1 Tipos de audio

Sección 6.2 Extracción de características a largo plazo

Sección 6.3 Extracción de características a corto plazo

Sección 6.4 Combinación de formas de realización y escenarios de aplicación

Sección 6.5 Métodos de clasificación de audio

Parte 7: Clasificadores de VoIP y Métodos de clasificación

Sección 7.1 Clasificación de contexto basado en un segmento a corto plazo

Sección 7.2 Clasificación utilizando voz de VoIP y ruido de VoIP

Sección 7.3 Fluctuación de alisado

Sección 7.4 Combinación de formas de realización y escenarios de aplicación

Sección 7.5 Métodos de clasificación de VoIP

Parte 1: Aparatos y métodos de procesamiento de audio

La Figura 1 ilustra un marco general de un aparato de procesamiento de audio de contenido adaptativo 100 que soporta la configuración automática de al menos un dispositivo de mejora de audio 400 con parámetros mejorados sobre la base del contenido de audio en la reproducción. Comprende tres componentes principales: un clasificador de audio 200, una unidad de ajuste 300 y un dispositivo de mejora de audio 400.

El clasificador de audio 200 es para clasificar una señal de audio en al menos un tipo de audio en tiempo real. Identifica automáticamente los tipos de audio del contenido en la reproducción. Cualesquiera tecnologías de clasificación de audio tales como un procesamiento de señal pasante, aprendizaje de máquina y reconocimiento de modelos, pueden aplicarse para identificar el contenido de audio. Los valores de confianza, que representan las probabilidades del contenido de audio con respecto a un conjunto de tipos de audio objetivo predefinidos, se estiman generalmente al mismo tiempo.

El dispositivo de mejora de audio 400 es para mejorar la experiencia de la audiencia en la realización del procesamiento de la señal de audio, y se describirá en detalle más adelante.

La unidad de ajuste 300 es para ajustar al menos un parámetro del dispositivo de mejora de audio en una manera continua basada en el velocidad del al menos un tipo de audio. Está diseñada para controlar el comportamiento del dispositivo de mejora de audio 400. Estima los parámetros más adecuados del dispositivo de mejora de audio correspondiente sobre la base de los resultados obtenidos a partir del clasificador de audio 200.

Varios dispositivos de mejora de audio pueden aplicarse en este aparato. La Figura 2 ilustra un sistema, a modo de ejemplo, que comprende cuatro dispositivos de mejora de audio, incluyendo un Mejorador de Diálogos (DE) 402, un Virtualizador de sonido envolvente (SV) 404, un Nivelador de volumen (VL) 406 y un Ecualizador (EQ) 408. Cada dispositivo de mejora de audio puede ajustarse automáticamente de una manera continua, sobre la base de los resultados (tipos de audio y/o valores de confianza) obtenidos en el clasificador de audio 200.

Por supuesto, los aparatos de procesamiento de audio pueden no incluir necesariamente todas las clases de dispositivos mejora de audio, sino que pueden incluir solamente uno o más de ellos. Por otro lado, los dispositivos de mejora de audio no están limitados a los dispositivos descritos en la presente idea inventiva y pueden incluir más clases de dispositivos de mejora de audio que estén también dentro del alcance de la presente idea inventiva. Además, los nombres de dichos dispositivos de mejora de audio examinados en la presente idea inventiva, incluyendo un Mejorador de diálogos (DE) 402, un Virtualizador de sonido envolvente (SV) 404, un Nivelador de volumen (VL) 406 y un Ecualizador (EQ) 408, no constituirán una limitación y cada uno de ellos deberá interpretarse como que cubre cualesquiera otros dispositivos que realicen las mismas o funciones similares.

1.1 Tipos de audio

Para controlar adecuadamente varias clases de dispositivos de mejora de audio, la presente idea inventiva proporciona, además, una nueva arquitectura de tipos de audio, aunque los tipos de audio de la técnica anterior son también aquí aplicables.

Más concretamente, los tipos de audio de diferentes niveles semánticos son objeto de modelado, incluyendo elementos de audio de bajo nivel que representan las componentes fundamentales en señales de audio y géneros de audio de alto nivel que representan los contenidos de audio más populares en las aplicaciones de entretenimiento de usuarios en la vida real. Lo anterior puede denominarse también como "tipo de contenido". Los tipos de contenidos de audio

fundamentales pueden incluir, voz, música (incluyendo canción), sonidos de fondo (o efectos sonoros) y ruido.

El significado de voz y música es evidente por sí mismo. El ruido en la presente idea inventiva significa ruido físico y no ruido semántico. El ruido físico en la presente idea inventiva puede incluir los ruidos de, a modo de ejemplo, sistemas de aire acondicionado y otros ruidos que tienen su origen en razones técnicas, tales como ruidos de bajo nivel, denominados 'ruido rosa' debido a la ruta de transmisión de señales. Por el contrario, los "sonidos de fondo" en la presente solicitud de patente son los efectos sonoros que pueden ser eventos auditivos que suceden alrededor del objetivo básico de la atención del oyente. A modo de ejemplo, en una señal de audio en una llamada telefónica, además de la voz de la persona que habla, puede existir algunos otros sonidos no previstos, tales como las voces de algunas otras personas no intervinientes en la llamada telefónica, sonidos de teclados, sonidos de pasos, etc. Estos sonidos no deseados se refieren como "sonidos de fondo" y no como ruido. Dicho de otro modo, podemos definir los "sonidos de fondo" como los sonidos que no son el objetivo (o el objetivo básico de la atención del oyente) o incluso no son deseados, pero siguen teniendo algún significado semántico; mientras que el "ruido" puede definirse como los sonidos no deseados con la excepción de los sonidos objetivos y los sonidos de fondo.

A veces, los sonidos de fondo no son realmente "indeseados" sino que se crean intencionadamente e incluyen alguna información de utilidad, tal como los sonidos de fondo en una película cinematográfica, un programa de TV o un programa de radiodifusión. Por ello, a veces pueden también referirse como "efectos sonoros". En adelante, en la presente idea inventiva, solamente se utiliza el término de "sonidos de fondo" para no ser concisos y pueden abreviarse también como "fondo".

Además, la música puede clasificarse, además, como música sin fuentes dominantes y música con fuentes dominantes. Si existe una fuente (voz o un instrumento) que es mucho más intensa que las demás fuentes en una pieza musical, se refiere como una "música con fuente dominante", de no ser así, se refiere como "música sin fuente dominante". A modo de ejemplo, en una música polifónica acompañada con voz de cantantes y varios instrumentos, si está armónicamente equilibrados, o la energía de varias fuentes más notorias son comparables entre sí, se considera que es una música sin fuente dominante; por el contrario, si una fuente (p.ej., una voz) es mucho más intensa mientras que las demás son mucho más silenciosas, se considera que contiene una fuente dominante. A modo de otro ejemplo, los tonos de instrumentos singulares o distintivos son "música con fuente dominante".

La música puede clasificarse, además, en diferentes tipos sobre la base de normas distintas. Puede clasificarse sobre la base de géneros de la música, tales como rock, jazz, rap y folk, pero sin que suponga una limitación. Se pueden clasificar también sobre la base de instrumentos, tales como música vocal y música instrumental. La música instrumental puede incluir varias músicas ejecutadas con diferentes instrumentos, tales como música de piano y música de guitarra. Otras normas ejemplo incluyen ritmo, *tempo*, timbre de la música y/o cualesquiera otros atributos musicales, de modo que la música se puede agrupar junta sobre la base de la similitud de estos atributos. A modo de ejemplo, en función del timbre, la música vocal puede clasificarse como tenor, barítono, bajo, soprano, mezzo soprano y alto.

El tipo de contenido de una señal de audio puede clasificarse con respecto a segmentos de audio a corto plazo, tales como los constituidos por una pluralidad de tramas. En general, una trama de audio es de una longitud de múltiples milisegundos, tales como 20 ms, y la longitud de un segmento de audio a corto plazo a clasificarse por el clasificador de audio puede tener una duración de varios cientos de milisegundos hasta varios segundos, tal como 1 segundo.

Para controlar el dispositivo de mejora de audio en una manera de contenido-adaptativo, la señal de audio puede clasificarse en tiempo real. Para el tipo de contenido establecido anteriormente, el tipo de contenido del segmento de audio a corto plazo actual representa el tipo de contenido de la señal de audio actual. Puesto que la duración de un segmento de audio a corto plazo no es tan larga, la señal de audio puede dividirse como segmentos de audio a corto plazo no solapados, uno tras otros. Sin embargo, los segmentos de audio a corto plazo pueden muestrearse también de forma continua/semi-continua a lo largo de la línea de tiempos de la señal de audio. Es decir, los segmentos de audio a corto plazo pueden muestrearse con una ventana con una longitud predeterminada (longitud prevista del segmento de audio a corto plazo) que se desplazan a lo largo de la línea de tiempos de la señal de audio en una magnitud de tonos de una o más tramas.

Los géneros de audio de alto nivel pueden nombrarse también como "tipos de contextos", puesto que indican un tipo a largo plazo de la señal de audio, y pueden considerarse como un entorno o contexto del evento sonoro instantáneo, que puede clasificarse en los tipos de contenidos según se indicó con anterioridad. De conformidad con la presente idea inventiva, el tipo de contexto puede incluir las aplicaciones de audio más populares, tales como multimedia similar a películas cinematográficas, música (incluyendo canción), juego y VoIP (Protocolo de Voz sobre Internet).

El significado de música, juego y VoIP es evidente por sí mismo. Los soportes similares a una película cinematográfica pueden incluir una película cinematográfica, un programa de TV, programas de radiodifusión o cualquier otro soporte de audio similar a los anteriormente mencionados. La característica principal de multimedia similar a cine es una mezcla de posibles voces, música y varias clases de sonido de fondo (efectos sonoros).

Conviene señalar que el tipo de contenido y el tipo de contexto incluyen música (incluyendo canción). En adelante, en la presente idea inventiva, utilizamos los términos "música a corto plazo" y "música a largo plazo" para distinguirlos

respectivamente.

Para algunas formas de realización de la presente idea inventiva, se proponen también algunas otras arquitecturas de tipo de contexto.

A modo de ejemplo, una señal de audio puede clasificarse como audio de alta calidad (tal como los soportes a modo de película cinematográfica y CD de música) o audio de baja calidad (tal como VoIP, audio de flujo continuo en línea de tasa binaria baja y contenido generado por el propio usuario), que se pueden referir colectivamente como “tipos de calidad de audio”.

A modo de otro ejemplo, una señal de audio puede clasificarse como VoIP o no VoIP, lo que puede considerarse como una transformación de la arquitectura del tipo de contexto 4 anteriormente mencionada (VoIP, soportes a modo de película cinematográfica, música (largo plazo) y juegos). En relación con el contexto de VoIP o de no VoIP, una señal de audio puede clasificarse como tipos de contenidos relacionados con VoIP, tales como voz de VoIP, voz no de VoIP, ruido de VoIP y ruido no de VoIP. La arquitectura de los tipos de contenidos de audio de VoIP son de utilidad particular para diferenciar los contextos de VoIP y de no VoIP puesto que el contexto de VoIP suele ser el escenario operativo de aplicación más exigente de un nivelador de volumen (una clase de dispositivo de mejora de audio).

Por lo general, el tipo de contexto de una señal de audio puede clasificarse con respecto a segmentos de audio a largo plazo de mayor duración que los segmentos de audio a corto plazo. Un segmento de audio a largo plazo está constituido por una pluralidad de tramas en un número superior al número de tramas en un segmento de audio a corto plazo. Un segmento de audio a largo plazo puede comprender también una pluralidad de segmentos de audio a corto plazo. Por lo general, un segmento de audio a largo plazo puede tener una duración del orden de magnitud de los segundos, tal como varios segundos a varias decenas de segundos, a modo de ejemplo 10 segundos.

De modo similar, para controlar el dispositivo de mejora de audio en una manera adaptativa, la señal de audio puede clasificarse en tipos de contextos en tiempo real. De modo similar, el tipo de contexto del segmento de audio a largo plazo actual representa el tipo de contexto de la señal de audio actual. Puesto que la longitud de un segmento de audio a largo plazo es relativamente larga, la señal de audio puede muestrearse de forma continua/semi-continua a lo largo de la línea de tiempos de la señal de audio para evitar un cambio brusco de su tipo de contexto y de este modo, un cambio brusco de los parámetros funcionales de los dispositivos de mejora de audio. Es decir, los segmentos de audio a largo plazo pueden muestrearse como una ventana con una longitud predeterminada (longitud prevista de un segmento de audio a largo plazo) que se desplaza a lo largo de la línea de tiempos de la señal de audio con una magnitud de los tonos de una o más tramas, o uno o más segmentos a corto plazo.

Lo que antecede se ha descrito con respecto al tipo de contenido y al tipo de contexto. En las formas de realización de la presente idea inventiva, la unidad de ajuste 300 puede ajustar al menos un parámetro de los dispositivos de mejora de audio sobre la base de al menos uno de los tipos de contenidos y/o y al menos uno de los diversos tipos de contextos. Por lo tanto, según se ilustra en la Figura 3, en una variante de la forma de realización ilustrada en la Figura 1, el clasificador de audio 200 puede comprender un clasificador de contenido de audio 202 o un clasificador de contexto de audio 204 o ambos a la vez.

Anteriormente se han mencionado diferentes tipos de audio basados en diferentes normas (tales como para los tipos de contextos), así como diferentes tipos de audio basados en diferentes niveles jerárquicos (tales como para los tipos de contenidos). Sin embargo, las normas y los niveles jerárquicos son solamente por comodidad de descripción en este caso y por supuesto, no tienen carácter de limitación. Dicho de otro modo, en la presente idea inventiva, cualesquiera dos o más tipos de audio anteriormente mencionados pueden identificarse por el clasificador de audio 200 al mismo tiempo y considerarse por la unidad de ajuste 300 al mismo tiempo, según se describirá más adelante. Dicho de otro modo, todos los tipos de audio en los diferentes niveles jerárquicos pueden ser paralelos o estar en el mismo nivel.

1.2 Valores de confianza de tipos de audio y arquitectura de clasificadores

El clasificador de audio 200 puede proporcionar, a la salida, resultados de decisiones difíciles, o la unidad de ajuste 300 puede considerar los resultados del clasificador de audio 200 como resultados de decisiones difíciles. Incluso para la decisión difícil, múltiples tipos de audio pueden asignarse a un segmento de audio. A modo de ejemplo, un segmento de audio puede etiquetarse mediante, a la vez, ‘voz’ y ‘música a corto plazo’ puesto que puede ser una señal mezcla de voz y música a corto plazo. Las etiquetas obtenidas pueden utilizarse directamente para controlar los dispositivos de mejora de audio 400. Un ejemplo simple es activar el dispositivo mejorador de diálogos 402 cuando la voz está presente y desactivarlo cuando la voz está ausente. Sin embargo, este método de decisión difícil puede introducir alguna falta de naturalidad en los puntos de transición desde un tipo de audio a otro, si no se dispone de un sistema de alisado cuidadoso (lo que se describirá más adelante).

Con el fin de tener más flexibilidad y sintonizar los parámetros de los dispositivos de mejora de audio de una manera continua, el valor de confianza de cada tipo de audio objetivo se puede estimar (decisión programada). Un valor de confianza representa el nivel adaptado entre el contenido de audio a identificarse y el tipo de audio objetivo, con valores desde 0 a 1.

Según se indicó con anterioridad, numerosas técnicas de clasificación pueden proporcionar valores de confianza directamente. El valor de confianza puede calcularse también a partir de varios métodos, que pueden considerarse como una parte del clasificador. A modo de ejemplo, si los modelos de audio se forman mediante algunas tecnologías de modelado probabilístico, tal como Modelos de Mezcla Gaussiana (GMM), la probabilidad posterior puede utilizarse para representar el valor de confianza, como

$$p(c_i \mid x) = \frac{p(x \mid c_i)}{\sum_{i=1}^{N} p(x \mid c_i)} \qquad (1)$$

en donde x es un elemento de segmento de audio, c_i es un tipo de audio objetivo, N es el número de tipos de audio objetivos, $p(x|c_i)$ es la probabilidad de que el segmento de audio x sea del tipo de audio c_i, y $p(c_i|x)$ es la probabilidad posterior correspondiente.

Por otro lado, si los modelos de audio se forman a partir de algunos métodos discriminativos, tales como la denominada Máquina de Vectores Soporte (SVM) y adaBoost, solamente se obtienen puntuaciones (valores reales) a partir de la comparación de modelos. En estos casos, una función sigmoidal se suele utilizar para establecer una correspondencia de la puntuación obtenida (teóricamente desde -∞ a ∞) a la confianza prevista (desde 0 a 1):

$$conf = \frac{1}{1+e^{Ay+B}} \qquad (2)$$

en donde el valor de y es la puntuación de salida desde SVM o adaBoost, A y B son dos parámetros que necesitan estimarse a partir de un conjunto de datos de formación utilizando algunas tecnologías bien conocidas.

Para algunas formas de realización de la presente idea inventiva, la unidad de ajuste 300 puede utilizar más de dos tipos de contenidos y/o más de dos tipos de contextos. A continuación, el clasificador de contenido de audio 202 necesita identificar más de dos tipos de contenidos y/o el clasificador de contexto de audio 204 necesita identificar más de dos tipos de contextos. En tal situación, el clasificador de contenido de audio 202 o el clasificador de contexto de audio 204 puede ser un grupo de clasificadores organizados en alguna arquitectura.

A modo de ejemplo, si la unidad de ajuste 300 necesita la totalidad de las cuatro clases de tipos de contextos de multimedia similar a cine, música a largo plazo, juego y VoIP, entonces, el clasificador de contexto de audio 204 puede tener las arquitecturas diferentes siguientes:

En primer lugar, el clasificador de contexto de audio 204 puede comprender 6 clasificadores binarios del tipo uno a uno (cada clasificador discrimina un tipo de audio objetivo a partir de otro tipo de audio objetivo) organizado según se ilustra en la Figura 4, 3 clasificadores binarios del tipo uno a otros (cada clasificador discrimina un tipo de audio objetivo a partir de los demás) organizado según se ilustra en la Figura 5 y 4 clasificadores del tipo uno a otros organizado según se ilustra en la Figura 6. Existen también otras arquitecturas tales como la arquitectura de Gráfico Acíclico dirigido por la Decisión (DDAG). Conviene señalar que, en las Figuras 4 a 6 y en la descripción correspondiente siguiente, los términos “película cinematográfica” en lugar de “multimedia similar a cine” se utilizan para mayor concisión.

Cada clasificador binario proporcionará una puntuación de confianza H(x) para su salida (x representa un segmento de audio). Después de que se obtengan las salidas de cada clasificador binario, necesitamos establecer una correspondencia entre ellas con respecto a los valores de confianza finales de los tipos de contextos identificados.

En general, se supone que la señal de audio ha de clasificarse en M tipos de contexto (M es un número entero positivo). La arquitectura de tipo ‘uno a uno’ convencional construye $\mathcal{M}(\mathcal{M}-1)/2$ clasificadores en donde cada uno se forma sobre datos procedentes de dos clases, a continuación, cada clasificador del tipo ‘uno a uno’ emite un voto para su clase preferida, y el resultado final es la clase con la mayor cantidad de votos entre las clasificaciones de $\mathcal{M}(\mathcal{M}-1)/2$ de los clasificadores. En comparación, con la arquitectura de tipo ‘uno a uno’ convencional, la arquitectura jerárquica ilustrada en la Figura 4 necesita también construir $\mathcal{M}(\mathcal{M}-1)/2$ clasificadores. Sin embargo, las iteraciones de pruebas pueden acortarse a $\mathcal{M}-1$, puesto que el segmento x será determinado como siendo/no siendo de la clase correspondiente a cada nivel jerárquico y el conteo de nivel global es $\mathcal{M}-1$. Los valores de confianza finales para varios tipos de contextos pueden calcularse a partir de la confianza de clasificación binaria $H_k(x)$, a modo de ejemplo (k=1,2,...6, que representan diferentes tipos de contexto):

$$C_{MOVIE} = \big(1 - H_1(x)\big) \cdot (1 - H_3(x)) \cdot (1 - H_6(x))$$

$$C_{VOIP} = H_1(x) \cdot H_2(x) \cdot H_4(x)$$

$$C_{MUSIC} = H_1(x) \cdot \big(1 - H_2(x)\big) \cdot \big(1 - H_5(x)\big) + H_3(x) \cdot \big(1 - H_1(x)\big) \cdot \big(1 - H_5(x)\big) + H_6(x) \cdot (1 - H_1(x)) \cdot (1 - H_3(x))$$

$$C_{GAME} = H_1(x) \cdot H_2(x) \cdot \big(1 - H_4(x)\big) + H_1(x) \cdot H_5(x) \cdot \big(1 - H_2(x)\big) + H_3(x) \cdot H_5(x) \cdot \big(1 - H_1(x)\big)$$

En la arquitectura ilustrada en la Figura 5, la función de mapeado de correspondencia desde los resultados de clasificación binarios $H_k(x)$ a los valores de confianza finales pueden definirse en el ejemplo siguiente.

$$C_{MOVIE} = H_1(x)$$

$$C_{MUSIC} = H_2(x) \cdot \big(1 - H_1(x)\big)$$

$$C_{VOIP} = H_3(x) \cdot \big(1 - H_2(x)\big) \cdot \big(1 - H_1(x)\big)$$

$$C_{GAME} = (1 - H_3(x)) \cdot \big(1 - H_2(x)\big) \cdot \big(1 - H_1(x)\big)$$

En la arquitectura ilustra en la Figura 6, los valores de confianza finales pueden ser iguales a los resultados de clasificaciones binarias correspondientes $H_k(x)$, o si la suma de los valores de confianza para todas las clases se requiere que sea 1, entonces, los valores de confianza finales pueden simplemente normalizarse sobre la base de los resultados $H_k(x)$ estimados:

$$C_{MOVIE} = H_1(x)/(H_1(x) + H_2(x) + H_3(x) + H_4(x))$$

$$C_{MUSIC} = H_2(x)/(H_1(x) + H_2(x) + H_3(x) + H_4(x))$$

$$C_{VOIP} = H_3(x)/(H_1(x) + H_2(x) + H_3(x) + H_4(x))$$

$$C_{GAME} = H_4(x)/(H_1(x) + H_2(x) + H_3(x) + H_4(x))$$

Los uno o más con los valores de confianza máxima pueden determinarse para ser la clase identificada final.

Conviene señalar que en las arquitecturas ilustradas en las Figuras 4 a 6, la secuencia de clasificadores binarios diferentes no son necesariamente como se ilustran, sino que pueden ser otras secuencias, que pueden seleccionarse mediante asignación manual o aprendizaje automático en conformidad con diferentes requisitos de varias aplicaciones.

Las descripciones anteriores están dirigidas a clasificadores de contexto de audio 204. Para el clasificador de contenido de audio 202, la situación es similar.

Como alternativa, el clasificador de contenido de audio 202 o el clasificador de contexto de audio 204 pueden ponerse en práctica como un clasificador único que identifica todos los tipos de contenido/tipos de contexto al mismo tiempo y proporcionan los valores de confianza correspondientes al mismo tiempo. Existen numerosas técnicas para realizar esta operación.

Utilizando el valor de confianza, la salida del clasificador de audio 200 puede representarse como un vector, con cada dimensión representando el valor de confianza de cada tipo de audio objetivo. A modo de ejemplo, si los tipos de audio objetivos (voz, música a corto plazo, ruido, fondo) de forma secuencial, un resultado de salida ejemplo podría ser (0.9, 0.5, 0.0, 0.0), lo que indica que es un 90 % seguro que el contenido de audio sea de voz, y un 50 % seguro que el audio sea música. Conviene señalar que la suma de todas las dimensiones en el vector de salida no es necesario que sea de valor uno (a modo de ejemplo, los resultados de la Figura 6 no son necesariamente normalizados), lo que significa que la señal de audio puede ser una señal de mezcla de voz y de música a corto plazo.

Más adelante, en la Parte 6 y en la Parte 7, se describirá, en detalle, una nueva puesta en práctica de la clasificación del contexto de audio y la clasificación de contenidos de audio.

1.3 Alisado de valores de confianza de tipos de audio

De modo opcional, después de que se haya clasificado cada segmento de audio en los tipos de audio predefinidos, un

paso adicional es el alisado de los resultados de la clasificación a lo largo de la línea de tiempos para evitar un salto brusco desde un tipo a otro y para realizar una estimación más alisada de los parámetros en los dispositivos de mejora de audio. A modo de ejemplo, un extracto largo se clasifica como multimedia similar a cine, exceptuado para solamente un segmento clasificado como VoIP, siendo, entonces, la decisión de VoIP brusca que puede revisarse para multimedia similar a cine mediante el alisado correspondiente.

Por lo tanto, en una variante de la forma de realización según se ilustra en la Figura 7, una unidad de alisado tipo 712 está provista, además, para cada tipo de audio, alisando el valor de confianza de la señal de audio en el momento actual.

Un método de alisado común está basado en una media ponderada, tal como se calcula una suma ponderada del valor de confianza real en el momento actual y un valor de confianza alisado de la última vez, como sigue:

$$smoothConf(t) = \beta \cdot smoothConf(t-1) + (1-\beta) \cdot conf(t) \quad (3)$$

en donde t representa el tiempo actual (el segmento de audio actual), t-1 representa la última vez (el último segmento de audio), β es el peso, conf y smoothConf son los valores de confianza antes y después del alisado, respectivamente.

Desde el punto de vista de los valores de confianza, los resultados de una decisión difícil de los clasificadores pueden representarse también con valores de confianza, con los valores siendo 0 o 1. Es decir, si un tipo de audio objetivo se elige y asigna a un segmento de audio, el valor de confianza correspondiente es 1; de no ser así, el valor de confianza es 0. Por lo tanto, aun cuando el clasificador de audio 200 no proporcione el valor de confianza, sino que simplemente proporcione una decisión difícil con respecto al tipo de audio, el ajuste continuo de la unidad de ajuste 300 es todavía posible mediante la operación de alisado de la unidad de alisado tipo 712.

El algoritmo de alisado puede ser 'asimétrico' utilizando diferentes ponderaciones de alisado para diferentes casos. A modo de ejemplo, las ponderaciones para calcular la suma ponderada pueden cambiarse de forma adaptativa sobre la base del valor de confianza del tipo de audio de la señal de audio. El valor de confianza del segmento actual es mayor, siendo también mayor su ponderación.

Desde otro punto de vista, las ponderaciones para calcular la suma ponderada pueden cambiarse, de forma adaptativa, sobre la base de diferentes pares de transición desde un tipo de audio a otro tipo de audio, en particular cuando el dispositivo de mejora de audio está ajustado sobre la base de múltiples tipos de contenido según se identifica por el clasificador de audio 200, en lugar de basarse en la presencia o ausencia de un tipo de contenido único. A modo de ejemplo, para una transición desde un tipo de audio que aparece con mayor frecuencia en algún contexto a otro tipo de audio que no aparece tan frecuentemente en el contexto, el valor de confianza de este último puede alisarse de modo que no aumente tan rápido, puesto que podría ser simplemente una interrupción ocasional.

Otro factor es la tendencia de cambio (aumento o disminución), incluyendo la tasa de cambio. Se supone que tenemos más cuidado respecto a la latencia cuando un tipo de audio se hace presente (es decir, cuando aumenta su valor de confianza), podemos designar el algoritmo de alisado en la forma siguiente:

$$smoothConf(t) = \begin{cases} conf(t) & conf(t) \geq smoothConf(t-1) \\ \beta \cdot smoothConf(t-1) + (1-\beta) \cdot conf(t) & de\ lo\ contrario \end{cases} \quad (4)$$

La fórmula anterior permite que el valor de confianza alisado responda con rapidez al estado actual cuando aumenta el valor de confianza y lentamente alisado cuando disminuye el valor de confianza. Variantes de las funciones de alisado pueden fácilmente diseñarse en forma similar. A modo de ejemplo, la fórmula (4) puede revisarse de modo que el valor de ponderación de conf(t) se haga mayor cuando se verifique conf(t)>=smoothConf(t-1). De hecho, en la fórmula (4) se puede considerar que β =0 y el valor de ponderación de conf(t) se hace el mayor, es decir 1.

Desde un punto de vista diferente, la consideración de que la tendencia cambiante de algún tipo de audio es simplemente un ejemplo específico de consideración de diferentes pares de transición de tipos de audio. A modo de ejemplo, aumentando el valor de confianza del tipo A puede considerarse como una transición desde no A a A y la disminución del valor de confianza de tipo A puede considerarse como una transición de A a no A.

1.4 Ajuste de parámetros

La unidad de ajuste 300 está diseñada para estimar o ajustar los parámetros adecuados para los dispositivos de mejora de audio 400 sobre la base de los resultados obtenidos a partir del clasificador de audio 200. Diferentes algoritmos de ajuste pueden diseñarse para diferentes dispositivos de mejora de audio, utilizando el tipo de contenido o el tipo de contexto, o ambos a la vez, para una decisión conjunta. A modo de ejemplo, con la información del tipo de contexto tal como multimedia similar a cine y música a largo plazo, los preajustes, según fueron anteriormente mencionados, pueden seleccionarse y aplicarse automáticamente sobre el contenido correspondiente. Con la información del tipo de contenido

disponible, los parámetros de cada dispositivo de mejora de audio pueden sintonizarse en una manera más fina, según se ilustra en las partes posteriores. La información del tipo de contenido y la información del contexto pueden utilizarse conjuntamente, además, en la unidad de ajuste 300 para equilibrar la información a largo plazo y la información a corto plazo. El algoritmo de ajuste específico para un dispositivo de mejora de audio específico puede considerarse como una unidad de ajuste separada, o los algoritmos de ajuste diferentes pueden considerarse colectivamente como una unidad de ajuste unificada.

Es decir, la unidad de ajuste 300 puede configurarse para ajustar el al menos un parámetro del dispositivo de mejora de audio sobre la base del valor de confianza de al menos un tipo de contenido y/o el valor de confianza de al menos un tipo de contexto. Para un dispositivo de mejora de audio específico, algunos tipos de audio son informativos, y algunos de los tipos de audio son interferentes. En consecuencia, los parámetros del dispositivo de mejora de audio específico pueden estar en correlación positiva o negativa respecto a los valores de confianza de los tipos de audio informativos o los tipos de audio interferentes. En este caso, el término de “correlación positiva” significa los aumentos o disminuciones de parámetros con el aumento o disminución del valor de confianza del tipo de audio, en una manera lineal o en una manera no lineal. El término de “correlación negativa” significa los aumentos o disminuciones de parámetros con, respectivamente, la disminución o aumento del valor de confianza del tipo de audio, en una manera lineal o en una manera no lineal.

En este caso, la disminución y aumento del valor de confianza son directamente “transferidos” a los parámetros a ajustarse mediante la correlación positiva o negativa. En matemáticas, dicha correlación o “transferencia” puede materializarse como proporción lineal o proporción inversa, operación de más o menos (adición o sustracción), operación de multiplicación o división o función no lineal. Todas estas formas de correlación pueden referirse como “función de transferencia”. Para determinar el aumento o disminución del valor de confianza, podemos comparar también el valor de confianza actual o su transformada matemática con el último valor de confianza o una pluralidad de valores de confianza históricos, o sus transformadas matemáticas. En el contexto de la presente idea inventiva, el término “comparar” significa la comparación mediante una operación de sustracción o la comparación mediante una operación de división. Podemos determinar un aumento o disminución determinando si la diferencia es mayor que 0 o si la relación es mayor que 1.

En puestas en práctica específicas, podemos relacionar directamente los parámetros con los valores de confianza o sus relaciones o diferencias por intermedio de un algoritmo adecuado (tal como una función de transferencia) y no es necesaria la presencia de un “observador externo” para conocer explícitamente si un valor de confianza específico y/o un parámetro específico ha aumentado o disminuido. Algunos ejemplos específicos se proporcionarán en las Partes 2 a 5 posteriores en relación con los dispositivos de mejora de audio específicos.

Según se describió en la sección anterior, con respecto al mismo segmento de audio, el clasificador 200 puede identificar múltiples tipos de audio con valores de confianza respectivos, cuyos valores de confianza pueden no ser necesariamente de valor 1, puesto que el segmento de audio puede comprender múltiples componentes al mismo tiempo, tal como música y voz y sonido de fondo. En tal situación, los parámetros de los dispositivos de mejora de audio se equilibrarán entre diferentes tipos de audio. A modo de ejemplo, la unidad de ajuste 300 puede configurarse para considerar al menos algunos de los múltiples tipos de audio mediante la ponderación de los valores de confianza del al menos un tipo de audio sobre la base de la importancia del al menos un tipo de audio. Cuanto más importante es un tipo de audio específico, tanto mayor será la influencia de los parámetros correspondientes.

El valor de ponderación puede reflejar también un efecto informativo e interferente de un tipo de audio. A modo de ejemplo, para un tipo de audio interferente, puede proporcionarse una ponderación de signo menos. Algunos ejemplos específicos se proporcionarán en las Partes 2 a 5 posteriores sobre los dispositivos de mejora de audio específicos.

Conviene señalar que el contexto de la presente idea inventiva, el término “ponderación” tiene un significado más amplio que los coeficientes en un polinomio. Además, los coeficientes en un polinomio, pueden adoptar también la forma de exponente o potencia. Cuando los coeficientes adoptan la forma polinomial, los coeficientes de ponderación pueden estar, o no, normalizados. En resumen, la ponderación simplemente representa cuánta influencia tiene el objeto ponderado sobre los parámetros que han de ajustarse.

En algunas otras formas de realización, para los múltiples tipos de audio contenidos en el mismo segmento de audio, los valores de confianza pueden convertirse en ponderaciones mediante su normalización y luego, el parámetro final puede determinarse mediante el cálculo de una suma de valores prestablecidos de parámetros predefinidos para cada tipo de audio y ponderados por los valores de ponderación basados en los valores de confianza. Es decir, la unidad de ajuste 300 puede configurarse para considerar los múltiples tipos de audio mediante ponderación de los efectos de los múltiples tipos de audio sobre la base de los valores de confianza.

Como un ejemplo específico de ponderación, la unidad de ajuste está configurada para considerar al menos un tipo de audio dominante sobre la base de los valores de confianza. Para dichos tipos de audio que tienen valores de confianza demasiado bajos (menor que un valor umbral), pueden no considerarse. Esto es equivalente al hecho de que las ponderaciones de los otros tipos de audio, cuyos valores de confianza son menores que el valor umbral, se ajustan como cero. Algunos ejemplos específicos se proporcionarán en las Partes 2 a 5 siguientes sobre dispositivos de mejora de audio específicos.

El tipo de contenido y el tipo de contexto pueden considerarse juntos. En una forma de realización, pueden considerarse como al mismo nivel y sus valores de confianza pueden tener sus ponderaciones respectivas. En otra forma de realización, simplemente como indica la nominación, el "tipo de contexto" es el contexto o entorno en donde está ubicado el "tipo de contenido" y por lo tanto, la unidad de ajuste 200 puede configurarse de modo que el tipo de contenido en una señal de audio de un tipo de contexto diferente se le asigne un valor de ponderación diferente dependiendo del tipo de contexto de la señal de audio. En términos generales, cualquier tipo de audio puede constituir un contexto de otro tipo de audio y en consecuencia, la unidad de ajuste 200 puede configurarse para modificar el valor de ponderación de un tipo de audio con el valor de confianza de otro tipo de audio. Algunos ejemplos específicos se proporcionarán en las Partes 2 a 5 siguientes sobre dispositivos de mejora de audio específicos.

En el contexto de la presente idea inventiva, el término "parámetro" tiene un significado más amplio que su significado literal. Además de que un parámetro tenga un valor único, puede significar también un preajuste según se mencionó con anterioridad, incluyendo un conjunto de diferentes parámetros, un vector constituido por diferentes parámetros o un perfil. Más concretamente, en las Partes 2 a 5 posteriores, los siguientes parámetros se examinarán pero la presente idea inventiva no está limitada a este respecto: el nivel de mejorador de diálogos, los umbrales para determinar bandas de frecuencia para mejorador de diálogos, el nivel de fondo, la magnitud de refuerzo del sonido envolvente, la frecuencia de inicio para el virtualizador de sonido envolvente, la ganancia dinámica o la gama de la ganancia dinámica de un nivelador de volumen, los parámetros que indican el grado de la señal de audio que es un nuevo evento auditivo perceptible, el nivel de ecualización, los perfiles de ecualización y los preajustes de equilibrio espectral.

1.5 Alisado de parámetros

En la Sección 1.3, hemos examinado el alisado del valor de confianza de un tipo de audio para evitar su cambio brusco y de este modo, evitar un cambio brusco de los parámetros de los dispositivos de mejora de audio. Otras medidas son también posibles. Una consiste en el alisado del parámetro ajustado sobre la base del tipo de audio y se examinará en esta sección; la otra es configurar el clasificador de audio y/o la unidad de ajuste para retardar el cambio de los resultados del clasificador de audio y esta circunstancia se examinará en la sección 1.6.

En una forma de realización, el parámetro puede alisarse, además, para evitar un cambio rápido que puede introducir artefactos sonoros audibles en puntos de transición, como

$$\tilde{L}(t) = \tau\tilde{L}(t-1) + (1-\tau)L(t) \qquad (3')$$

en donde $\tilde{L}(t)$ es el parámetro alisado, $L(t)$ es el parámetro no alisado, τ es un coeficiente que representa una constante de tiempo, t es el tiempo actual y t-1 es el último tiempo.

Es decir, según se ilustra en la Figura 8, el aparato de procesamiento de audio puede comprender una unidad de alisado de parámetros 814 para, un parámetro del dispositivo de mejora de audio (tal como al menos uno de entre el dispositivo de mejorador de diálogos 402, el virtualizador de sonido envolvente 404, el nivelador de volumen 406 y el ecualizador 408) ajustados por la unidad de ajuste 300, el alisado del valor de parámetro determinado por la unidad de ajuste 300 en el momento actual calculando una suma ponderada del valor del parámetro que se determina por la unidad de ajuste en el momento actual y un valor de parámetro alisado de la última vez.

La constante de tiempo τ puede ser un valor fijo basado en el requisito específico de una aplicación y/o puesta en práctica del dispositivo de mejora de audio 400. Puede cambiarse también, de forma adaptativa, sobre la base del tipo de audio, en particular sobre la base de los diferentes tipos de transición desde un tipo de audio a otro, tal como desde música a voz, y desde voz a música.

Tomaremos un ecualizador como un ejemplo (detalles adicionales pueden consultarse en la Parte 5). La ecualización es adecuada para aplicarse sobre el contenido de música pero no sobre un contenido de voz. De este modo, para el alisado de nivel de acuse de recibo, la constante de tiempo puede ser relativamente pequeña cuando la señal de audio transita desde música a voz, de modo que un nivel de ecualización más pequeño puede aplicarse sobre el contenido de voz con mayor rapidez. Por otro lado, la constante de tiempo para la transición desde voz a música puede ser relativamente grande con el fin de evitar los artefactos sonoros audibles en los puntos de transición.

Para estimar el tipo de transición (p.ej., desde voz a música o desde música a voz) los resultados de la clasificación de contenidos pueden utilizarse de forma directa. Es decir, la clasificación del contenido de audio en música o voz se hace más sencilla para obtener el tipo de transición. Para estimar la transición de una manera más continua, podemos confiar también en el nivel de ecualización no alisado estimado, en lugar de comparar directamente la decisión difícil de los tipos de audio. La idea general es, si está aumentando el nivel de ecualización no alisado, ello indica una transición desde voz a música (o más música similar); de no ser así, es más como una transición desde música a voz (o más voz similar). Diferenciando los tipos de transición diferentes, la constante de tiempo puede establecerse en correspondencia, siendo un ejemplo:

$$\tau(t) = \begin{cases} \tau_1 & L(t) \geq L(t-1) \\ \tau_2 & L(t) < L(t-1) \end{cases} \qquad (4')$$

en donde τ (t) es la constante de tiempo variable con el tiempo dependiendo del contenido, τ 1 y τ 2 son dos valores de constante de tiempo prestablecidos, normalmente que satisfacen la relación τ 1 > τ 2. De forma intuitiva, la función anterior indica una transición relativamente lenta cuando aumenta el nivel de ecualización y una transición relativamente rápida cuando disminuye el nivel de ecualización, pero la presente idea inventiva no está limitada a este respecto. Además, el parámetro no está limitado al nivel de ecualización, sino que pueden ser otros parámetros. Es decir, la unidad de alisado de parámetros 814 puede configurarse de modo que los valores de ponderación para calcular la suma ponderada se cambien, de forma adaptativa, sobre la base de una tendencia de aumento o disminución del valor del parámetro que se determina por la unidad de ajuste 300.

1.6 Transición de tipos de audio

Con referencia a las Figuras 9 y 10, se describirá otro sistema para evitar un cambio brusco del tipo de audio y de este modo, evitar un cambio brusco de los parámetros de los dispositivos de mejora de audio.

Según se ilustra en la Figura 9, el aparato de procesamiento de audio 100 pueden comprender, además, un temporizador 916 para medir el tiempo de duración durante el cual el clasificador de audio 200 proporciona continuamente el mismo nuevo tipo de audio, en donde la unidad de ajuste 300 puede configurarse para seguir utilizando el tipo de audio presente hasta que la longitud del tiempo de duración del nuevo tipo de audio alcance un valor umbral.

Dicho de otro modo, se introduce una fase de observación (o de sostenimiento), según se ilustra en la Figura 10. Con la fase de observación (correspondiente al valor umbral de la longitud del tiempo de duración), el cambio de tipo de audio se controla, además, durante una cantidad de tiempo consecutiva para confirmar si el tipo de audio ha cambiado realmente, antes de que la unidad de ajuste 300 utilice realmente el nuevo tipo de audio.

Según se ilustra en la Figura 10, la flecha (1) ilustra la situación en donde el estado actual es el tipo A y el resultado del clasificador de audio 200 no cambia.

Si el estado actual es el tipo A y el resultado del clasificador de audio 200 se hace de tipo B, en tal caso, el temporizador 916 inicia la temporización o, según se ilustra en la Figura 10, el proceso entra en una fase de observación (la flecha (2) y un valor inicial del conteo de persistencia cnt se ajusta, indicando la magnitud de la duración de observación (igual al valor umbral).

A continuación, si el clasificador de audio 200 proporciona continuamente el tipo B, entonces, el valor de cnt disminuye continuamente (la flecha (3)) hasta que el valor de cnt es igual a 0 (es decir, la longitud del tiempo de duración del nuevo tipo B alcanza el valor umbral), entonces, la unidad de ajuste 300 puede utilizar el nuevo tipo de audio B (la flecha (4)) o, dicho de otro modo, solamente hasta ahora puede considerarse el tipo de audio que ha cambiado realmente al tipo B.

De no ser así, si antes de que el valor de cnt se haga cero (antes de que la longitud del tiempo de duración alcance el valor umbral), la salida del clasificador de audio 200 retorna al tipo A antiguo, luego se termina la fase de observación y la unidad de ajuste 300 sigue utilizando el tipo A antiguo (la flecha (5)).

El cambio desde el tipo B al tipo A puede ser similar al proceso anteriormente descrito.

En el proceso anterior, el valor umbral (o el conteo de persistencia) puede establecerse sobre la base del requisito de aplicación. Puede ser un valor fijo predefinido. Puede ser también un ajuste adaptativo. En una variante, el valor umbral es diferente para los distintos pares de transición desde un tipo de audio a otro tipo de audio. A modo de ejemplo, cuando cambia desde el tipo A al tipo B, el valor umbral puede ser un primer valor; y cuando se cambia desde el tipo B al tipo A, el valor umbral puede ser un segundo valor.

En otra variante, el conteo de persistencia (umbral) puede estar en correlación negativa con el valor de confianza del nuevo tipo de audio. La idea general es que, si el valor de confianza se muestra confuso entre dos tipos (p.ej., cuando el valor de confianza es solamente alrededor de 0.5), la duración de la observación necesita ser larga; de no ser así, la duración puede ser relativamente corta. Siguiendo esta directriz, un conteo de persistencia ejemplo puede establecerse por la fórmula siguiente,

$$HangCnt = C \cdot |0.5 - Conf| + D$$

en donde HangCnt es la duración persistencia o el valor umbral, C y D son dos parámetros que pueden ajustarse sobre la base del requisito de aplicación, siendo normalmente C un valor negativo mientras que D es un valor positivo.

A tal propósito, el temporizador 916 (y de este modo, el proceso de transición anteriormente descrito) ha sido descrito anteriormente como una parte del aparato de procesamiento de audio pero fuera del clasificador de audio 200. En algunas otras formas de realización, puede considerarse como una parte del clasificador de audio 200, según se describe en la Sección 7.3.

1.7 Combinación de formas de realización y escenarios de aplicación

Todas las formas de realización y las variantes, anteriormente descritas, pueden ponerse en práctica en cualquiera de sus combinaciones, y cualesquiera componentes mencionados en diferentes partes/formas de realización, pero teniendo las mismas o similares funciones, puede ponerse en práctica como los mismos componentes o componentes separados.

Más concretamente, cuando se describen las formas de realización y sus variantes en la presente descripción, los componentes que tienen señales de referencia similares a las ya descritas en formas de realización anteriores o variantes se omiten, y solamente se describen los componentes diferentes. De hecho, estas componentes diferentes pueden combinarse con los componentes de otras formas de realización o variantes, o constituir soluciones separadas por sí mismas. A modo de ejemplo, cualesquiera dos o más de las soluciones descritas con referencia a las Figuras 1 a 10, pueden combinarse entre sí. Como la solución más completa, los aparatos de procesamiento de audio pueden comprender el clasificador de contenido de audio 202 y el clasificador de contexto de audio 204, así como la unidad alisado de tipo 712, la unidad de alisado de parámetros 814 y el temporizador 916.

Según se indicó con anterioridad, los dispositivos de mejora de audio 400 pueden incluir el dispositivo de mejorador de diálogos 402, el virtualizador de sonido envolvente 404, el nivelador de volumen 406 y el ecualizador 408. Los aparatos de procesamiento de audio 100 pueden incluir cualquiera o más de ellos, con la unidad de ajuste 300 adaptada a los mismos. Cuando se implican múltiples dispositivos de mejora de audio 400, la unidad de ajuste 300 puede considerarse como incluyendo múltiples sub-unidades 300A a 300D (Figuras 15, 18, 20 y 22) específicas para los respectivos dispositivos de mejora de audio 400, o considerarse todavía como una sola unidad de ajuste unificada. Cuando es específico para un dispositivo de mejora de audio, la unidad de ajuste 300 junto con el clasificador de audio 200, así como otros posibles componentes, pueden considerarse como el controlador del dispositivo de mejora de audio específico, que se examinará en detalle en las Partes 2 a 5 siguientes.

Además, los dispositivos de mejora de audio 400 no están limitados a los ejemplos que se mencionan y pueden incluir cualquier otro dispositivo de mejora de audio.

Además, cualesquiera soluciones ya examinadas o cualquiera de sus combinaciones pueden combinarse, además, con cualquier forma de realización descrita o implícita en las otras Partes de esta idea inventiva. En particular, las formas de realización de los clasificadores de audio, según se describirá en las Partes 6 y 7, pueden utilizarse en los aparatos de procesamiento de audio.

1.8 Método de procesamiento de audio

En el proceso de descripción de los aparatos de procesamiento de audio en las presentes formas de realización, evidentemente se dan a conocer también algunos procesos o métodos. A continuación se proporciona un sumario de estos métodos sin repetir algunos de los datos ya descritos con anterioridad, pero conviene señalar que aunque los métodos se dan a conocer en el proceso de descripción de los aparatos de procesamiento de audio, los métodos no adoptan necesariamente los componentes según se describen o no se ejecutan necesariamente mediante dichos componentes. A modo de ejemplo, las formas de realización de los aparatos de procesamiento de audio pueden realizarse, parcial o completamente, con hardware y/o firmware, mientras que es posible que el método de procesamiento de audio descrito a continuación pueda realizarse totalmente mediante un programa ejecutable por ordenador, aunque los métodos pueden adoptar también el hardware y/o firmware de los aparatos de procesamiento de audio.

Los métodos se describirán a continuación haciendo referencia a las Figuras 11 a 14. Conviene señalar que en correspondencia con la propiedad de flujo continuo de la señal de audio, diversas operaciones se repiten cuando el método se pone en práctica en tiempo real, y diferentes operaciones no son necesarias con respecto a un mismo segmento de audio.

En una forma de realización, según se ilustra en la Figura 11, un método de procesamiento de audio se da a conocer. En primer lugar, la señal de audio a procesarse se clasifica en al menos un tipo de audio en tiempo real (operación 1102). Sobre la base del valor de confianza del al menos un tipo de audio, puede ajustarse continuamente al menos un parámetro para la mejora de audio (operación 1104). La mejora de audio puede ser una mejorador de diálogos (operación 1106), virtualizador de sonido envolvente (operación 1108), nivelador de volumen (1110) y/o ecualización (operación 1112). En correspondencia, el al menos un parámetro puede comprender al menos un parámetro para al menos uno de entre el procesamiento de mejorador de diálogos, el procesamiento de virtualización de la envolvente, el procesamiento de nivelación de volumen y el procesamiento de ecualización.

En este caso, los términos "en tiempo real" y "continuamente" significa el tipo de audio y en consecuencia, el parámetro cambiará en tiempo real con el contenido específico de la señal de audio y “continuamente” significa también que el ajuste es un ajuste continuo sobre la base del valor de confianza y no un ajuste brusco o discreto.

El tipo de audio puede comprender el tipo de contenido y/o el tipo de contexto. En correspondencia, la operación 1104 de ajuste puede configurarse para ajustar el al menos un parámetro sobre la base del valor de confianza de al menos un tipo de contenido y el valor de confianza de al menos un tipo de contexto. El tipo de contenido puede comprender, además, al menos uno de los tipos de contenidos de música a corto plazo, voz, sonido de fondo y ruido. El tipo de contexto puede comprender, además, al menos uno de entre los tipos de contextos de música a largo plazo, multimedia similar a cine, juego y VoIP.

Algunos otros sistemas de tipo de contexto son también propuestos, tales como tipos de contextos relacionados con VoIP incluyendo VoIP y no VoIP y los tipos de calidad de audio que incluyen audio de alta calidad o audio de baja calidad.

La música a corto plazo puede clasificarse también en sub-tipos en conformidad con normas diferentes. Dependiendo de la presencia de una fuente dominante, puede comprender música sin fuentes dominantes y música con fuentes dominantes. Además, la música a corto plazo puede comprender al menos una agrupación basada en el género o al menos una agrupación basada en los instrumentos o al menos una agrupación musical clasificada sobre la base del ritmo, *tempo*, timbre de música y/o cualesquiera otros atributos musicales.

Cuando se identifican los tipos de contenidos y los tipos de contextos, la importancia de un tipo de contenido puede determinarse por el tipo de contexto en donde está situado el tipo de contenido. Es decir, el tipo de contenido en una señal de audio de un tipo de contexto diferente se le asigna una ponderación diferente dependiendo del tipo de contexto de la señal de audio. Más en general, un tipo de audio puede influir o puede ser una premisa de otro tipo de audio. Por lo tanto, la operación de ajustar 1104 puede configurarse para modificar la ponderación del tipo de audio con el valor de confianza de otro tipo de audio.

Cuando una señal de audio se clasifica en múltiples tipos de audio al mismo tiempo (es decir, con respecto al mismo segmento de audio), la operación de ajustar 1104 puede considerar algunos o la totalidad de los tipos de audio identificados para ajustar los parámetros para mejorar ese segmento de audio. A modo de ejemplo, la operación de ajuste 1104 puede configurarse para la ponderación de los valores de confianza de los al menos un tipo de audio sobre la base de la importancia del al menos un tipo de audio. O bien, la operación de ajuste 1104 puede configurarse para considerar al menos algunos de los tipos de audio mediante su ponderación sobre la base de sus valores de confianza. En un caso especial, la operación de ajuste 1104 puede configurarse para considerar el al menos un tipo de audio dominante sobre la base de los valores de confianza.

Para evitar cambios bruscos de los resultados, pueden introducirse sistemas de alisado.

El valor de parámetro ajustado puede ser alisado (operación 1214 en la Figura 12). A modo de ejemplo, el valor de parámetro determinado por la operación de ajuste 1104 en el momento presente puede sustituirse con una suma ponderada del valor de parámetro determinado por la operación de ajuste en el momento actual y un valor de parámetro alisado de la última vez. De este modo, mediante la operación de alisado en iteración, el valor del parámetro es objeto de alisado en la línea de tiempos.

Los valores de ponderación para calcular la suma ponderada pueden cambiarse, de forma adaptativa, sobre la base del tipo de audio de la señal de audio o basarse en diferentes pares de transición desde un tipo de audio a otro tipo de audio. Como alternativa, los valores de ponderación para calcular la suma ponderada se cambian, de forma adaptativa sobre la base de un aumento o disminución en la tendencia del valor de parámetro determinado por la operación de ajuste.

Otro sistema de alisado se ilustra en la Figura 13. Es decir, el método puede comprender, además, para cada tipo de audio, el alisado del valor de confianza de la señal de audio en el momento actual calculando una suma ponderada del valor de confianza real en el presente y un valor de confianza alisado de la última vez (operación 1303). De modo similar a la operación de alisado de parámetros 1214, los valores de ponderación para calcular la suma ponderada pueden cambiarse, de forma adaptativa, sobre la base del valor de confianza del tipo de audio de la señal de audio o basarse en diferentes pares de transición desde un tipo de audio a otro tipo de audio.

Otro sistema de alisado es un mecanismo de memorización intermedia para retardar la transición desde un tipo de audio a otro tipo de audio, aun cuando cambie la salida de la operación de clasificación de audio 1102. Es decir, la operación de ajuste 1104 no utiliza el nuevo tipo de audio de forma inmediata, sino que espera a la estabilización de la salida de la operación de clasificación de audio 1102.

Más concretamente, el método puede comprender la medición del tiempo de duración durante el cual la operación de clasificación proporciona continuamente el mismo nuevo tipo de audio (operación 1403 en la Figura 14), en donde la operación de ajuste 1104 está configurada para seguir utilizando el presente tipo de audio ("N" en la operación 14035 y en la operación 11041) hasta que la longitud del tiempo de duración del nuevo tipo de audio alcance un valor umbral ("Y"

en la operación 14035 y en la operación 11042). Más concretamente, cuando la salida del tipo de audio desde la operación de clasificación de audio 1102 cambia con respecto al tipo de audio actual utilizado en la operación de ajuste de parámetros de audio 1104 ("Y" en la operación 14031), entonces se inicia la temporización (operación 14032). Si la operación de clasificación de audio 1102 sigue proporcionando el nuevo tipo de audio, es decir, si la determinación en la operación 14031 sigue siendo "Y", entonces continua la temporización (operación 14032). Por último, cuando el tiempo de duración del nuevo tipo de audio alcanza un valor umbral ("Y" en la operación 14035), la operación de ajuste 1104 utiliza el nuevo tipo de audio (operación 11042), y la temporización es objeto de restablecimiento (operación 14034) para preparar la siguiente conmutación del tipo de audio. Antes de alcanzar el valor umbral ("N" en la operación 14035), la operación de ajuste 1104 sigue utilizando el tipo de audio actual (operación 11041).

En este caso, la temporización puede ponerse en práctica con el mecanismo de un temporizador (conteo ascendente o conteo descendente). Si después de que se inicie la temporización, pero antes de que alcance el valor umbral, la salida de la operación de clasificación de audio 1102 se hace retornar al tipo de audio actual utilizado en la operación de ajuste 1104, debe considerarse que no existe ningún cambio ("N" en la operación 14031) con respecto al tipo de audio actual utilizado por la operación de ajuste 1104. Pero el resultado de clasificación actual (correspondiente al segmento de audio presente a clasificarse en la señal de audio) cambia con respecto a la salida anterior (correspondiente al segmento de audio anterior a clasificarse en la señal de audio) de la operación de clasificación de audio 1102 ("Y" en la operación 14033), de modo que la temporización se restablece (operación 14034), hasta que el siguiente cambio ("Y" en la operación 14031) inicie la temporización. Por supuesto, si el resultado de la clasificación de la operación de clasificación de audio 1102 no cambia con respecto al tipo de audio actual utilizado por la operación de ajuste de parámetros de audio 1104 ("N" en la operación 14031), ni cambia con respecto a la clasificación anterior ("N" en la operación 14033), ello demuestra que la clasificación de audio está en un estado estable y se sigue utilizando el tipo de audio actual.

El valor umbral aquí utilizado puede ser también diferente para distintos parámetros de transición desde un tipo de audio a otro tipo de audio, porque cuando el estado no es estable, en general podemos preferir que el dispositivo de mejora de audio esté en sus condiciones por defecto en lugar de en otras. Por otro lado, si el valor de confianza del nuevo tipo de audio es relativamente alto, es más seguro transitar al nuevo tipo de audio. Por lo tanto, el valor umbral puede estar en correlación negativa con el valor de confianza del nuevo tipo de audio. Cuanto más alto sea el valor de confianza, tanto más bajo será el valor umbral, lo que significa que el tipo de audio puede transitar al nuevo más rápido.

De modo similar a las formas de realización de los aparatos de procesamiento de audio, cualquier combinación de las formas de realización del método de procesamiento de audio y sus variantes son prácticas por un lado y por el otro lado, cualquier aspecto de las formas de realización del método de procesamiento de audio y sus variantes pueden ser soluciones separadas. En particular, en todos los métodos de procesamiento de audio, los métodos de clasificación de audio según se describen en las Partes 6 y 7 pueden utilizarse a este respecto.

Parte 2: Controlador del dispositivo de mejorador de diálogos y método de control

Un ejemplo del dispositivo de mejora de audio es el dispositivo mejorador de diálogos (DE), cuyo objetivo es controlar continuamente el audio en la reproducción, detectar la presencia de diálogo y mejorar el diálogo para aumentar su claridad e inteligibilidad (haciendo el diálogo más fácil de oír y entender), en particular las personas de edad avanzada con disminución de su capacidad auditiva. Además de detectar si un diálogo está presente, las frecuencias más importantes para la inteligibilidad son detectadas también si un diálogo está presente y luego, se mejora en correspondencia (con un reequilibrio espectral dinámico). Un método de mejora de diálogo ejemplo se presenta en H. Muesch. *"Mejora de la voz en un audio de entretenimiento"* publicada como documento WO 2008/106036 A2:, que se incorpora aquí en su integridad por referencia.

Una configuración manual común del dispositivo de mejora de diálogo es que se suele activar en contenidos de multimedia similar a cine pero desactivar en un contenido de música, puesto que la mejora del diálogo puede iniciar falsamente demasiado en señales musicales.

Con la información del tipo de audio disponible, el nivel de mejora de diálogo y otros parámetros pueden ajustarse sobre la base del valor de confianza de los tipos de audio identificados. Como un ejemplo específico de los aparatos de procesamiento de audio y del método anteriormente descrito, el dispositivo de mejora de diálogo puede hacer uso de todas las formas de realización examinadas en la Parte 1 y de cualesquiera combinaciones de dichas formas de realización. Más concretamente, en el caso de controlar el dispositivo de mejora de diálogo, el clasificador de audio 200 y la unidad de ajuste 300 en el aparato de procesamiento de audio 100 según se ilustra en las Figuras 1 a 10 pueden constituir un controlador de dispositivo de mejora de diálogo 1500 según se ilustra en la Figura 15. En esta forma de realización, puesto que la unidad de ajuste es específica para el dispositivo de mejora de diálogo, puede referirse 300A. Y, según se describe en la parte anterior, el clasificador de audio 200 puede comprender al menos uno de entre el clasificador de contenido de audio 202 y el clasificador de contexto de audio 204 y el controlador del dispositivo de mejora de diálogo 1500 puede comprender, además, al menos uno de entre la unidad de alisado de tipo 712, la unidad de alisado de parámetros 814 y el temporizador 916.

Por lo tanto, en esta parte, no repetiremos los contenidos ya descritos en la parte anterior y simplemente proporcionaremos algunos de sus ejemplos específicos.

Para un dispositivo de mejora de diálogo, los parámetros ajustables incluyen, sin limitación, el nivel de mejora de diálogo, el nivel de fondo y los umbrales para determinar las bandas de frecuencias que han de mejorarse. Véase el documento de H. Muesch. *"Mejora de la voz en audio de entretenimiento"*, publicada como documento WO 2008/106036 A2:, que se incorpora aquí en su integridad por referencia.

2.1 Nivel de mejora de los diálogos

Cuando se implica el nivel de mejora de diálogo, la unidad de ajuste 300A puede configurarse para establecer una correlación positiva del nivel de mejora de diálogo del dispositivo de mejora de diálogo con el valor de confianza de la voz. Además o como alternativa, el nivel puede estar en correlación negativa para el valor de confianza de los otros tipos de contenidos. De este modo, el nivel de mejora de diálogo puede establecerse para ser proporcional (de forma lineal o no lineal) con la confianza de la voz, de modo que la mejora de diálogo sea menos efectiva en señales sin voz, tales como música y sonido de fondo (efectos sonoros).

En cuanto al tipo de contexto, la unidad de ajuste 300A puede configurarse para establecer una correlación positiva del nivel de mejora de diálogo del dispositivo de mejora de diálogo con el valor de confianza de multimedia similar a cine y/o VoIP, y/o una correlación negativa del nivel de mejora de diálogo del dispositivo de mejora de diálogo con el valor de confianza de la música a largo plazo y/o juego. A modo de ejemplo, el nivel de mejora de diálogo puede establecerse para ser proporcional (en forma lineal o no lineal) con el valor de confianza de multimedia similar a cine. Cuando el valor de confianza de multimedia similar a cine es 0 (p.ej., en el contenido de música), el nivel de mejora de diálogo es también 0, lo que es equivalente a desactivar el dispositivo de mejora de diálogo.

Según se describe en la parte anterior, el tipo de contenido y el tipo de contexto pueden considerarse conjuntamente.

2.2 Umbrales para determinar las bandas de frecuencias a mejorarse

Durante el funcionamiento del dispositivo de mejora de diálogo, existe un valor umbral (normalmente umbral de energía o de intensidad) para cada banda de frecuencias para determinar si necesita mejorarse, es decir, las bandas de frecuencia por encima de los respectivos valores umbral de energía/intensidad deberán mejorarse. Para ajustar los valores umbral, la unidad de ajuste 300A puede configurarse para una correlación positiva de los valores umbral con un valor de confianza de la música a corto plazo y/o ruido y/o sonidos de fondo y/o correlación negativa de los valores umbral con un valor de confianza de la voz. A modo de ejemplo, los valores umbral pueden disminuirse si la confianza de la voz es alta, suponiendo una detección de la voz más fiable, para permitir que se mejoren más bandas de frecuencias; por el contrario, cuando el valor de confianza de la música es alto, los valores umbral pueden aumentarse para hacer que se mejoren menos bandas de frecuencias (y de este modo, menor presencia de artefactos).

2.3 Ajuste al nivel de fondo

Otra componente en el dispositivo de mejora del diálogo es la unidad de seguimiento mínimo 4022, según se ilustra en la Figura 15, que se utiliza para la estimación del nivel de fondo en la señal de audio (para estimación de la relación de señal a ruido SNR, y la estimación del umbral de bandas de frecuencia según se menciona en la Sección 2.2). Puede sintonizarse también sobre la base de los valores de confianza de tipos de contenidos de audio. A modo de ejemplo, si la confianza de la voz es alta, la unidad de seguimiento mínimo puede ser más fiable para establecer el nivel de fondo al mínimo actual. Si la confianza de la música es alta, el nivel de fondo puede establecerse a un valor algo más alto que el mínimo actual, o en otra manera, establecerlo a una media ponderada del mínimo actual y de la energía de la trama actual, con una ponderación grande sobre el mínimo actual. Si la confianza del fondo y del ruido es alta, el nivel de fondo puede establecerse mucho más elevado que el valor mínimo actual o de otra manera, establecerse a una media ponderada del mínimo actual y la energía de la trama actual, con una pequeña ponderación del mínimo actual.

De este modo, la unidad de ajuste 300A puede configurarse para asignar un ajuste a un nivel de fondo estimado por la unidad de seguimiento mínimo, en donde la unidad de ajuste está configurada, además, para la correlación positiva del ajuste con un valor de confianza de música a corto plazo y/o ruido y/o sonido de fondo y/o en correlación negativa con el ajuste con un valor de confianza de la voz. En una variante, la unidad de ajuste 300A puede configurarse para establecer la correlación del ajuste con el valor de confianza del ruido y/o fondo de forma más positiva que la música a corto plazo.

2.4 Combinación de formas de realización y escenarios de aplicación

De forma similar a lo establecido en la Parte 1, todas las formas de realización y variantes anteriormente descritas pueden ponerse en práctica en cualquiera de sus combinaciones y cualesquiera componentes mencionados en diferentes partes o formas de realización, pero teniendo las mismas o similares funciones, pueden ponerse en práctica como los mismos o componentes separados.

A modo de ejemplo, cualesquiera dos o más de las soluciones descritas en las secciones 2.1 a 2.3 pueden combinarse entre sí. Y estas combinaciones pueden combinarse, además, con cualquier forma de realización descrita o implícita en la Parte 1 y las otras Partes que se describirán más adelante. En particular, numerosas fórmulas son realmente

aplicables a cada clase de dispositivo de mejora de audio o su método correspondiente, pero no son necesariamente descritas o citadas en cada parte de esta idea inventiva. En tal situación, puede hacerse una referencia cruzada entre las Partes de esta idea inventiva para aplicar una fórmula específica incluida en una parte a otra parte, con solamente el ajuste adecuado de los parámetros, coeficientes, potencias (exponentes) y ponderaciones pertinentes en conformidad con los requisitos específicos de la aplicación concreta.

2.5 Método de control del dispositivo de mejora de diálogos

De forma similar a la Parte 1, en el proceso de describir el controlador del dispositivo de mejora de diálogo en las formas de realización aquí incluidas, se dan a conocer también algunos procesos o métodos. En este caso, se proporciona un resumen de estos métodos sin repetir algunos de los datos ya descritos con anterioridad.

En primer lugar, las formas de realización del método de procesamiento de audio según se describe en la Parte 1 pueden utilizarse para un dispositivo de mejora de diálogo, cuyos parámetros es uno de los objetivos que han de ajustase mediante el método de procesamiento de audio. Desde este punto de vista, el método de procesamiento de audio es también un método control del dispositivo de mejora de diálogo.

En esta sección, solamente se describirán aspectos específicos para el control del dispositivo de mejora de diálogo. Para los aspectos generales del método de control, puede hacerse referencia a la Parte 1.

En conformidad con una forma de realización, el método de procesamiento de audio puede comprender, además, un procesamiento de mejora de diálogo y la operación de ajuste 1104 comprende una correlación positiva del nivel de mejora de diálogo con el valor de confianza de multimedia similar a cine y/o VoIP, y/o una correlación negativa del nivel de mejora de diálogo con el valor de confianza de la música a largo plazo y/o juego. Es decir, la mejora de diálogo está principalmente orientada a la señal de audio en el contexto de multimedia similar a cine o VoIP.

Más concretamente, la operación de ajuste 1104 puede comprender la correlación positiva del nivel de mejora de diálogo del dispositivo de mejora de diálogo con el valor de confianza de la voz.

La presente idea inventiva puede ajustar también las bandas de frecuencias que han de mejorarse en el procesamiento de mejora de diálogo. Según se ilustra en la Figura 16, los valores umbrales (normalmente energía o intensidad) para determinar si las bandas de frecuencia respectivas que han de mejorarse pueden ajustarse sobre la base de los valores de confianza de los tipos de audio identificados (operación 1602) de conformidad con la presente idea inventiva. A continuación, dentro del dispositivo de mejora de diálogo, sobre la base de los valores umbrales ajustados, se seleccionan (operación 1604) y se mejoran (operación 1606) las bandas de frecuencias por encima de los valores umbrales respectivos.

Más concretamente, la operación de ajuste 1104 puede incluir la correlación positiva de los valores umbrales con un valor de confianza de música a corto plazo y/o ruido y/o sonido de fondo y/o una correlación negativa de los valores umbrales con una valor de confianza de la voz.

El método de procesamiento de audio (en particular, el procesamiento de mejora de diálogo) suele comprender, además, la estimación del nivel de fondo en la señal de audio, que se suele realizar mediante una unidad de seguimiento mínimo 4022 realizada en el dispositivo de mejora de diálogo 402 y utilizada en la estimación de la relación señal a ruido SNR o la estimación del valor umbral de las bandas de frecuencias. La presente idea inventiva puede utilizarse también para ajustar el nivel de fondo. En tal situación, después de que se estime el nivel de fondo (operación 1702), se ajusta primero sobre la base de los valores de confianza de los tipos de audio (operación 1704), y luego, se utiliza en la estimación de la relación SNR y/o estimación de valores umbrales de bandas de frecuencias (operación 1706). Más concretamente, la operación de ajuste 1104 puede configurarse para asignar un ajuste al nivel de fondo estimado, en donde la operación de ajuste 1104 puede configurarse, además, para una correlación positiva del ajuste con un valor de confianza de una música a corto plazo y/o ruido y/o sonido de fondo y/o una correlación negativa del ajuste con un valor de confianza de la voz.

Más concretamente, la operación de ajuste 1104 puede configurarse para establecer una correlación del ajuste con el valor de confianza del ruido y/o fondo de forma más positiva que la música a corto plazo.

De modo similar a las formas de realización del aparato de procesamiento de audio, cualquier combinación de las formas de realización del método de procesamiento de audio y sus variantes son de efecto práctico por un lado; y por el otro lado, cada aspecto de las formas de realización del método de procesamiento de audio y sus variantes pueden ser soluciones separadas. Además, cualesquiera dos o más soluciones descritas en esta sección pueden combinarse entre sí, y estas combinaciones pueden combinarse, además, con cualquier forma de realización descrita o implícita en la Parte 1 y las otras partes que se describirán más adelante.

Parte 3: Controlador del virtualizador de sonido envolvente y método de control

Un virtualizador de sonido envolvente permite que una señal sonora envolvente (tal como los multicanales 5.1 y 7.1) se

presente a través de los altavoces internos del PC o a través de auriculares. Es decir, con dispositivos estéreo tales como altavoces o auriculares de ordenador portátil internos, crea un efecto de envolvente virtual y proporciona una experiencia cinemática para los consumidores. Las denominadas Funciones de Transferencias Relacionadas con los Auriculares (HRTFs) se suelen utilizar en el virtualizador de sonido envolvente para simular la llegada de sonido en los oídos procedente de las diversas ubicaciones de altavoces asociadas con la señal de audio multicanal.

Aunque el virtualizador de sonido envolvente actual funciona adecuadamente con los auriculares, funciona de forma distinta con diferentes contenidos con los altavoces incorporados. En general, el contenido de multimedia similar a cine permite que el virtualizador de sonido envolvente se active para altavoces, aunque la música es posible que no suelen con intensidad.

Puesto que los mismos parámetros en el virtualizador de sonido envolvente no pueden crear una buena imagen acústica para el contenido de multimedia similar a cine y el contenido musical simultáneamente, los parámetros han de ajustarse sobre la base del contenido de forma más precisa. Con la información del tipo de audio disponible, en particular, el valor de confianza de la música y el valor de confianza de la voz, así como algunas otras informaciones del tipo de contenido y del tipo de contexto, puede conseguirse un funcionamiento adecuado con la presente idea inventiva.

De modo similar a la Parte 2, a modo de un ejemplo específico del aparato de procesamiento de audio y del método descrito en la Parte 1, el virtualizador de sonido envolvente 404 puede hacer uso de la totalidad de las formas de realización descritas en la Parte 1 y cualesquiera combinaciones de dichas formas de realización que en dicha parte se dan a conocer. Más concretamente, en el caso de control del virtualizador de sonido envolvente 404, el clasificador de audio 200 y la unidad de ajuste 300 en el aparato de procesamiento de audio 100 según se ilustra en las Figuras 1 a 10, pueden constituir un controlador de virtualizador de sonido envolvente 1800 según se ilustra en la Figura 18. En esta forma de realización, puesto que la unidad de ajuste es específica para el virtualizador de sonido envolvente 40, puede referirse como 300B. Y, de modo similar a la Parte 2, el clasificador de audio 200 puede comprender al menos uno de entre el clasificador de contenido de audio 202 y el clasificador de contexto de audio 204, y el controlador de virtualizador de sonido envolvente 1800 puede comprender, además, al menos uno entre la unidad de alisado de tipo 712, la unidad de alisado de parámetros 814 y el temporizador 916.

Por lo tanto, en esta Parte, no repetiremos los contenidos ya descritos en la Parte 1, y simplemente proporcionaremos algunos de sus ejemplos específicos.

Para un virtualizador de sonido envolvente, los parámetros ajustables incluyen, sin limitación, a la magnitud reforzadora de envolvente y la frecuencia inicial para el virtualizador de sonido envolvente 404.

3.1 Magnitud de refuerzo de la envolvente

Cuando está implicada la magnitud de refuerzo de la envolvente, la unidad de ajuste 300B puede configurarse para una correlación positiva de la magnitud de refuerzo de la envolvente del virtualizador de sonido envolvente 404 con un valor de confianza del ruido y/o fondo y/o voz y/o una correlación negativa de la magnitud de refuerzo de la envolvente con un valor de confianza de la música a corto plazo.

Más concretamente, para modificar el virtualizador de sonido envolvente 404 con el fin de que la música (tipo de contenido) suene de forma aceptable, una puesta en práctica a modo de ejemplo de la unidad de ajuste 300B podría ajustar la magnitud de refuerzo de la envolvente sobre la base del valor de confianza de la música a corto plazo, tal como:

$$SB \propto (1 - Conf_{music}) \qquad (5)$$

en donde SB indica la magnitud de refuerzo de la envolvente, $Conf_{music}$ es el valor de confianza de la música a corto plazo.

Sirve de ayuda para disminuir el refuerzo de la envolvente para la música y evitar que suene en la forma de 'lavado':

De modo similar, el valor de confianza de la voz puede utilizarse también, a modo de ejemplo:

$$SB \propto (1 - Conf_{music}) * Conf_{speech}^{\alpha} \qquad (6)$$

en donde $Conf_{speech}$ es el valor de confianza de la voz, α es el coeficiente de ponderación en la forma de exponente, y puede estar en el margen de 1-2. Esta fórmula indica que la magnitud de refuerzo de la envolvente será alta para solamente la voz pura (confianza de la voz alta y confianza de la música baja).

O bien, podemos considerar solamente el valor de confianza de la voz:

$$SB \propto Conf_{speech} \qquad (7)$$

Diversas variantes pueden diseñarse de una forma similar. En particular, para el ruido o el sonido de fondo, pueden establecerse fórmulas similares a la fórmula (5) a (7). Además, los efectos de los cuatros tipos de contenidos pueden considerarse juntos en cualquier combinación. En tal situación, el ruido y el sonido de fondo son sonidos ambientales y son más seguros para tener una gran magnitud de refuerzo; la voz puede tener una magnitud de refuerzo media, suponiendo que la persona que habla suele estar sentada frente a la pantalla; y la música utiliza menos magnitud de refuerzo. Por lo tanto, la unidad de ajuste 300B puede configurarse para establecer una correlación de la magnitud de refuerzo de la envolvente con el valor de confianza del ruido y/o fondo de forma más positiva que el contenido del tipo de voz.

Suponiendo que definimos previamente una magnitud de refuerzo prevista (que es equivalente a una ponderación) para cada tipo de contenido, se puede aplicar también otra alternativa:

$$\hat{a} = \frac{a_{speech} \cdot Conf_{speech} + a_{music} \cdot Conf_{music} + a_{noise} \cdot Conf_{noise} + a_{bkg} \cdot Conf_{bkg}}{Conf_{speech} + Conf_{music} + Conf_{noise} + Conf_{bkg}} \qquad (8)$$

en donde â es una magnitud de refuerzo estimada, α con un subíndice del tipo de contenido es la magnitud de refuerzo (ponderación) prevista/predefinida del tipo de contenido, *Conf* con un subíndice del tipo de contenido es el valor de confianza del tipo de contenido (en donde bkg representa al "sonido de fondo"). Dependiendo de las situaciones, a_{music} puede (pero no necesariamente) establecerse como 0, lo que indica que el virtualizador de sonido envolvente 404 estará desactivado para la música pura (tipo de contenido).

Desde otro punto de vista, el valor de α con un subíndice del tipo de contenido en la fórmula (8) es la magnitud de refuerzo prevista/predefinida del tipo de contenido, y el cociente del valor de confianza del tipo de contenido correspondiente dividido por la suma de los valores de confianza de todos los tipos de contenidos identificados puede considerarse como una ponderación normalizada de la magnitud de refuerzo predefinida/prevista del tipo de contenido correspondiente. Es decir, la unidad de ajuste 300B puede configurarse para considerar al menos algunos de los múltiples tipos de contenidos mediante ponderación de las magnitudes de refuerzo predefinidas de los múltiples tipos de contenidos sobre la base de los valores de confianza.

En cuanto el tipo de contexto, la unidad de ajuste 300B puede configurarse para una correlación positiva de la magnitud de refuerzo de la envolvente del virtualizador de sonido envolvente 404 con un valor de confianza de multimedia similar a cine y/o juego, y/o una correlación negativa de la magnitud de refuerzo de la envolvente con un valor de confianza de la música a largo plazo y/o VoIP. A continuación, pueden establecerse las fórmulas similares a (5) a (8).

A modo de ejemplo especial, el virtualizador de sonido envolvente 404 puede activarse para una condición pura de multimedia similar a cine y/o juego, pero desactivarse para música y/o VoIP. Asimismo, la magnitud de refuerzo del virtualizador de sonido envolvente 404 puede establecerse de forma diferente para multimedia similar a cine y juego. En multimedia similar a cine se utiliza una magnitud de refuerzo alta y el juego utiliza menos. Por lo tanto, la unidad de ajuste 300B puede configurarse para establecer una correlación de la magnitud de refuerzo de la envolvente con el valor de confianza de multimedia similar a cine de forma más positiva que con el juego.

De modo similar, al tipo de contenido, la magnitud de la envolvente de una señal de audio puede establecerse también a un valor medio ponderado de los valores de confianza de los tipos de contextos:

$$\hat{a} = \frac{a_{MOVIE} \cdot Conf_{MOVIE} + a_{MUSIC} \cdot Conf_{MUSIC} + a_{GAME} \cdot Conf_{GAME} + a_{VOIP} \cdot Conf_{VOIP}}{Conf_{MOVIE} + Conf_{MUSIC} + Conf_{GAME} + Conf_{VOIP}} \qquad (9)$$

en donde *â* es la magnitud de refuerzo estimada, α con un subíndice del tipo de contexto es la magnitud de refuerzo prevista/predefinida (ponderación) del tipo de contexto, *Conf* con un subíndice del tipo de contexto es el valor de confianza del tipo de contexto. Dependiendo de las situaciones, a_{MUSIC} y a_{VOIP} pueden (pero no necesariamente) establecerse como 0, lo que indica que el virtualizador de sonido envolvente 404 será desactivado para música pura (tipo de contexto) y/o VoIP pura.

De nuevo, de modo similar al tipo de contenido, el valor de α con un subíndice del tipo de contexto en la fórmula (9) es la magnitud de refuerzo prevista/predefinida del tipo de contexto y el cociente del valor de confianza del tipo de contexto correspondiente dividido por la suma de los valores de confianza de todos los tipos de contextos identificados puede considerarse como una ponderación normalizada de la magnitud de refuerzo predefinida/prevista del tipo de contexto correspondiente. Es decir, la unidad de ajuste 300B puede configurarse para considerar al menos algunos de los

múltiples tipos de contextos mediante la ponderación la magnitud de refuerzo predefinida de los múltiples tipos de contextos sobre la base de los valores de confianza.

3.2 Frecuencia inicial

Otros parámetros pueden modificarse también en el virtualizador de sonido envolvente, tal como la frecuencia inicial. En general, los componentes de alta frecuencia en una señal de audio son más adecuados para presentarse de forma espacial. A modo de ejemplo, en la música, el bajo se presenta espacialmente para tener más efectos de envolvente. Por lo tanto, para una señal de audio específica, el virtualizador de sonido envolvente necesita determinar un valor umbral de frecuencia, con los componentes por encima de dicho valor presentados espacialmente mientras se retienen los componentes inferiores. El valor umbral de frecuencia es la frecuencia inicial.

En conformidad con una forma de realización de la presente idea inventiva, la frecuencia inicial para el virtualizador de sonido envolvente puede aumentarse en el contenido de música de modo que se pueda retener más bajos para las señales musicales. A continuación, la unidad de ajuste 300B puede configurarse para establecer una correlación positiva de la frecuencia inicial del virtualizador de sonido envolvente con un valor de confianza de la música a corto plazo.

3.3 Combinación de formas de realización y escenarios de aplicación

De forma similar a la Parte 1, todas las formas de realización y variantes anteriormente descritas pueden ponerse en práctica en cualquiera de sus combinaciones, y cualesquiera componentes que se mencionen en diferentes partes/formas de realización pero teniendo las mismas o funciones similares, pueden ponerse en práctica como los mismos o componentes separados.

A modo de ejemplo, cualquiera dos o más de las soluciones descritas en las secciones 3.1 y 3.2 pueden combinarse entre sí. Y cualquiera de las combinaciones puede combinarse, además, con cualquier forma de realización descrita o implícita en la Parte 1, Parte 2 y más demás partes que se describirán más adelante.

3.4 Método de control del virtualizador de sonido envolvente

De forma similar a la Parte 1, en el proceso de describir el controlador del virtualizador de sonido envolvente en las formas de realización aquí descritas, evidentemente son también algunos procesos o métodos pertinentes. Más adelante se proporciona un sumario de estos métodos sin repetir algunos de los detalles ya descritos con anterioridad.

En primer lugar, las formas de realización del método de procesamiento de audio según se describe en la Parte 1 pueden utilizarse para un virtualizador de sonido envolvente, cuyos parámetros es uno de los objetivos que han de ajustarse por el método de procesamiento de audio. Desde este punto de vista, el método de procesamiento de audio es también un método de control del virtualizador de sonido envolvente.

En esta sección, solamente se describirán los aspectos específicos para el control del virtualizador de sonido envolvente. Para aspectos generales del método de control, puede hacerse referencia a la Parte 1.

En conformidad con una forma de realización, el método de procesamiento de audio puede comprender, además, un procesamiento de virtualización de envolvente, y la operación de ajuste 1104 puede configurarse para establecer una correlación positiva de la magnitud de refuerzo de envolvente del procesamiento de virtualización de envolvente con un valor de confianza del ruido y/o fondo y/o voz y/o una correlación negativa de la magnitud de refuerzo de envolvente con un valor de confianza de la música a corto plazo.

Más concretamente, la operación de ajuste 1104 puede configurarse para establecer una correlación de la magnitud de refuerzo de envolvente con el valor de confianza de ruido y/o sonido de fondo de forma más positiva que con el tipo de contexto de voz.

Como alternativa o de forma adicional, la magnitud de refuerzo de envolvente puede ajustarse también sobre la base de los valores de confianza de los tipos de contextos. Más concretamente, la operación de ajuste 1104 puede configurarse para establecer una correlación positiva de la magnitud de refuerzo de envolvente del procesamiento de virtualización de envolvente con un valor de confianza de multimedia similar a cine y/o juego, y/o una correlación negativa de la magnitud de refuerzo de envolvente con un valor de confianza de la música a largo plazo y/o VoIP.

Más concretamente, la operación de ajuste 1104 puede configurarse para establecer una correlación de la magnitud de refuerzo de envolvente con el valor de confianza de multimedia similar a cine de forma más positiva que con el juego.

Otro parámetro a ajustarse es la frecuencia inicial para el procesamiento de virtualización de envolvente. Según se ilustra en la Figura 19, la frecuencia inicial se ajusta, en primer lugar, sobre la base de los valores de confianza de los tipos de audio (operación 1902), y a continuación, el virtualizador de sonido envolvente procesa las componentes de audio por encima de la frecuencia inicial (operación 1904). Más concretamente, la operación de ajuste 1104 puede configurarse para establecer una correlación positiva de la frecuencia inicial del procesamiento de virtualizador de sonido envolvente

con un valor de confianza de la música a corto plazo.

De modo similar a las formas de realización del aparato de procesamiento de audio, cualquier combinación de las formas de realización del método de procesamiento de audio y sus variantes son prácticas por un lado; y por el otro lado, cada aspecto de las formas de realización del método de procesamiento de audio y sus variantes pueden ser soluciones separadas. Además, cualesquiera dos o más soluciones descritas en esta sección pueden combinarse entre sí, y dichas combinaciones pueden combinarse, además, con cualquier forma de realización descrita o implícita en las otras partes de esta idea inventiva.

Parte 4: Controlador del nivelador de volumen y método de control

El volumen de diferentes fuentes de audio o diferentes elementos en la misma fuente de audio cambia mucho en algunas ocasiones. Es enojoso puesto que los usuarios han de ajustar con frecuencia el volumen. El dispositivo de nivelador de volumen (VL) tiene como objetivo ajustar el volumen del contenido de audio en la reproducción y mantenerlo casi constante a través de la línea de tiempo sobre la base de un valor de intensidad objetivo. Niveladores de volumen a modo de ejemplo se presentan en el documento A. J. Seefeldt et al. Titulado: *"Cálculo y ajuste de la intensidad percibida y/o el equilibrio espectral percibido de una señal de audio"*, publicada como US2009/0097676A1; B. G. Grockett et al. *"Control de la ganancia de audio utilizando detección de eventos auditivos sobre la base de la intensidad específica"*, publicado como WO2007/127023A1; y A. Seefeldt et al. *"Procesamiento de audio utilizando análisis de escenas auditivas y sesgado espectral"*, publicado como WO 2009/011827 A1. Los tres documentos se incorporan aquí en sus integridades por referencia.

El nivelador de volumen mide continuamente la intensidad de una señal de audio en alguna manera y luego, modifica la señal en una magnitud de *ganancia* que es un factor de escala para modificar la intensidad de una señal de audio y suele ser una función de la intensidad medida, la intensidad objetivo deseada y varios otros factores. Varios factores necesitan considerarse para la estimación de una ganancia adecuada, con criterios subyacentes para aproximarse a la intensidad objetivo y mantenerse el margen dinámico. Suele comprender varios subelementos tales como control automático de la ganancia (AGC), detección de evento auditorio, control del margen dinámico (DRC).

Una señal de control se suele aplicar en el nivelador de volumen para controlar la “ganancia” de la señal de audio. A modo de ejemplo, una señal de control puede ser un indicador del cambio en la magnitud de la señal de audio derivada mediante análisis de señales puras. Puede ser también un indicador de evento de audio para representar si aparece un nuevo evento de audio, mediante un análisis psico-acústico, tal como un análisis de escenas de auditorio o detección de eventos de auditorio sobre la base de una intensidad específica. Dicha señal de control se aplica en el nivelador de volumen para controlar la ganancia, a modo de ejemplo, asegurando que la ganancia sea casi constante dentro de un evento de auditorio y confinando gran parte del cambio de la ganancia a la proximidad de un límite de evento, con el fin de reducir la presencia de posibles artefactos audibles debido a un cambio rápido de la ganancia en la señal de audio.

Sin embargo, los métodos convencionales de derivación de señales de control no pueden diferenciar los eventos de auditorio informativos de los eventos de auditorio no informativos (interferente). En este caso, el evento de auditorio informativo significa el evento de audio que contiene información significativa y puede recibir más atención por parte de los usuarios, tal como diálogo y música, mientras que la señal no informativa no contiene información significativa para los usuarios, tal como ruido en VoIP. En consecuencia, las señales no informativas pueden aplicarse también mediante una ganancia de gran magnitud y reforzarse para cerrar la intensidad objetivo. Resultará desagradable en algunas aplicables. A modo de ejemplo, en las llamadas de VoIP, la señal de ruido que aparece en la pausa de una conversación se suele reforzar hasta un volumen intenso después de procesarse por un nivelador de volumen. Esto es indeseable por los usuarios.

Con el fin de resolver este problema al menos en parte, la presente idea inventiva propone el control del nivelador de volumen sobre la base de las formas de realización dadas a conocer en la Parte 1.

De modo similar a la Parte 2 y la Parte 3, como un ejemplo específico del aparato de procesamiento de audio y del método descritos en la Parte 1, el nivelador de volumen 406 puede hacer uso de la totalidad de las formas de realización descritas en la Parte 1 y cualesquiera combinaciones de dichas formas de realización aquí dadas a conocer. Más concretamente, en el caso de control del nivelador de volumen 406, el clasificador de audio 200 y la unidad de ajuste 300 en el aparato de procesamiento de audio 100 según se ilustra en las Figuras 1 a 10, puede constituir un controlador 2000 de nivelador de volumen 406 según se ilustra en la Figura 20. En esta forma de realización, puesto que la unidad de ajuste es específica para el nivelador de volumen 406, puede referirse como 300C.

Es decir, sobre la base de la descripción de la Parte 1, un controlador del nivelador de volumen 2000 puede comprender un clasificador de audio 200 para identificar continuamente el tipo de audio (tal como tipo de contenido y/o tipo de contexto) de una señal de audio; y una unidad de ajuste 300C para ajustar un nivelador de volumen en una manera continua sobre la base del valor de confianza del tipo de audio que se identifica. De modo similar, el clasificador de audio 200 puede comprender al menos uno de entre el clasificador de contenido de audio 202 y el clasificador de contexto de audio 204, y el controlador de nivelador de volumen 2000 puede comprender, además, al menos de entre la unidad de alisado de tipo 712, la unidad de alisado de parámetros 814 y el temporizador 916.

Por lo tanto, en esta Parte, no se repetirán los contenidos ya descritos en la Parte 1 y simplemente se proporcionarán algunos de sus ejemplos específicos.

Diferentes parámetros en el nivelador de volumen 406 pueden ajustarse de forma adaptativa, sobre la base de los resultados de clasificación. Se pueden ajustar los parámetros directamente relacionados con la ganancia dinámica o la gama de la ganancia dinámica, a modo de ejemplo, reduciendo la ganancia para las señales no informativas. Se pueden ajustar también los parámetros que indican el grado de la señal que es un nuevo evento de audio perceptible, y luego, controlar indirectamente, la ganancia dinámica (la ganancia cambiará lentamente dentro de un evento de audio, pero puede cambiar con rapidez en el límite de separación de dos eventos de audio). En esta aplicación, se presentan varias formas de realización del ajuste de parámetros o del mecanismo de control del nivelador de volumen.

4.1 Tipos de contenido informativo e interferente

Según se indicó con anterioridad, en relación con el control del nivelador de volumen, los tipos de contenidos de audio pueden clasificarse como tipos de contenidos informativos y tipos de contenidos interferentes. Y la unidad de ajuste 300C puede configurarse para establecer una correlación positiva de la ganancia dinámica del nivelador de volumen con los tipos de contenidos informativos de la señal de audio, y una correlación negativa de la ganancia dinámica del nivelador de volumen con tipos de contenidos interferentes de la señal de audio.

A modo de ejemplo, suponiendo que el ruido es interferente (no informativo) y será enojoso que se refuerce en un volumen mayor, el parámetro que controla directamente la ganancia dinámica o el parámetro que indica nuevos eventos de audio pueden establecerse para ser proporcionales a una función decreciente del valor de confianza del ruido ($Conf_{noise}$), tal como

$$\text{GainControl} \propto 1 - Conf_{noise} \qquad (10)$$

En este caso, para mayor simplicidad, se utiliza el símbolo GainControl para representar todos los parámetros (o sus efectos) relacionado con el control de la ganancia, en el nivelador de volumen, puesto que diferentes puestas en práctica del nivelador de volumen pueden utilizar distintos nombres de parámetros con diferente significado subyacente. La utilización del término único GainControl puede tener una expresión corta sin perder su generalidad. En esencia, el ajuste de estos parámetros es equivalente a aplicar una ponderación sobre la ganancia original, lineal o no lineal. A modo de ejemplo, el GainControl puede utilizarse directamente para poner a escala la ganancia de modo que la ganancia será pequeña si el valor de GainControl es pequeño. A modo de otro ejemplo específico, la ganancia está indirectamente controlada mediante la puesta a escala con GainControl de la señal de control de eventos descrita en el documento B.G. Grockett et al. *"Control de la ganancia de audio utilizando la detección de eventos de auditorio basados en una intensidad específica"*, publicado como documento WO2007/127023A1, que se incorpora aquí en su integridad por referencia. En este caso, cuando GainControl es de valor pequeño, los controles de la ganancia del nivelador de volumen se modifican para impedir que la ganancia cambie significativamente en el transcurso del tiempo. Cuando el valor de GainControl es alto, los controles se modifican de modo que la ganancia del nivelador de volumen sea permitido que cambie con más libertad.

Con el control de la ganancia descrito en la fórmula (10) (poniendo directamente a escala la ganancia original o la señal de control de eventos), la ganancia dinámica de una señal de audio se pone en correlación (lineal o no lineal) para su valor de confianza del ruido. Si la señal es ruido con un alto valor de confianza, la ganancia final será pequeña debido al factor (1-$Conf_{noise}$). De este modo, se impide reforzar una señal de ruido en un volumen intenso desagradable.

A modo de una variante ejemplo de la fórmula (10), si el sonido de fondo tampoco es de interés en una aplicación (tal como en VoIP), puede negociarse de forma similar y aplicarse también mediante una ganancia pequeña. Una función de control puede considerar, a la vez, el valor de confianza del ruido ($Conf_{noise}$) y el valor de confianza del fondo ($Conf_{bkg}$), a modo de ejemplo

$$\text{GainControl} \propto (1 - Conf_{noise}) \cdot (1 - Conf_{bkg}) \qquad (11)$$

En la fórmula anterior, puesto que no son deseados ni el ruido ni los sonidos de fondo, el valor de GainControl es igualmente afectado por el valor de confianza del ruido y el valor de confianza del fondo, y puede considerarse que el ruido y los sonidos de fondo tienen la misma ponderación. Dependiendo de las situaciones, pueden tener diferentes ponderaciones. A modo de ejemplo, puede proporcionarse los valores de confianza del ruido y del sonido de fondo (o su diferencia con 1) con diferentes coeficientes o diferentes exponentes (α e γ). Es decir, la fórmula (11) puede reescribirse como:

$$\text{GainControl} \propto (1 - Conf_{noise})^{\alpha} \cdot (1 - Conf_{bkg})^{\gamma} \qquad (12)$$

o

$$\text{GainControl} \propto (1 - Conf_{noise}{}^{\alpha}) \cdot (1 - Conf_{bkg}{}^{\gamma}) \qquad (13)$$

Como alternativa, la unidad de ajuste 300C puede configurarse para considerar al menos un tipo de contenido dominante sobre la base de los valores de confianza. A modo de ejemplo:

$$\text{GainControl} \propto 1 - \max(Conf_{noise}, Conf_{bkg}) \qquad (14)$$

Tanto la fórmula (11) (y sus variantes) y la fórmula (14) indican una ganancia pequeña para señales de ruido y señales de sonido de fondo, y el comportamiento original del nivelador de volumen se mantiene solamente con, a la vez, la confianza del ruido y la confianza del sonido de fondo que se mantenga en un pequeño valor (tal como en la voz y la señal musical) de modo que GainControl tenga un valor próximo a uno.

La realización ejemplo anterior ha de considerar el tipo de contenido interferente dominante. Dependiendo de la situación, la unidad de ajuste 300C puede configurarse también para considerar el tipo de contenido informativo dominante sobre la base de los valores de confianza. En términos más generales, la unidad de ajuste 300C puede configurarse para considerar al menos un tipo de contenido dominante sobre la base de los valores de confianza, sin importar que los tipos de audio identificados sean/incluyan tipos de audio informativos y/o tipos de audio interferentes.

A modo de otro ejemplo variante de la fórmula (10), se supone que la señal de voz es el contenido más informativo y necesita menos modificación sobre el comportamiento por defecto del nivelador de volumen, la función controladora puede considerar, a la vez, el valor de confianza del ruido ($Conf_{noise}$) y el valor de confianza de la voz ($Conf_{speech}$), como

$$\text{GainControl} \propto 1 - Conf_{noise} \cdot (1 - Conf_{speech}) \qquad (15)$$

Con esta función, se obtiene un pequeño valor de GainControl solamente para las señales con alto valor de confianza de ruido y bajo valor de confianza de la voz (p.ej., ruido puro) y el valor de GainControl estará próximo a 1 si el valor de confianza de la voz es alto (y de este modo, se mantiene el comportamiento original del nivelador de volumen). En términos más generales, puede considerarse que la ponderación de un tipo de contenido (tal como $Conf_{noise}$) puede modificarse con el valor de confianza de al menos otro tipo de contenido (tal como $Conf_{speech}$). En la fórmula (15) anterior, puede considerarse que la confianza de la voz cambia el coeficiente de ponderación de la confianza del ruido (otra clase de ponderación si se compara con las ponderaciones en las fórmulas (12 y 13)). Dicho de otro modo, en la fórmula (10), el coeficiente de $Conf_{noise}$ puede considerarse como 1; mientras que en la fórmula (15), algunos otros tipos de audio (tales como la voz, pero sin limitación) afectarán a la importancia del valor de confianza del ruido, por lo que se puede afirmar que la ponderación de $Conf_{noise}$ se modifica por el valor de confianza de la voz. Dentro del contexto de la presente idea inventiva, el término de "ponderación" se interpretará como que incluye esta circunstancia. Es decir, indica la importancia de un valor, pero no necesariamente normalizado. Puede hacerse referencia a la sección 1.4.

Desde otro punto de vista, similar a las fórmulas (12) y (13), ponderaciones en la forma de exponentes pueden aplicarse sobre los valores de confianza en la función anterior para indicar la prioridad (o importancia) de diferentes señales de audio, a modo de ejemplo, la fórmula (15) puede cambiarse a:

$$\text{GainControl} \propto 1 - Conf_{noise}{}^{\alpha} \cdot (1 - Conf_{speech})^{\gamma} \qquad (16)$$

en donde los valores de α y γ son dos ponderaciones, que pueden establecerse más pequeñas si está previsto que sean más sensibles para modificar los parámetros del nivelador.

Las fórmulas (10) a (16) pueden combinarse libremente para formar varias funciones de control que pueden ser adecuadas en diferentes aplicaciones. Los valores de confianza de otros tipos de contenidos de audio, tales como el valor de confianza de la música, pueden incorporarse también fácilmente en las funciones de control de una manera similar.

En el caso en donde el parámetro de GainControl se utiliza para ajustar los parámetros que indican el grado en que la señal se hace un nuevo evento de audio perceptible y luego, controlan indirectamente la ganancia dinámica (la ganancia cambiará lentamente dentro de un evento de audio pero puede cambiar con rapidez en el límite de dos eventos de audio), pudiendo considerarse que existe otra función de transferencia entre el valor de confianza de tipos de contenidos y la ganancia dinámica final.

4.2 Tipos de contenidos en diferentes contextos

Las funciones de control anteriores en la fórmula (10)-(16) toman en consideración los valores de confianza de tipos de contenidos de audio, tales como ruido, sonido de fondo, música a corto plazo y voz, pero no consideran sus contextos de audio en donde los sonidos de procedencia tal como multimedia similar a cine y VoIP. Es posible que el mismo tipo de

contenido de audio pudiera necesitar procesarse de forma distinta en contextos de audio diferentes, a modo de ejemplo, los sonidos de fondo. El sonido de fondo comprende varios sonidos tales como un motor de vehículo, explosión y aplausos. Puede no ser importante en una llamada de VoIP pero podría ser importante en un multimedia similar a cine. Esto indica que los contextos de audio interesados necesitan identificarse y diferentes funciones de control necesitan diseñarse para distintos contextos de audio.

Por lo tanto, la unidad de ajuste 300C puede configurarse para considerar el tipo de contenido de la señal de audio como informativo o interferente sobre la base del tipo de contexto de la señal de audio. A modo de ejemplo, considerando el valor de confianza del ruido y el valor de confianza del sonido de fondo, y diferenciando los contextos de VoIP y no VoIP, una función de control dependiente del contexto de audio puede ser:

Si el contexto de audio es VoIP

$$\text{GainControl} \propto 1 - \max(Conf_{noise}, Conf_{bkg})$$

de lo contrario (17)

$$\text{GainControl} \propto 1 - Conf_{noise}$$

Es decir, en el contexto de VoIP, el ruido y los sonidos de fondo se consideran como tipos de contenidos interferentes; mientras que el contexto de no VoIP, los sonidos de fondo se consideran como un tipo de contenido informativo.

A modo de otro ejemplo, una función de control dependiente del contexto de audio que considera los valores de confianza de la voz, ruido y sonido de fondo y diferenciando los contextos de VoIP y no VoIP, podría ser:

Si el contexto de audio es VoIP

$$\text{GainControl} \propto 1 - \max(Conf_{noise}, Conf_{bkg})$$

de lo contrario (18)

$$\text{GainControl} \propto 1 - Conf_{noise} \cdot (1 - Conf_{speech})$$

En este caso, la voz es resaltada como un tipo de contenido informativo.

Suponiendo que la música es también un tipo informativo importante en el contexto de no VoIP, se puede ampliar la segunda parte de la fórmula (18) a:

$$\text{GainControl} \propto 1 - Conf_{noise} \cdot (1 - \max(Conf_{speech}, Conf_{music})) \quad (19)$$

De hecho, cada una de las funciones de controles en (10)-(16) o sus variantes pueden aplicarse en contextos de audio distintos/correspondientes. De este modo, puede generarse un gran número de combinaciones para formar funciones de control dependientes del contexto de audio.

Además, los contextos de VoIP y no VoIP según se diferencian y utilizan en la fórmula (17) y (18), otros contextos de audio, tales como multimedia similar a cine, música a largo plazo y juego o audio de baja calidad y audio de alta calidad, pueden utilizarse de una forma similar.

4.3 Tipo de contextos

Los tipos de contextos pueden utilizarse directamente también para controlar el nivelador de volumen para que evitar sonidos desagradables, tales como ruido, se refuercen en una magnitud excesiva. A modo de ejemplo, el valor de confianza de VoIP puede utilizarse para controlar el nivelador de volumen, haciéndole menos sensible cuando su valor de confianza es alto.

Más concretamente, con el valor de confianza de VoIP $Conf_{VOIP}$ el nivel del nivelador de volumen puede establecerse para ser proporcional a (1 -$Conf_{VOIP}$). Es decir, el nivelador de volumen está casi desactivado en el contenido de VoIP (cuando el valor de confianza de VoIP es alto), lo que es coherente con el ajuste manual tradicional (preset) que desactiva el nivelador de volumen para el contexto de VoIP.

Como alternativa, se puede establecer diferentes gamas de ganancia dinámica para diferentes contextos de señales de audio. En general, una magnitud del VL (nivelador de volumen) ajusta, además, la magnitud de la ganancia aplicada en una señal de audio y puede considerarse como otra ponderación (no lineal) sobre la ganancia dinámica. En una forma de realización, un ajuste puede ser:

Tabla 1

	MULTIMEDIA SIMILAR A CINE	MÚSICA A LARGO PLAZO	VoIP	JUEGO
Magnitud de VL	Alta	Media	Off (o más baja)	Baja

Además, suponiendo que una magnitud de VL prevista sea predefinida para cada tipo de contexto. A modo de ejemplo, la magnitud de VL se establece como 1 para multimedia similar a cine, 0 para VoIP, 0.6 para música y 0.3 para juego, pero la presente idea inventiva no está limitada a este respecto. De conformidad con el ejemplo, si la gama de la ganancia dinámica de multimedia similar a cine es del 100 %, entonces, la gama de la ganancia dinámica de VoIP es del 60 % y así sucesivamente. Si la clasificación del clasificador de audio 200 está basada en una decisión difícil, en tal caso, la gama de la ganancia dinámica puede establecerse directamente como en el ejemplo anterior. Si la clasificación del clasificador de audio 200 está basada en una decisión programada, en tal caso, la gama puede ajustarse sobre la base del valor de confianza del tipo de contexto.

De modo similar, el clasificador de audio 200 puede identificar múltiples tipos de contextos desde la señal de audio, y la unidad de ajuste 300C puede configurarse para ajustar la gama de la ganancia dinámica mediante la ponderación de los valores de confianza de los múltiples tipos de contenidos sobre la base de la importancia de los múltiples tipos de contenidos.

En general, para el tipo de contexto, las funciones similares a (10)-(16) pueden utilizarse también, en este caso, para establecer la magnitud de VL adecuada de forma adaptativa, con los tipos de contenidos allí sustituidos con tipos de contextos, y realmente, la tabla 1 refleja la importancia de un tipo de contexto diferente.

Desde otro punto de vista, el valor de confianza puede utilizarse para derivar una ponderación normalizada, según se describe en la sección 1.4. En el supuesto de que una magnitud específica se predefine para cada tipo de contexto en la tabla 1, entonces, se puede aplicar también una fórmula similar a la fórmula (9). De forma imprevista, se pueden aplicar también soluciones similares a múltiples tipos de contenidos y cualesquiera otros tipos de audio.

4.4 Combinación de formas de realización y escenarios de aplicación

De modo similar a la Parte 1, la totalidad de las formas de realización y variantes, según fueron anteriormente descritas, pueden ponerse en práctica en cualquiera de sus combinaciones, y cualesquiera componentes mencionados en diferentes partes/formas de realización, pero teniendo las mismas o similares funciones que pueden ponerse en práctica como los mismos o componentes separados. A modo de ejemplo, cualesquiera dos o más soluciones descritas en las secciones 4.1 a 4.3 pueden combinarse entre sí. Y cualquiera de las combinaciones puede combinarse, además, con cualquier forma de realización descrita o implícita en las Partes 1-3 y las otras partes que se describirán a continuación.

La Figura 21 ilustra el efecto del controlador de nivelador de volumen dado a conocer en la presente idea inventiva en comparación con un segmento a largo plazo original (Figura 21(A)), el segmento a corto plazo procesado por un nivelador de volumen convencional sin modificación de parámetros (Figura 21(B)), y el segmento a corto plazo procesado por un nivelador de volumen según se presenta en esta idea inventiva (Figura 21(C)). Como puede observarse, en el nivelador de volumen convencional según se ilustra en la Figura 21(B), el volumen del ruido (la segunda mitad de la señal de audio) se refuerza también y resulta enojoso. Por el contrario, en el nuevo nivelador de volumen según se ilustra en la Figura 21(C), el volumen de la parte efectiva de la señal de audio es reforzado sin reforzar evidentemente el volumen del ruido, lo que proporciona una buena experiencia para la audiencia.

4.5 Método de control del nivelador de volumen

De modo similar a la Parte 1, en el proceso de descripción del controlador del nivelador de volumen en las formas de realización anteriormente descritas, se dan a conocer, evidentemente, también algunos procesos o métodos. A continuación, se proporciona un resumen de estos métodos sin repetir algunos de los detalles ya descritos con anterioridad.

En primer lugar, las formas de realización del método de procesamiento de audio según se describe en la Parte 1 pueden utilizarse para un nivelador de volumen, los parámetros es uno de los objetivos a ajustarse por el método de procesamiento de audio. Desde este punto de vista, el método de procesamiento de audio es también un método de control del nivelador de volumen A.

En esta sección, solamente se describirán los aspectos específicos para el control del nivelador de volumen. Para los aspectos generales del método de control puede hacerse referencia a la Parte 1.

En conformidad con la presente idea inventiva, se da a conocer un método de control del nivelador de volumen A, incluyendo la identificación del tipo de contenido de una señal en tiempo real, y ajustando un nivelador de volumen en una manera continua sobre la base del tipo de contenido que se identifica, estableciendo una correlación positiva de la

ganancia dinámica del nivelador de volumen con los tipos de contenidos informativos de la señal de audio y una correlación negativa de la ganancia dinámica del nivelador de volumen con los tipos de contenidos interferentes de la señal de audio.

El tipo de contenido puede comprender voz, música a corto plazo, ruido y sonido de fondo. En términos generales, el ruido se considera como un tipo de contenido interferente.

Cuando se ajusta la ganancia dinámica del nivelador de volumen, puede ajustarse directamente sobre la base del valor de confianza del tipo de contenido o puede ajustarse por intermedio de una función de transferencia del valor de confianza del tipo de contenido.

Como fue ya descrito, la señal de audio puede clasificarse en múltiples tipos de audio al mismo tiempo. Cuando se implican múltiples tipos de contenidos, la operación de ajuste 1104 puede configurarse para considerar al menos algunos de los múltiples tipos de contenidos de audio mediante la ponderación de los valores de confianza de los múltiples tipos de contenidos sobre la base de la importancia de los múltiples tipos de contenidos, o mediante la ponderación de los efectos de los múltiples tipos de contenidos sobre la base de los valores de confianza. Más concretamente, y la operación de ajuste 1104 puede configurarse para considerar al menos un tipo de contenido dominante sobre la base de los valores de confianza. Cuando la señal de audio contiene, a la vez, tipos de contenidos interferentes y tipos de contenidos informativos, la operación de ajuste puede configurarse para considerar al menos un tipo de contenido interferente dominante sobre la base de los valores de confianza y/o considerar al menos un tipo de contenido informativo dominante sobre la base de los valores de confianza.

Diferentes tipos de audio pueden tener una influencia mutua. Por lo tanto, la operación de ajuste 1104 puede considerarse para modificar la ponderación un tipo de contenido con el valor de confianza de al menos otro tipo de contenido.

Según se describió en la Parte 1, el valor de confianza del tipo de audio de la señal de audio puede ser objeto de alisado. Para conocer los detalles de la operación de alisado, puede hacerse referencia a la Parte 1.

El método puede incluir, además, la identificación del tipo de contexto de la señal de audio, en donde la operación de ajuste 1104 puede configurarse para ajustar la gama de la ganancia dinámica sobre la base del valor de confianza del tipo de contexto.

La función de un tipo de contenido está limitada por el tipo de contexto en donde está situado. Por lo tanto, cuando se obtienen, a la vez, información del tipo de contenido e información del tipo de contexto para una señal de audio al mismo tiempo (es decir, para el mismo segmento de audio), el tipo de contenido de la señal de audio puede determinarse como informativo o interferente sobre la base del tipo de contexto de la señal de audio. Además, al tipo de contenido en una señal de audio de un tipo de contexto diferente puede asignarse una ponderación diferente dependiendo del tipo de contexto de la señal de audio. Desde otro punto de vista, se puede utilizar una ponderación diferente (mayor o menor, plusvalía o minusvalía) para reflejar la naturaleza informativa o la naturaleza interferente de un tipo de contenido.

El tipo de contexto de la señal de audio puede comprender VoIP, multimedia similar a cine, música a largo plazo y juego. Y en la señal de audio del tipo de contexto VoIP, el sonido de fondo se considera como un tipo de contenido interferente; mientras en la señal de audio del tipo de contexto no VoIP, el sonido de fondo y/o voz y/o música se consideran como un tipo de contenido informativo. Otros tipos de contextos pueden incluir audio de alta calidad o audio de baja calidad.

De modo similar a los múltiples tipos de contenidos, cuando la señal de audio se clasifica en múltiples tipos de contextos con valores de confianza correspondientes al mismo tiempo (con respecto a un mismo segmento de audio), la operación de ajuste 1104 puede configurarse para considerar al menos algunos de los múltiples tipos de contexto mediante ponderación de los valores de confianza de los múltiples tipos de contextos sobre la base de la importancia de los múltiples tipos de contextos o mediante la ponderación de los efectos de los múltiples tipos de contextos sobre la base de los valores de confianza. Más concretamente, la operación de ajuste puede configurarse para considerarse al menos un tipo de contexto dominante sobre la base de los valores de confianza.

Por último, las formas de realización del método que se describen en esta sección pueden utilizar el método de clasificación de audio según se describirá en las Partes 6 y 7 y la descripción detallada se omite aquí.

De modo similar a las formas de realización del aparato de procesamiento de audio, cualquier combinación de las formas de realización del método de procesamiento de audio y sus variantes son de uso práctico por un lado; y por el otro lado, cada aspecto de las formas de realización del método de procesamiento de audio y sus variantes pueden ser soluciones separadas. Además, cualesquiera dos o más soluciones descritas en esta sección pueden combinarse entre sí y estas combinaciones pueden combinarse, además, con cualquier forma de realización descrita o implícita en las otras partes de la presente idea inventiva.

Parte 5: Controlador de ecualizador y método de control

La ecualización se suele aplicar sobre una señal de música para ajustar o modificar su equilibrio espectral, también conocido como "tono" o "timbre". Un ecualizador tradicional permite a los usuarios configurar el perfil global (curva o forma) de la respuesta de frecuencia (ganancia) en cada banda de frecuencia individual, con el fin de resaltar algunos instrumentos o eliminar sonidos indeseados. Los reproductores de música populares, tales como un reproductor de multimedia Windows, suelen proporcionar un ecualizador de gráficos para ajustar la ganancia en cada banda de frecuencias, y proporcionar también un conjunto de preajustes del ecualizador para diferentes géneros de música, tales como Rock, Rap, Jazz y Folk, para obtener la mejor experiencia en la escucha para diferentes géneros de música. Una vez que se seleccione un preajuste, o se establezca un perfil, las mismas ganancias de ecualización se aplicarán sobre la señal, hasta que el perfil se modifique manualmente.

Por el contrario, un ecualizador dinámico proporciona una forma de ajustar automáticamente las ganancias de ecualización en cada banda de frecuencias con el fin de mantener una coherencia global del equilibrio espectral con respecto a un timbre o tono deseado. Esta coherencia se consigue controlando continuamente el equilibrio espectral de la señal de audio, comparándola con un equilibrio espectral prestablecido deseado y ajustando dinámicamente las ganancias de ecualización aplicadas para transformar el equilibrio espectral original del audio en el equilibrio espectral deseado. El equilibrio espectral deseado se selecciona manualmente o se establece previamente antes del procesamiento.

Ambas clases de los ecualizadores comparten el inconveniente siguiente: el mejor perfil de ecualización, el equilibrio espectral deseado o los parámetros relacionados han de seleccionarse manualmente, y no se pueden modificar automáticamente sobre la base del contenido de audio en la reproducción. La discriminación de los tipos de contenidos de audio será muy importante para proporcionar una buena calidad global para diferentes clases de señales de audio. A modo de ejemplo, diferentes piezas musicales necesitan diferentes perfiles de ecualización, tal como los correspondientes a diferentes géneros.

En un sistema de ecualizador en el que son de entrada posible cualquier clase de señales de audio (no solamente de música), los parámetros del ecualizador necesitan ajustarse sobre la base de los tipos de contenidos. A modo de ejemplo, el ecualizador se suele activar con señales de música, pero se desactiva con señales de la voz, puesto que puede cambiar el timbre de la voz demasiado y en correspondencia, obtener un sonido de la señal no natural.

Con el fin de resolver este problema al menos en parte, la presente idea inventiva propone el control del ecualizador sobre la base de las formas de realización descritas en la Parte 1.

De modo similar a las Partes 2 -4, como un ejemplo específico del aparato de procesamiento de audio y del método descrito en la Parte 1, el ecualizador 408 puede hacer uso de todas las formas de realización descritas en la Parte 1 y cualesquiera combinaciones de dichas formas de realización que se dieron a conocer. Más concretamente, en el caso de control del ecualizador 408, el clasificador de audio 200 y la unidad de ajuste 300 en el aparato de procesamiento de audio 100 según se ilustra en las Figuras 1 a 10, pueden constituir un controlador 2200 del ecualizador 408 según se ilustra en la Figura 22. En esta forma de realización, puesto que la unidad de ajuste es específica para el ecualizador 408 puede referirse como 300D.

Es decir, sobre la base de la idea inventiva descrita en la Parte 1, un controlador de ecualizador 2200 puede comprender un clasificador de audio 200 para identificar continuamente el tipo de audio de una señal de audio; y una unidad de ajuste 300D para ajustar un ecualizador en manera continua sobre la base del valor de confianza del tipo de audio que se identifica. De modo similar, el clasificador de audio 200 puede comprender al menos uno de entre el clasificador de contenido de audio 202 y el clasificador de contexto de audio 204, y el controlador de ecualizador de volumen 2200 puede comprender, además, al menos uno de entre la unidad de alisado de tipo 712, la unidad de alisado de parámetros 814 y el temporizador 916.

Por lo tanto, en esta parte, no se repetirán los contenidos ya descritos en la Parte 1, y solamente se proporcionarán algunos de sus ejemplos específicos.

5.1 Control basado en el tipo de contenido

En términos generales, para los tipos de contenidos de audio generales, tales como la música, la voz, sonido de fondo y ruido, el ecualizador debe establecerse diferentemente en distintos tipos de contenidos. De modo similar a la configuración tradicional, el ecualizador puede activarse automáticamente sobre las señales de música, pero desactivarse en la presencia de voz; o en una manera más continua, establecer un nivel alto de ecualización sobre señales musicales y un bajo nivel de ecualización en las señales de voz. De este modo, el nivel de ecualización de un ecualizador puede establecerse automáticamente para diferentes contenidos de audio.

Más concretamente, para la música, se observa que el ecualizador no funciona tan bien en una pieza musical que tenga una fuente dominante, puesto que el timbre de la fuente dominante puede cambiar notablemente y tener un sonido no natural si se aplica una ecualización inadecuada. Considerando esta circunstancia, sería conveniente establecer un nivel de ecualización bajo sobre las piezas musicales con fuentes dominantes, mientras que el nivel de ecualización puede mantenerse alto en relación con piezas musicales sin fuentes dominantes. Con esta información, el ecualizador puede

establecer automáticamente el nivel de ecualización para diferentes contenidos musicales.

La música puede agruparse también sobre la base de diferentes propiedades, tales como género, instrumentos y características músicas generales incluyendo ritmo, *tempo*, y timbre. Del mismo modo que diferentes preajustes del ecualizador se utilizan para diferentes géneros musicales, estos grupos/agrupaciones musicales pueden tener también sus propios perfiles de ecualización óptima o curvas del ecualizador (en el ecualizador tradicional) o un equilibrio espectral deseado óptimo (en el ecualizador dinámico).

Según se mencionó con anterioridad, el ecualizador se suele activar sobre el contenido musical pero desactivarse sobre la voz, puesto que el ecualizador puede hacer que un diálogo no suene adecuadamente debido al cambio del timbre. Un método para conseguirlo automáticamente es relacionar el nivel de ecualización con el contenido, en particular, el valor de confianza de la música y/o el valor de confianza de la voz que se obtienen a partir del módulo de clasificación de contenido de audio. En este caso, el nivel de ecualización puede explicarse como la ponderación de las ganancias del ecualizador aplicadas. Cuando más alto sea el nivel, tanto más fuerte será la ecualización aplicada. A modo de ejemplo, si el nivel de ecualización es 1, se obtiene la aplicación de un perfil de ecualización total; si el nivel de ecualización es cero, todas las ganancias son, en correspondencia, 0 dB y de este modo, se aplica la no ecualización. El nivel de ecualización puede representarse por diferentes parámetros en diferentes puestas en práctica de los algoritmos del ecualizador. Una forma de realización a modo de ejemplo de estos parámetros es la ponderación del ecualizador según se pone en práctica en el documento de A. Seefeldt et.al. *"Cálculo y ajuste de la intensidad recibida y/o el equilibrio espectral percibido de una señal de audio"*, publicado como US 2009/0097676 A1:, que se incorpora aquí en su integridad por referencia.
.

Varios sistemas de control pueden designarse para ajustar el nivel de ecualización. A modo de ejemplo, con la información del tipo de contenido de audio, el valor de confianza de la voz o el valor de confianza de la música pueden utilizarse para establecer el nivelación de ecualización, como

$$L_{eq} \propto Conf_{music} \qquad (20)$$

O

$$L_{eq} \propto 1 - Conf_{speech} \qquad (21)$$

en donde L_{eq} es el nivel de ecualización y $Conf_{music}$ y $Conf_{speech}$ corresponden al valor de confianza de la música y de la voz respectivamente.

Es decir, la unidad de ajuste 300D puede configurarse para establecer una correlación positiva en el nivel de ecualización con un valor de confianza de la música a corto plazo o una correlación negativa del nivel de ecualización con un valor de confianza de la voz.

El valor de confianza de la voz y el valor de confianza de la música pueden utilizarse, además, conjuntamente para establecer el nivel de ecualización. La idea general es que el nivel de ecualización debe ser alto solamente cuando el valor de confianza de la música es alto y el valor de confianza de la voz es bajo, y de no ser así, el nivel de ecualización es bajo. A modo de ejemplo,

$$L_{eq} = Conf_{music}(1 - Conf_{speech}^{\alpha}) \qquad (22)$$

en donde el valor de confianza de la voz se establece para el valor de α con el fin de negociar con la confianza de la voz no cero en las señales musicales, lo que puede suceder con frecuencia. Con la fórmula anterior, la ecualización será completamente aplicada (con el nivel igual a 1) sobre las señales musicales puras sin ningún componente de la voz. Según se establece en la Parte 1, el valor de α puede considerarse como un coeficiente de ponderación basado en la importancia del tipo de contenido y puede ser normalmente establecido de 1 a 2.

Si se asigna una mayor ponderación sobre el valor de confianza de la voz, la unidad de ajuste 300D puede configurarse para desactivar el ecualizador 408 cuando el valor de confianza para el tipo de contenido de la voz es mayor que un valor umbral.

En la descripción anterior, los tipos de contenidos de música y voz se toman a modo de ejemplo. Como alternativa o de forma adicional, los valores de confianza del sonido de fondo y/o ruido pueden considerarse también a este respecto. Más concretamente, la unidad de ajuste 300D puede configurarse para establecer una correlación positiva de un nivel de ecualización con un valor de confianza del sonido de fondo y/o una correlación negativa del nivel de ecualización con un valor de confianza del ruido.

A modo de otra forma de realización, el valor de confianza puede utilizarse para derivar una ponderación normalizada según se describe en la sección 1.4. Suponiendo que un nivel de ecualización previsto se predefine para cada tipo de contenido (p.ej., 1 para música, 0 para la voz, 0.5 para ruido y sonido de fondo), se puede aplicar exactamente una fórmula similar a la fórmula (8).

El nivel de ecualización puede alisarse, además, para evitar que un cambio rápido pueda introducir artefactos audibles en los puntos de transición. Lo que antecede puede realizarse con la unidad de alisado de parámetros 814 según se describe en la sección 1.5.

5.2 Probabilidad de fuentes dominantes en la música

Con el fin de evitar que a la música con fuentes dominantes se le aplique un nivel de ecualización alto, el nivel de ecualización puede, además, ponerse en correlación con el valor de confianza $Conf_{dom}$ lo que indica si una pieza musical contiene una fuente dominante, a modo de ejemplo

$$L_{eq} = 1 - Conf_{dom} \qquad (23)$$

De este modo, el nivel de ecualización será bajo en las piezas musicales con fuentes dominantes y alto en las piezas musicales sin fuentes dominantes.

En este caso, aunque el valor de confianza de la música con una fuente dominante se describe, se puede utilizar también el valor de confianza de la música sin una fuente dominante. Es decir, la unidad de ajuste 300D puede configurarse para establecer una correlación positiva de un nivel de ecualización con un valor de confianza de música a corto plazo sin fuente dominante y/o una correlación negativa del nivel de ecualización con un valor de confianza de música a corto plazo con fuentes dominantes.

Según se establece en la sección 1.1, aunque la música y la voz por un lado, y la música con o sin fuentes dominantes por otro lado, son tipos de contenidos en diferentes niveles jerárquicos, pueden considerarse en paralelo. Considerando conjuntamente el valor de confianza de fuentes dominantes y los valores de confianza de la voz y de la música, según se describió con anterioridad, el nivel de ecualización puede establecerse combinando al menos una de entre las fórmulas (20)-(21) con (23). Un ejemplo consiste en combinar la totalidad de las tres fórmulas:

$$L_{eq} = Conf_{music}(1 - Conf_{speech})(1 - Conf_{dom}) \qquad (24)$$

Diferentes ponderaciones basadas en la importancia del tipo de contenido pueden aplicarse, además, a diferentes valores de confianza para su generalidad, tal como en la manera de la fórmula (22).

A modo de otro ejemplo, suponiendo que $Conf_{dom}$ se calcula solamente cuando la señal de audio es música, una función escalonada puede designarse como

$$L_{eq} = \begin{cases} (1 - Conf_{dom}) & Conf_{music} > threshold \\ Conf_{music}(1 - conf_{speech}^{\alpha}) & de\ lo\ contrario \end{cases} \qquad (25)$$

Esta función establece el nivel de ecualización basado en el valor de confianza de puntuaciones dominantes si el sistema de clasificación determina fielmente que el audio es música (el valor de confianza de la música es mayor que un valor umbral); de no ser así, se establece sobre la base de los valores de confianza de la música y de la voz. Es decir, la unidad de ajuste 300D puede configurarse para considerar la música a corto plazo sin/con fuentes dominantes cuando el valor de confianza para la música a corto plazo es mayor que un valor umbral. Por supuesto, la primera o la segunda mitad en la fórmula (25) puede modificarse en la manera de la fórmula (20) o (24).

El mismo sistema de alisado según se describe en la sección 1.5 puede aplicarse también a este respecto y la constante de tiempo α puede basarse además, sobre la base del tipo de transición, de modo que la transición desde música con fuente dominante a música sin fuente dominante, o la transición desde música sin fuentes dominantes a música con fuentes dominantes. Para esta finalidad, se puede aplicar también una fórmula similar a la fórmula (4').

5.3 Preajustes del ecualizador

Además del ajuste adaptativo del nivel de ecualización sobre la base de los valores de confianza de tipos de contenidos de audio, pueden elegirse automáticamente preajustes del equilibrio espectral deseado o perfiles de ecualización adecuados para diferentes contenidos de audio, dependiendo de su género, instrumento u otras características. La música con el mismo género, que contienen el mismo instrumento, o que tienen las mismas características musicales,

pueden compartir los mismos perfiles de ecualización o preajustes de equilibrio espectral deseados.

Para su generalidad, utilizamos el término "agrupaciones musicales" para representar a los grupos musicales con el mismo género, el mismo instrumento o atributos musicales similares y puede considerarse como otro nivel jerárquico de tipos de contenidos de audio según se establece en la sección 1.1. El perfil de ecualización adecuado, el nivel de ecualización y/o preajuste de equilibrio espectral deseado, pueden asociarse a cada agrupación musical. El perfil de ecualización es la curva de ganancia aplicada sobre la señal musical y puede ser cualquiera de los preajustes del ecualizador que se utilizan para diferentes géneros musicales (tales como Clásica, Rock, Jazz, y Folk), y el preajuste del equilibrio espectral deseado representa el timbre deseado para cada agrupación musical. La Figura 23 ilustra varios ejemplos de preajustes del equilibrio espectral deseado según se implantan en las tecnologías de Dolby Home Theater. Cada uno describe la forma espectral deseada a través de la gama de frecuencias audibles. Esta forma se compara continuamente con la forma espectral del audio entrante, y las ganancias de ecualización se calculan a partir de esta comparación para transformar la forma espectral del audio entrante en la que se ha prestablecido.

Para una nueva pieza musical, la agrupación más próxima puede determinarse (decisión difícil) o el valor de confianza con respecto a cada agrupación musical puede calcularse (decisión programada). Sobre la base de esta información, un perfil de ecualización adecuado, o un preajuste de equilibrio espectral deseado, pueden determinarse para la pieza musical dada. La manera más simple es asignar el perfil correspondiente de la mejor agrupación adaptada, como

$$P_{eq} = P_{c^*} \tag{26}$$

en donde P_{eq} es el perfil de ecualización estimado o el preajuste del equilibrio espectral deseado y c* es el índice de la mejor agrupación musical adaptada (el tipo de audio dominante), que pueden obtenerse captando la agrupación con el más alto valor de confianza.

Además, puede existir más de una agrupación musical que tenga un valor de confianza que sea mayor que cero, lo que significa que la pieza musical tiene atributos más o menos similares a los que tienen las agrupaciones. A modo de ejemplo, una pieza musical puede tener múltiples instrumentos, o puede tener atributos de múltiples géneros. Ello inspira otra forma de estimar el perfil de ecualización adecuado considerando todas las agrupaciones, en lugar de utilizar solamente la agrupación más próxima. A modo de ejemplo, una suma ponderada puede utilizarse:

$$P_{eq} = \sum_{c=1}^{N} w_c P_c \tag{27}$$

en donde N es el número de agrupaciones predefinidas y w_c es la ponderación del perfil designado P_c con respecto a cada agrupación musical predefinida (con el índice c), que debe normalizarse a 1 sobre la base de sus valores de confianza correspondientes. De este modo, el perfil estimado sería una mezcla de los perfiles de las agrupaciones musicales. A modo de ejemplo, para una pieza musical que tenga ambos atributos de Jazz y Rock, el perfil estimado sería algo comprendido entre ambos.

En algunas aplicaciones, puede no ser deseable implicar todas las agrupaciones según se ilustra en la fórmula (27). Solamente un subconjunto de las agrupaciones – las agrupaciones más relaciones con la pieza musical actual – necesitan considerarse, la fórmula (27) puede revisarse ligeramente a:

$$P_{eq} = \sum_{c'=1}^{N'} w_{c'} P_{c'} \tag{28}$$

en donde N' es el número de agrupaciones a considerarse y c' es el índice de la agrupación después de clasificar, en forma decreciente, las agrupaciones basadas en sus valores de confianza. Utilizando un subconjunto, se puede concentrar más en las agrupaciones más relacionadas y excluir las menos pertinentes. Dicho de otro modo, la unidad de ajuste 300D puede configurarse para considerar al menos algún tipo de audio dominante sobre la base de los valores de confianza.

En la descripción anterior, las agrupaciones musicales se toman a modo de ejemplo. De hecho, las soluciones son aplicables a tipos de audio a cualquier nivel jerárquico según se describió en la sección 1.1. De este modo, en general, la unidad de ajuste 300D puede configurarse para asignarse un nivel de ecualización y/o un perfil de ecualización y/o preajuste de equilibrio espectral a cada tipo de audio.

5.4 Control basado en el Tipo de contexto

En las secciones anteriores, la descripción se concentró en varios tipos de contenidos. En más formas de realización a describirse en esta misma sección, el tipo de contexto puede considerarse como alternativa o de forma adicional.

En general, el ecualizador se activa para la música pero se desactiva para el contenido de multimedia similar a cine, puesto que el ecualizador puede hacer que los diálogos en multimedia similar a cine no suenen tan bien debido a un cambio de timbre obvio. Ello indica que el nivel de ecualización puede relacionarse con el valor de confianza de la música a largo plazo y/o el valor de confianza de multimedia similar a cine:

$$L_{eq} \propto Conf_{MUSIC} \tag{29}$$

O

$$L_{eq} \propto 1 - Conf_{MOVIE} \tag{30}$$

en donde L_{eq} es el nivel de ecualización, $Conf_{MUSIC}$ y $Conf_{MOVIE}$ se refieren al valor de confianza de la música a largo plazo y multimedia similar a cine.

Es decir, la unidad de ajuste 300D puede configurarse para establecer una correlación positiva en un nivel de ecualización con un valor de confianza de la música a largo plazo, o una correlación negativa del nivel de ecualización con un valor de confianza de multimedia similar a cine.

Es decir, para una señal de multimedia similar a cine, el valor de confianza de multimedia similar a cine es alto (o la confianza de música es baja) y de este modo, el nivel de ecualización es bajo; por otro lado, para una señal de música, el valor de confianza de multimedia similar a cine será bajo (o la confianza de la música es alta) y en consecuencia, el nivel de ecualización es alto.

Las soluciones ilustradas en las fórmulas (29) y (30) pueden modificarse del mismo modo que las fórmulas (22) a (25), y/o pueden combinarse con cualquiera de las soluciones ilustradas en las fórmulas (22) a (25).

Además o de forma alternativa, la unidad de ajuste 300D puede configurarse para establecer una correlación negativa en el nivel de ecualización con un valor de confianza del juego.

Como otra forma de realización, el valor de confianza puede utilizarse para derivar una ponderación normalizada según se describe en la sección 1.4. Suponiendo que se predefine un nivel/perfil de ecualización previsto para cada tipo de contexto (los perfiles de ecualización se muestran en la tabla 2 siguiente), puede aplicarse también una fórmula similar a la fórmula (9).

Tabla 2:

	MULTIMEDIA SIMILAR A CINE	MÚSICA A LARGO PLAZO	VoIP	JUEGO
Perfil de ecualización	Perfil 1	Perfil 2	Perfil 3	Perfil 4

En este caso, en algunos perfiles, todas las ganancias pueden establecerse a cero, como una forma para desactivar el ecualizador para ese cierto tipo de contexto, tal como multimedia similar a cine y juego.

5.5 Combinación de formas de realización y escenarios de aplicación

De modo similar a la Parte 1, todas las formas de realización y variantes anteriormente descritas pueden ponerse en práctica en cualquiera de sus combinaciones y cualesquiera componentes mencionados en diferentes partes/formas de realización, pero teniendo la misma o funciones similares, pueden ponerse en práctica como el mismo o componentes separados.

A modo de ejemplo, cualesquiera dos o más de las soluciones descritas en las secciones 5.1 a 5.4 pueden combinarse entre sí. Y cualquiera de las combinaciones puede combinarse, además, con cualquier forma de realización descrita o implícita en las partes 1-4 y las demás partes que se describirán más adelante.

5.6 Método de control del ecualizador

De modo similar a la Parte 1, en el proceso de describir el controlador del ecualizador en las formas de realización anteriormente descritas, se dan a conocer, evidentemente, también algunos procesos o métodos. A continuación se proporciona un resumen de estos métodos sin repetir algunos de los detalles ya descritos con anterioridad.

En primer lugar, las formas de realización del método de procesamiento de audio según se describe en la Parte 1 puede utilizarse para un ecualizador, los parámetros es uno de los objetivos a ajustarse por el método de procesamiento de audio. Desde este punto de vista, el método de procesamiento de audio es también un método de control del ecualizador.

En esta sección, solamente se describirán los aspectos específicos para el control del ecualizador. Para los aspectos generales del método de control, puede hacerse referencia a la Parte 1.

De conformidad con las formas de realización, un método de control del ecualizador puede incluir la identificación del tipo de audio de una señal de audio en tiempo real, y ajustar un ecualizador en una manera continua sobre la base del valor de confianza del tipo de audio identificado.

De modo similar a las otras partes de la presente idea inventiva, cuando se implica a múltiples tipos de audio con valores de confianza correspondientes, la operación de ajuste 1104 puede configurarse para considerar al menos algunos de los múltiples tipos de audio mediante la ponderación de los valores de confianza de los múltiples tipos de audio sobre la base de la importancia de los múltiples tipos de audio, o mediante una ponderación de los efectos de los múltiples tipos de audio sobre la base de los valores de confianza. Más concretamente, la operación de ajuste 1104 puede configurarse para considerar al menos un tipo de audio dominante sobre la base de los valores de confianza.

Según se describió en la Parte 1, el valor del parámetro ajustado puede ser alisado. Puede hacerse referencia a la sección 1.5 y la sección 1.8 y se omite aquí una descripción detallada.

El tipo de audio puede ser del tipo de contenido o del tipo de contexto o ambos a la vez. Cuando se implica el tipo de contenido, la operación de ajuste 1104 puede configurarse para establecer una correlación positiva de un nivel de ecualización con un valor de confianza de música a corto plazo y/o una correlación negativa del nivel de ecualización con un valor de confianza de la voz. De forma adicional o como alternativa, la operación de ajuste puede configurarse para establecer una correlación positiva de un nivel de ecualización con un valor de confianza del fondo y/o una correlación negativa del nivel de ecualización con un valor de confianza del ruido.

Cuando se implica el tipo de contexto, la operación de ajuste 1104 puede configurarse para establecer una correlación positiva en un nivel de ecualización con un valor de confianza de la música a largo plazo y/o una correlación negativa del nivel de ecualización con un valor de confianza de multimedia similar a cine y/o juego.

Para el tipo de contenido de música a corto plazo, la operación de ajuste 1104 puede configurarse para establecer una correlación positiva en un nivel de ecualización con un valor de confianza de la música a corto plazo sin fuentes dominantes y/o una correlación negativa del nivel de ecualización con un valor de confianza de la música a corto plazo con fuentes dominantes. Lo que antecede puede realizarse solamente cuando el valor de confianza para la música a corto plazo es mayor que un valor umbral.

Además de ajustar el nivel de ecualización, otros aspectos de un ecualizador pueden ajustarse sobre la base de los valores de confianza de los tipos de audio de una señal de audio. A modo de ejemplo, la operación de ajuste 1104 puede configurarse para asignar un nivel de ecualización y/o un perfil de ecualización y/o un preajuste de equilibrio espectral para cada tipo de audio.

Con respecto a las instancias específicas de los tipos de audio, puede hacerse referencia a la Parte 1.

De modo similar a las formas de realización del aparato de procesamiento de audio, cualquier combinación de las formas de realización del método de procesamiento de audio y sus variantes son prácticas por un lado; y por el otro lado, cada aspecto de las formas de realización del método de procesamiento de audio y sus variantes pueden ser soluciones separadas. Además, cualesquiera dos o más soluciones descritas en esta sección pueden combinarse entre sí y estas combinaciones pueden combinarse, además, con cualquier forma de realización descrita o implícita en las otras partes de esta idea inventiva.

Parte 6: Clasificadores de audio y métodos de clasificación

Según se indicó en las secciones 1.1 y 1.2, los tipos de audio descritos en la presente idea inventiva, incluyendo varios niveles jerárquicos de tipos de contenidos y tipos de contextos, pueden clasificarse o identificarse con cualquier sistema de clasificación existente, incluyendo métodos basados en el aprendizaje de la máquina. En esta parte y en la parte siguiente, la presente idea inventiva propone algunos nuevos aspectos de clasificadores y métodos para clasificar tipos de contextos según se menciona en las partes anteriores.

6.1 Clasificador de contexto sobre la base de la clasificación de tipos de contenidos

Según se establece en las partes anteriores, el clasificador de audio 200 se utiliza para identificar el tipo de contenido de una señal de audio y/o para identificar el tipo de contexto de la señal de audio. Por lo tanto, el clasificador de audio 200

puede comprender un clasificador de contenido de audio 202 y/o un clasificador de contexto de audio 204. Cuando se adoptan las técnicas existentes para poner en práctica el clasificador de contenido de audio 202 y el clasificador de contexto de audio 204, los dos clasificadores pueden ser independientes entre sí, aunque pueden compartir algunas características y de este modo, pueden compartir algunos sistemas para extraer las características.

En esta parte y en la Parte 7 siguiente, en conformidad con el nuevo aspecto propuesto en la presente idea inventiva, el clasificador de contexto de audio 204 puede hacer uso de los resultados del clasificador de contenido de audio 202, es decir, el clasificador de audio 200 puede comprender: un clasificador de contenidos de audio 202 para identificar el tipo de contenido de una señal de audio; y un clasificador de contexto de audio 204 para identificar el tipo de contexto de la señal de audio sobre la base de los resultados del clasificador de contenido de audio 202. De este modo, los resultados de la clasificación del clasificador de contenido de audio 202 pueden utilizarse por el clasificador de contexto de audio 204 y la unidad de ajuste 300 (las unidades de ajuste 300A a 300D) según se describió en las partes anteriores. Sin embargo, aunque no se ilustra en los dibujos, el clasificador de audio 200 puede contener también dos clasificadores de contenido de audio 202 a utilizarse, respectivamente, por la unidad de ajuste 300 y el clasificador de contexto de audio 204.

Además, según se describe en la sección 1.2, en particular, cuando se clasifican múltiples tipos de audio, el clasificador de contenido de audio 202 o el clasificador de contexto de audio 204 pueden comprender un grupo de clasificadores que cooperan entre sí, aunque también es posible ponerse en práctica un clasificador único.

Según se describe en la sección 1.1, el tipo de contenido es una clase de tipo de audio con respecto a segmentos de audio a corto plazo que suelen tener una longitud en el orden de varias decenas de tramas (tal como 1 s), y el tipo de contexto es una clase de tipo de audio con respecto a segmentos de audio a largo plazo que suelen tener una longitud en el orden de magnitud de varias decenas de segundos (tal como 10 s). Por lo tanto, en correspondencia con el "tipo de contenido " y "tipo de contexto", se utiliza los términos “a corto plazo” y “a largo plazo” respectivamente cuando sea necesario. Sin embargo, según se describirá en la Parte 7 siguiente, aunque el tipo de contexto es para indicar la propiedad de la señal de audio en una escala de tiempos relativamente larga, puede identificarse también sobre la base de las características extraídas de los segmentos de audio a corto plazo.

Ahora se retorna a las estructuras del clasificador de contenidos de audio 202 y el clasificador de contexto de audio 204 haciendo referencia a la Figura 24.

Según se ilustra en la Figura 24, el clasificador de contenido de audio 202 puede comprender un extractor de característica a corto plazo 2022 para extraer características a corto plazo a partir de segmentos de audio a corto plazo, comprendiendo cada uno una secuencia de tramas de audio; y un clasificador a corto plazo 2024 para clasificar una secuencia de segmentos a corto plazo en un segmento de audio a largo plazo en tipos de audio a corto plazo que utilizan características a corto plazo respectivas. El extractor de característica a corto plazo 2022 y el clasificador a corto plazo 2024 pueden ponerse con las técnicas existentes, pero también se proponen algunas modificaciones para extractor de características a corto plazo 2022 en la sección 6.3 siguiente.

El clasificador a corto plazo 2024 puede configurarse para clasificar cada una de entre las secuencias de segmentos a corto plazo en al menos uno de los tipos de audio a corto plazo siguientes (tipos de contenidos): voz, música a corto plazo, sonido de fondo y ruido, que han sido explicados en la sección 1.1. Cada uno del tipo de contenido puede clasificarse, además, en tipos de contenidos, con un nivel jerárquico más bajo, tal como se describe en la sección 1.1 pero sin limitación este respecto.

Según es conocido en esta técnica, los valores de confianza de los tipos de audio clasificados pueden obtenerse también por el clasificador a corto plazo 2024. En la presente idea inventiva, cuando se menciona la operación de cualquier clasificador, deberá entenderse que los valores de confianza se obtienen al mismo tiempo si fuera necesario, sin importar si se registran explícitamente o no se hace. Un ejemplo de clasificación de tipo de audio puede encontrarse en el documento de L. Lu, H.-J. Zhang, y S. Li, *"Clasificación de audio basado en contenidos y segmentación utilizando máquinas vectoriales de soporte"*, ACM Multimedia Systems Journal 8 (6), páginas 482-492, marzo 2003:, que se incorpora aquí en su integridad por referencia..

Por otro lado, según se ilustra en la Figura 24, el clasificador de contexto de audio 204 puede comprender un extractor de datos estadísticos 2042 para calcular la estadística de los resultados del clasificador a corto plazo con respecto a la secuencia de segmentos a corto plazo en el segmento de audio a largo plazo, como características a largo plazo; y un clasificador a largo plazo 2044 para, utilizando las características a largo plazo, clasificar el segmento de audio a largo plazo en tipos de audio a largo plazo. De modo similar, el extractor de datos estadísticos 2042 y el clasificador a largo plazo 2044 pueden ponerse con técnicas existentes, pero también se proponen algunas modificaciones para el extractor de datos estadísticos 2042 en la sección 6.2 siguiente.

El clasificador a largo plazo 2044 puede configurarse para clasificar el segmento de audio a largo plazo en al menos uno de los tipos de audio a largo plazo (tipos de contextos) siguientes: multimedia similar a cine, música a largo plazo, juego y VoIP, que han sido explicados en la sección 1.1. Como alternativa o de forma adicional, el clasificador a largo plazo 2044 puede configurarse para clasificar el segmento de audio a largo plazo en VoIP o no VoIP, lo que ha sido explicado en la

sección 1.1. De forma alternativa o adicional, el clasificador a largo plazo 2044 puede configurarse para clasificar el segmento de audio a largo plazo en audio de alta calidad o audio de baja calidad, lo que se explicó en la sección 1.1. En la práctica, varios tipos de audio objetivo pueden elegirse y formarse sobre la base de las necesidades de aplicación/sistema.

Con respecto al significado y selección del segmento a corto plazo y del segmento a largo plazo (así como la trama a describirse en la sección 6.3), puede hacerse referencia a la sección 1.1.

6.2 Extracción de características a largo plazo

Según se ilustra en la Figura 24, en una forma de realización, solamente el extractor de datos estadísticos 2042 se utiliza para extraer características a largo plazo a partir de los resultados del clasificador a corto plazo 2024. En cuanto a las características a largo plazo, al menos una de las siguientes puede calcularse por el extractor de datos estadísticos 2042: la media y la varianza de los valores de confianza de los tipos de audio a corto plazo de los segmentos a corto plazo en el segmento a largo plazo a clasificarse, la media y la varianza ponderadas por los grados de importancia de los segmentos a corto plazo, la frecuencia de ocurrencia de cada tipo de audio a corto plazo y la frecuencia de transiciones entre diferentes tipos de audio a corto plazo en el segmento a largo plazo han de clasificarse.

En la Figura 25 se ilustra la media de los valores de confianza de la voz y de la música a corto plazo en cada segmento a corto plazo (de una longitud de 1s). Para fines de comparación, los segmentos se extraen desde tres contextos de audio diferentes: multimedia similar a cine (Figura 25(A)), música a largo plazo (Figura 25(B)), y VoIP (Figura 25(C)). Puede observarse que para el contexto de multimedia similar a cine, se obtienen altos valores de confianza para el tipo de voz o para el tipo de música y alterna entre estos tipos de audio frecuentemente. Por el contrario, el segmento de música a largo plazo proporciona un valor estable y de alta confianza de la música a corto plazo y un bajo valor de confianza de la voz y relativamente estable. Mientras que el segmento de VoIP proporciona un valor de confianza de música a corto plazo bajo y establece, pero proporciona valores de confianza de la voz fluctuantes debido a las pausas durante la conversación de VoIP.

La varianza de los valores de confianza para cada tipo de audio es también una característica importante para clasificar diferentes contextos de audio. La Figura 26 proporciona histogramas de la varianza de los valores de confianza de la voz, música a corto plazo, música de fondo y ruido en contextos de multimedia similar a cine, música a corto plazo y de audio VoIP (en el eje de abscisas es la varianza de valores de confianza en un conjunto de datos y el eje de ordenada corresponde al número de ocurrencias de cada conjunto de valores de varianza s en el receptáculo de datos, lo que puede normalizarse para indicar la probabilidad de ocurrencia de cada receptáculo de valores de varianza). Para multimedia similar a cine, todas las varianzas de valor de confianza de la voz, música a corto plazo y sonido de fondo son relativamente altas y ampliamente distribuidas, lo que indica que los valores de confianza de esos tipos de audio están cambiando en gran medida; para la música a largo plazo, todas las varianzas del valor de confianza de la voz, música a corto plazo, sonido de fondo y ruido son relativamente bajas y estrechamente distribuidas, lo que indica que los valores de confianza de dichos tipos de audio se mantienen estables; el valor de confianza de la voz se mantiene constantemente bajo y el valor de confianza de la música se mantiene constantemente alto; para VoIP, las varianzas del valor de confianza de la música a corto plazo son bajas y estrechamente distribuidas, mientras que las que corresponden a la voz son relativamente distribuidas de forma amplia, lo que se debe a las frecuentes pausas durante las conversaciones de VoIP.

Con respecto a las ponderaciones utilizadas en el cálculo de la media ponderada y de la varianza, se determinan sobre la base del grado de importancia de cada segmento a corto plazo. El grado de importancia de un segmento a corto plazo puede medirse por su energía o intensidad, lo que puede estimarse con numerosas técnicas existentes.

La frecuencia de ocurrencia de cada tipo de audio a corto plazo en el segmento a largo plazo a clasificarse es el valor de conteo de cada tipo de audio para los que se han clasificado los segmentos a corto plazo en el segmento a largo plazo, normalizados con la longitud del segmento a largo plazo.

La frecuencia de las transiciones entre diferentes tipos de audio a corto plazo en el segmento a largo plazo a clasificarse es el conteo de cambios de tipos de audio entre segmentos a corto plazo adyacentes en el segmento a largo plazo a clasificarse, normalizados con la longitud del segmento a largo plazo.

Cuando se describe los valores de la media y de la varianza de los valores de confianza con referencia a la Figura 25, la frecuencia de ocurrencia de cada tipo de audio a corto plazo y la frecuencia de transición entre dichos diferentes tipos de audio a corto plazo son también tratados de hecho. Estas características son también de importancia para la clasificación de contextos de audio. A modo de ejemplo, la música a largo plazo contiene principalmente tipos de audio de música a corto plazo por lo que tiene una alta frecuencia de ocurrencia de música a corto plazo, mientras que el VoIP contiene principalmente la voz y pausas de modo que tiene una alta frecuencia de ocurrencia de la voz o del ruido. A modo de otro ejemplo, en multimedia similar a cine se transita entre diferentes tipos de audio a corto plazo con más frecuencia que la música a largo plazo o VoIP, por lo que suele tener una frecuencia de transición más alta entre la música a corto plazo, la voz y el sonido de fondo; VoIP suele transitar entre la voz y el ruido con más frecuencia que los demás, por lo que suele tener una frecuencia de transición más alta entre la voz y el ruido.

En términos generales, se supone que los segmentos a largo plazo son de la misma longitud en la misma aplicación/sistema. Si éste es el caso, entonces el conteo de ocurrencia de cada tipo de audio a corto plazo y el conteo de transición entre diferentes tipos de audio a corto plazo en el segmento a largo plazo pueden utilizarse directamente sin necesidad de normalización. Si la longitud del segmento a largo plazo es variable, entonces, la frecuencia de ocurrencia y la frecuencia de transiciones, según se mencionó con anterioridad, deberán utilizarse a este respecto. Y las reivindicaciones en la presente idea inventiva deberán interpretarse como cubriendo ambas situaciones.

De forma adicional o alternativa, el clasificador de audio 200 (o el clasificador de contexto de audio 204) pueden comprender, además, un extractor de característica a largo plazo 2046 (Figura 27) para extraer características a largo plazo adicionales a partir del segmento de audio a largo plazo sobre la base de las características a corto plazo de la secuencia de segmentos a corto plazo en el segmento de audio a largo plazo. Dicho de otro modo, el extractor de característica a largo plazo 2046 no utiliza los resultados de la clasificación del clasificador a corto plazo 2024, pero utiliza directamente las características a corto plazo extraídas por el extractor de características a corto plazo 2022 para derivar algunas características a largo plazo a utilizarse por el clasificador a largo plazo 2044. El extractor de características a largo plazo 2046 y el extractor de datos estadísticos 2042 pueden utilizarse de forma independiente o conjuntamente. Dicho de otro modo, el clasificador de audio 200 puede comprender el extractor de características a largo plazo 2046 o el extractor de datos estadísticos 2042 o ambos a la vez.

Cualesquiera características pueden extraerse por el extractor de características a largo plazo 2046. En la presente idea inventiva, se propone calcular, como las características a largo plazo, al menos una de las estadísticas siguientes de las características a corto plazo desde el extractor de características a largo plazo 2022: media, varianza, media ponderada, varianza de ponderación, media alta, media baja y relación (contraste) entre la media alta y la media baja.

El valor medio y la varianza de las características corto plazo extraídas desde los segmentos a corto plazo en el segmento a largo plazo a clasificarse,

La media ponderada y la varianza de las características a corto plazo se extrajeron a partir de los segmentos a corto plazo en el segmento a largo plazo a clasificarse. Las características a corto plazo se ponderan sobre la base del grado de importancia de cada segmento a corto plazo que se mide con su energía o intensidad según se acaba de mencionar;

Media alta: una media de las características a corto plazo seleccionadas extraídas a partir de los segmentos a corto plazo en el segmento a largo plazo a clasificarse. Las características a corto plazo se seleccionan cuando cumplen al menos una de las condiciones siguientes: mayor que un valor umbral; o dentro de una proporción predeterminada de características a corto plazo no más bajas que todas las demás características a corto plazo, a modo de ejemplo, el 10 % más alto de las características a corto plazo; y

Media baja: una media de características a corto plazo seleccionadas que se extrajeron a partir de los segmentos a corto plazo en el segmento a largo plazo a clasificarse. Las características a corto plazo se seleccionan cuando al menos una de las condiciones siguientes: más pequeña que un valor umbral; o dentro de una proporción predeterminada de las características a corto plazo no más altas que todas las demás características a corto plazo, a modo de ejemplo, el más bajo 10 % de las características a corto plazo; y

Contraste: una relación entre la media alta y la media baja para representar la dinámica de las características a corto plazo en un segmento a largo plazo.

El extractor de características a corto plazo 2022 puede ponerse en práctica con las técnicas existentes, y se pueden extraer de este modo cualesquiera características. No obstante, se proponen algunas modificaciones para el extractor de características a corto plazo 2022 en la sección 6.3 siguiente.

6.3 Extracción de características a corto plazo

Según se ilustra en la Figura 24 y la Figura 27, el extractor de características a corto plazo 2022 puede configurarse para extraer, como características a corto plazo, al menos una de las siguientes características directamente a partir de cada segmento de audio a corto plazo: características rítmicas, características de interrupciones/silenciamientos y características de calidad de audio a corto plazo.

Las características rítmicas pueden incluir intensidad de ritmo, regularidad del ritmo, claridad del ritmo (véase el documento de L. Lu, D. Liu, y H.-J. Zhang. Titulado: *"Detección automática del estado anímico y seguimiento de señales de audio musicales"*. IEEE Transactions on Audio, Speech, and Language Processing, 14(1):5 - 18, 2006:, que se incorpora aquí en su integridad por referencia) y la modulación de sub-bandas en 2D M.F McKinney and J. Breebaart. *"Características para clasificación de audio y de música"*, Proc. ISMIR, 2003:, que se incorpora aquí en su integridad por referencia).

Las características de interrupciones/silenciamientos pueden incluir interrupciones de la voz, descensos agudos, duración de silenciamientos, silencio no natural, media de silencio no natural, energía total de silencio no natural, etc.

Las características de calidad de audio a corto plazo son características de calidad de audio con respecto a los segmentos a corto plazo, que son similares a las características de calidad de audio extraídas de las tramas de audio, que han de describirse a continuación.

Como alternativa o de forma adicional, según se ilustra en la Figura 28, el clasificador de audio 200 puede comprender un extractor de características a nivel de trama 2012 para extraer características a nivel de trama de cada una de las secuencias de tramas de audio incluidas en el segmento a corto plazo y el extractor de características a corto plazo 2022 puede configurarse para calcular las características corto plazo sobre la base de las características a nivel de trama extraídas a partir de la secuencia de tramas de audio.

Como pre-procesamiento, la señal de audio de entrada puede mezclarse con una señal de audio monoaural. El pre-procesamiento es innecesario si la señal de audio es ya una señal monoaural. A continuación, se divide en tramas con una longitud predefinida (normalmente de 10 a 25 milisegundos). En correspondencia, las características a nivel de trama se extraen desde cada trama.

El extractor de características a nivel de trama 2012 puede configurarse para extraer al menos una de las características siguientes: características que caracterizan las propiedades de varios tipos de audio a corto plazo, frecuencia de corte, características de relación de señal a ruido estática (SNR), características de relación de señal a ruido segmental (SNR), descriptores vocales básicos y características del tracto vocal.

Las características que caracterizan las propiedades de varios tipos de audio a corto plazo (en particular, voz, música a corto plazo, sonido de fondo y ruido) pueden comprender al menos una de las características siguientes. Energía de trama, distribución espectral de sub-bandas, flujo espectral, Coeficientes Cepstrales en las frecuencias de Mel (MFCC), bajos, información residual, característica de croma y tasa de cruce por cero.

Para conocer más detalles de los coeficientes MFCC, puede hacerse referencia al documento de L. Lu, H.-J. Zhang, y S. Li, *"Clasificación de audio basada en el contenido y segmentación utilizando máquinas vectoriales de soporte"*, ACM Multimedia Systems Journal 8 (6), páginas 482-492, marzo 2003 que se incorpora aquí en su integridad por referencia. Para conocer más detalles de la característica de croma, puede hacerse referencia al documento de G. H. Wakefield, *"Representación matemática de distribuciones de croma en tiempos conjuntos"* en SPIE, 1999 que se incorpora aquí en su integridad por referencia.

La frecuencia de corte representa la más alta frecuencia de una señal de audio por encima de la cual la energía del contenido está próxima a cero. Está diseñada para tetar un contenido de banda limitada, que es de utilidad en esta aplicación para clasificación de contexto de audio. Una frecuencia de corte suele estar causada por codificación, puesto que la mayoría de los codificadores desechan las altas frecuencias a tasas binarias bajas o medias. A modo de ejemplo, el códec de MP3 tiene una frecuencia de corte de 16 kHz a 128 kbps; a modo de otro ejemplo, los *codecs* de VoIP tienen una frecuencia de corte de 8 kHz o 16 kHz.

Además de la frecuencia de corte, la degradación de la señal durante el proceso de codificación de audio se considera como otra característica para diferenciar varios contextos de audio tales como VoIP vs. no VoIP, contextos de audio de alta calidad vs. baja calidad. Las características que representan la calidad de audio, tales como la evaluación de calidad de la voz operativa (véase documento de Ludovic Malfait, Jens Berger y Martin Kastner, titulado *"P.563- La norma de ITU-T para la evaluación de la calidad de la voz de extremo único"*, IEEE, Transaction on Audio, Speech and Language Processing, vol. 14, nº 6, noviembre 2006 que se incorpora aquí en su integridad por referencia), puede extraerse, además, en múltiples niveles para captar características de mayor contenido. Ejemplos de las características de calidad de audio incluyen:

a) Características de relación SNR estáticas, incluyendo nivel de ruido de fondo estimado, claridad espectral, etc.

b) Características de SNR segmentales incluyendo desviación de nivel espectral, gama de nivel espectral, nivel inferior de ruido relativo, etc.

c) Descriptores básicos de la voz incluyendo media del tono, variación de nivel de sección de la voz, nivel de la voz, etc.

d) Características del tracto vocal, incluyendo robotización, potencia a través del tono, etc.

Para derivar las características a corto plazo a partir de las características a nivel de trama, el extractor de características a corto plazo 2022 puede configurarse para calcular datos estadísticos de las características a nivel de trama, como las características a corto plazo.

Ejemplos de las estadísticas de las características a nivel de trama incluyen el valor medio y la desviación estándar, que captan las propiedades rítmicas para diferenciar varios tipos de audio, tales como música a corto plazo, voz, sonido de fondo y ruido. A modo de ejemplo, la voz suele alternar entre sonidos vocales y no vocales a una tasa de sílabas,

mientras que la música no lo hace, lo que indica que la variación de las características a nivel de trama de la voz suele ser mayor que la variación de la música.

Otro ejemplo de las estadísticas es la media ponderada de las características a nivel de trama. A modo de ejemplo, para la frecuencia de corte, la media ponderada entre las frecuencias de corte derivadas de cada trama de audio en un segmento a corto plazo, con la energía o intensidad de cada trama como ponderación, sería la frecuencia de corte para ese segmento a corto plazo.

Como alternativa o de forma adicional, según se ilustra en la Figura 29, el clasificador de audio 200 puede comprender un extractor de característica a nivel de trama 2012 para extraer características a nivel de trama a partir de las tramas de audio un clasificador a nivel de trama 2014 para clasificar cada una de las secuencias de tramas de audio en tipos de audio a nivel de trama utilizando las características a nivel de trama respectivas, en donde el extractor de características a corto plazo 2022 puede configurarse para calcular las características a corto plazo sobre la base de los resultados del clasificador a nivel de trama 2014 con respecto a la secuencia de las tramas de audio.

Dicho de otro modo, además del clasificador de contenido de audio 202 y el clasificador de contexto de audio 204, el clasificador de audio 200 puede comprender, además, un clasificador de tramas 201. En dicha arquitectura, el clasificador de contenido de audio 202 clasifica un segmento a corto plazo sobre la base de los resultados de la clasificación a nivel de trama del clasificador de tramas 201 y el clasificador de contexto de audio 204 clasifica un segmento a largo plazo sobre la base de los resultados de la clasificación a corto plazo del clasificador de contenido de audio 202.

El clasificador a nivel de trama 2014 puede configurarse para clasificar cada una de las secuencias de tramas de audio en cualesquiera clases, que pueden referirse como “tipos de audio a nivel de trama”. En una forma de realización, los tipos de audio a nivel de trama pueden tener una arquitectura similar a la arquitectura de los tipos de contenidos descritos con anterioridad y tienen también un significado similar a los tipos de contenidos, y la única diferencia es los tipos de audio a nivel de trama y los tipos de contenidos que se clasifican a diferentes niveles de la señal de audio, esto es, a nivel de trama y a nivel de segmento a corto plazo. A modo de ejemplo, el clasificador a nivel de trama 2014 puede configurarse para clasificar cada una de las secuencias de tramas de audio en al menos uno de los tipos de audio a nivel de trama siguiente: voz, música, sonido de fondo y ruido. Por otro lado, los tipos de audio a nivel de tema pueden tener también una arquitectura parcial o completamente distinta de la arquitectura de los tipos de contenidos, más adecuada para la clasificación a nivel de trama, y más adecuada para utilizarse como las características a corto plazo para la clasificación a corto plazo. A modo de ejemplo, el clasificador a nivel de trama 2014 puede configurarse para clasificar cada una de las secuencias de tramas de audio en al menos uno de los tipos de audio a nivel de trama siguientes: con voz, sin voz y pausa.

Con respecto a cómo derivar características a corto plazo a partir de los resultados de la clasificación a nivel de trama, se puede adoptar un sistema similar haciendo referencia a la descripción contenida en la sección 6.2.

Como una alternativa, las características a corto plazo basadas en los resultados del clasificador a nivel de trama 2014 y las características a corto plazo directamente basadas en las características a nivel de tramas obtenidas por el extractor de características a nivel de trama 2012 pueden utilizarse por el clasificador a corto plazo 2024. Por lo tanto, el extractor de características a corto plazo 2022 puede configurarse para calcular las características a corto plazo sobre la base de las características a nivel de trama extraídas a partir de la secuencia de las tramas de audio y de los resultados del clasificador a nivel de trama con respecto a la secuencia de las tramas de audio.

Dicho de otro modo, el extractor de características a nivel de trama 2012 puede configurarse para calcular datos estadísticos similares a los descritos en la sección 6.2 y las características a corto plazo descritas en relación con la Figura 28, incluyendo al menos una de las características siguientes: características que definen las propiedades de varios tipos de audio a corto plazo, la frecuencia de corte, las características de la relación señal a ruido estáticas, características de la relación señal a ruido por segmentos, descriptores de la voz básicos y características del tracto vocal.

Para trabajar en tiempo real, en todas las formas de realización, el extractor de características a corto plazo 2022 puede configurarse para funcionar sobre los segmentos de audio a corto plazo formados con un deslizamiento de ventana móvil en la dimensión temporal del segmento de audio a largo plazo en una longitud de etapa predeterminada. Con respecto a la ventana móvil para el segmento de audio a corto plazo, así como para la trama de audio y la ventana móvil para el segmento de audio a largo plazo, puede hacerse referencia a la sección 1.1 para conocer más detalles.

6.4 Combinación de formas de realización y escenarios de aplicación

De modo similar a la Parte 1, todas las formas de realización y variantes anteriormente descritas pueden ponerse en práctica en cualquiera de sus combinaciones y cualesquiera componentes mencionados en diferentes partes/formas de realización, pero teniendo las mismas o similares funciones que puedan ponerse en práctica como los mismos o componentes separados.

A modo de ejemplo, cualesquiera dos o más soluciones descritas en las secciones 6.1 a 6.3 pueden combinarse entre sí. Y cualquiera de las combinaciones puede combinarse, además, con cualquier forma de realización descrita o implícita en las Partes 1 a 5 y las otras partes que se describirán más adelante. En particular, la unidad de alisado de tipo 712 descrita en la Parte 1 puede utilizarse en esta Parte como una componente del clasificador de audio 200, para el alisado de los resultados del clasificador de tramas 2014, o el clasificador de contenido de audio 202 o el clasificador de contexto de audio 204. Además, el temporizador 916 puede servir también como un componente del clasificador de audio 200 para evitar un cambio brusco de la salida del clasificador de audio 200.

6.5 Método de clasificación de audio

De forma similar a la Parte 1, en el proceso de describir el clasificador de audio en las formas de realización anteriormente descritas, evidentemente se dan a conocer también algunos procesos o métodos. A continuación se proporciona un resumen de estos métodos con repetir algunos de los detalles ya descritos con anterioridad.

En una forma de realización, según se ilustra en la Figura 30, se da a conocer un método de clasificación de audio. Para identificar el tipo de audio a largo plazo (es decir, el tipo de contexto) de un segmento de audio a largo plazo incluido en una secuencia de segmentos de audio a corto plazo (solapados o no solapados entre sí), los segmentos de audio a corto plazo se clasifican en primer lugar (operación 3004) en tipos de audio a corto plazo, es decir, tipos de contenidos y las características a largo plazo se obtiene calculando (operación 3006) las estadísticas de los resultados de la operación de clasificación con respecto a la secuencia de los segmentos a corto plazo en el segmento de audio a largo plazo. A continuación, la clasificación a largo plazo (operación 3008) puede realizarse utilizando las características a largo plazo. El segmento de audio a corto plazo puede incluir una secuencia de tramas de audio. Por supuesto, para identificar el tipo de audio a corto plazo de los segmentos a corto plazo, necesitan extraerse las características a corto plazo a partir de dichos segmentos (operación 3002).

Los tipos de audio a corto plazo (tipos de contenidos) pueden incluir, sin limitación, a la voz, música a corto plazo, sonido de fondo y ruido.

Las características a largo plazo pueden incluir, sin limitación, a: valor medio y varianza de los valores de confianza de los tipos de audio a corto plazo, el valor medio y la varianza ponderados por el grados de importancia de los segmentos a corto plazo, la frecuencia de ocurrencia de cada tipo de audio a corto plazo y la frecuencia de transición entre diferentes tipos de audio a corto plazo.

En una variante, según se ilustra en la Figura 31, pueden obtenerse características a largo plazo adicionales (operación 3107) directamente sobre la base de las características a corto plazo de la secuencia de segmentos a corto plazo en el segmento de audio a largo plazo. Dichas características adicionales a largo plazo pueden incluir, sin limitación, a las estadísticas siguientes de las características a corto plazo: valor medio, varianza, media ponderada, varianza de ponderación, media alta, media baja y relación entre la media alta y la media baja.

Existen diferentes maneras para extraer las características a corto plazo. Una es extraer directamente las características a corto plazo a partir del segmento de audio a corto plazo a clasificarse. Tales características incluyen, sin limitación, características rítmicas, características de interrupciones/silenciamiento y características de calidad de audio a corto plazo.

La segunda manera es extraer las características a nivel de trama a partir de las tramas de audio incluidas en cada segmento a corto plazo (operación 3201 en la Figura 32), y luego, calcular las características a corto plazo sobre la base de las características a nivel de trama, tal como calcular los datos estadísticos de las características a nivel de trama como las características a corto plazo. Las características a nivel de trama pueden comprender, sin limitación, a: características que definen las propiedades de varios tipos de audio a corto plazo, la frecuencia de corte, las características de relación señal a ruido estáticas, las características de la relación señal a ruido por segmentos, descriptores de la voz básicos y características del tracto vocal. Las características que definen las propiedades de varios tipos de audio a corto plazo pueden comprender, además, la energía de trama, la distribución espectral de sub-bandas, el flujo espectral, coeficientes cepstrales en la frecuencia de Mel, bajos, información residual, característica de croma y tasa de cruce por cero.

La tercera forma consiste en extraer las características a corto plazo en una manera similar a la extracción de las características a largo plazo: después de extraer las características a nivel de trama a partir de las tramas de audio en un segmento a corto plazo a clasificarse (operación 3201), clasificar cada trama de audio en tipos de audio a nivel de trama utilizando las características a nivel de trama respectivas (operación 32011 en la Figura 33); y las características a corto plazo pueden extraerse (operación 3002) calculando las características a corto plazo sobre la base de los tipos de audio a nivel de trama (incluyendo, de modo opcional, los valores de confianza). Los tipos de audio a nivel de trama pueden tener propiedades y una arquitectura similar al tipo de audio a corto plazo (tipo de contenido), y pueden incluir también a la voz, música, sonido de fondo y ruido.

La segunda y la tercera formas pueden combinarse juntas según se ilustra por la flecha de línea de trazos en la Figura 33.

Según se describió en la Parte 1, los segmentos de audio a corto plazo y los segmentos de audio a largo plazo pueden muestrearse con ventanas móviles. Es decir, la operación de extraer características a corto plazo (operación 3002) puede realizarse sobre segmentos de audio a corto plazo formados con un deslizamiento de la ventana móvil en la dimensión temporal del segmento de audio a largo plazo con una longitud de tono predeterminada, y la operación de extraer características a largo plazo (operación 3107) y la operación de calcular estadísticas de tipos de audio a corto plazo (operación 3006) pueden realizarse también sobre los segmentos de audio a largo plazo formados con un deslizamiento de la ventana móvil en la dimensión temporal de la señal de audio en una longitud de tono predeterminada.

De modo similar a las formas de realización del aparato de procesamiento de audio, cualquier combinación de las formas de realización del método de procesamiento de audio y sus variantes son prácticas por un lado; y por otro lado, cada aspecto de las formas de realización del método de procesamiento de audio y sus variantes pueden ser soluciones separadas. Además, cualesquiera dos o más soluciones descritas en esta sección pueden combinarse entre sí, y estas combinaciones pueden combinarse, además, con cualquier forma de realización descrita o implícita en las otras partes de esta idea inventiva. En particular, según se describió ya en la sección 6.4, los sistemas de alisado y el sistema de transición de los tipos de audio pueden ser una parte del método de clasificación de audio aquí descrito.

Parte 7: Clasificadores de VoIP y métodos de clasificación

En la Parte 6 se propone un nuevo clasificador de audio para clasificar una señal de audio en tipos de contextos de audio al menos basados en parte sobre los resultados de la clasificación de tipos de contenidos. En las formas de realización descritas en la Parte 6, las características a largo plazo se extraen a partir de un segmento a largo plazo de una longitud de varios segundos a varias decenas de segundos, con lo que la clasificación del contexto de audio puede causar una larga latencia. Es deseable que el contexto de audio pueda clasificarse también en tiempo real o casi en tiempo real, tal como al nivel de segmentos a corto plazo.

7.1 Clasificación de contextos basada en el segmento a corto plazo

Por lo tanto, según se ilustra en la Figura 34, un clasificador de audio 200A se da a conocer, que incluye un clasificador de contenido de audio 202A para identificar un tipo de contenido de un segmento a corto plazo de una señal de audio y un clasificador de contexto de audio 204A para identificar un tipo de contexto del segmento a corto plazo al menos basado en parte en el tipo de contenido identificado por el clasificador de contenido de audio.

En este caso, el clasificador de contenido de audio 202A puede adoptar las técnicas ya mencionadas en la Parte 6 pero puede adoptar también diferentes técnicas según se describirá más adelante en la sección 7.2. Además, el clasificador de contexto de audio 204A puede adoptar las técnicas ya mencionadas en la Parte 6, con una diferencia de que el clasificador de contextos 204A puede utilizar directamente los resultados del clasificador de contenido de audio 202A, en lugar de utilizar los datos estadísticos de los resultados procedentes del clasificador de contenido de audio 202A puesto que el clasificador de contexto de audio 204A y el clasificador de contenido de audio 202A están clasificando el mismo segmento a corto plazo. Además, de forma similar a la Parte 6, además de los resultados procedentes del clasificador de contenido de audio 202A, el clasificador de contexto de audio 204A puede utilizar otras características directamente extraídas a partir del segmento a corto plazo. Es decir, el clasificador de contexto de audio 204A puede configurarse para clasificar el segmento a corto plazo basado en un modelo de aprendizaje de máquina utilizando, como características, los valores de confianza de los tipos de contenidos del segmento a corto plazo y otras características extraídas a partir del segmento a corto plazo. Con respecto a las características extraídas desde el segmento a corto plazo, puede hacerse referencia a la Parte 6.

El clasificador de contenido de audio 200A puede etiquetar simultáneamente el segmento a corto plazo puesto que más tipos de audio que la voz/ruido de VoIP y/o voz/ruido de no VoIP (voz/ruido de VoIP y voz/ruido de no VoIP que se describirán a continuación en la sección 7.2), y cada uno de los múltiples tipos de audio pueden tener su propio valor de confianza según se describe en la sección 1.2. Lo que antecede puede conseguir una mejor exactitud de la clasificación puesto que se puede obtener información más abundante. A modo de ejemplo, la información conjunta de los valores de confianza de la voz y de la música a corto plazo da a conocer en qué medida el contenido de audio es probable que sea una mezcla de voz y música de fondo, de modo que pueda discriminarse a partir del contenido de VoIP puro.

7.2 Clasificación utilizando la voz de VoIP y el ruido de VoIP

Este aspecto de la presente idea inventiva es de utilidad especial en un sistema de clasificación de VoIP/no VoIP, que requeriría clasificar el segmento a corto plazo actual con respecto a la latencia de decisión corta.

Para esta finalidad, según se ilustra en la Figura 34, el clasificador de audio 200A está especialmente diseñado para la clasificación de VoIP/no VoIP. Para clasificar VoIP/no VoIP, un clasificador de voz VoIP 2026 y/o un clasificador de ruido de VoIP se desarrollan para generar resultados intermedios para una clasificación sólida final de VoIP/no VoIP por el clasificador de contexto de audio 204A.

Un segmento a corto plazo de VoIP contendría voz de VoIP y ruido de VoIP alternativamente. Se observa que puede conseguirse una alta precisión para clasificar un segmento a corto plazo de voz en voz VoIP o voz no VoIP, pero sin clasificar un segmento a corto plazo de ruido en ruido de VoIP o ruido de no VoIP. De este modo, puede llegarse a la conclusión de que se hará confusa la discriminalidad clasificando directamente el segmento a corto plazo en VoIP (incluyendo voz de VoIP y ruido de VoIP pero sin identificar específicamente la voz de VoIP y el ruido de VoIP) y no VoIP sin considerar la diferencia entre la voz y el ruido y de este modo, con las características de estos dos tipos de contenidos (voz y ruido) mezclados juntos.

Es razonable para los clasificadores conseguir más altas precisiones para la clasificación de voz de VoIP/voz de no VoIP que para la clasificación de ruido de VoIP/ruido de no VoIP puesto que la voz contiene más información que ruido y dichas características tales como frecuencia de corte son más efectivas para clasificar la voz. En función de la ponderación obtenida a partir del proceso de formación de adaBoost, las características a corto plazo ponderadas superiores para la clasificación de voz de VoIP/no VoIP son: desviación estándar de la energía logarítmica, frecuencia de corte, desviación estándar de intensidad rítmica y desviación estándar del flujo espectral. La desviación estándar de la energía logarítmica, desviación estándar de la intensidad rítmica y desviación estándar del flujo espectral suelen ser más altas para la voz de VoIP que para la voz de no VoIP. Una razón probable es que numerosos segmentos de voz a corto plazo en un contexto de no VoIP, tal como un dispositivo de multimedia similar a cine o un juego se suelen mezclar con otros sonidos tales como música de fondo o efecto sonoro, cuyos valores de las características anteriores son más bajos. Asimismo, la frecuencia de corte suele ser más baja para la voz de VoIP que para la voz de no VoIP, lo que indica la baja frecuencia de corte introducida por los numerosos y populares *códecs* de VoIP.

Por lo tanto, en una forma de realización, el clasificador de contenido de audio 202A puede incluir un clasificador de voz de VoIP 2026 para clasificar el segmento a corto plazo en la voz de VoIP de tipos de contenidos o la voz de no VoIP de tipos de contenidos; y el clasificador de contexto de audio 204A puede configurarse para clasificar el segmento a corto plazo en el VoIP de tipo de contexto o el no VoIP de tipo de contexto sobre la base de los valores de confianza de la voz de VoIP y de la voz de no VoIP.

En otra forma de realización, el clasificador de contenido de audio 202A puede comprender, además, un clasificador de ruido de VoIP 2028 para clasificar el segmento a corto plazo en el ruido de VoIP de tipo de contenido o el ruido de no VoIP del tipo de contenido; y el clasificador de contexto de audio 204A puede configurarse para clasificar el segmento a corto plazo en el VoIP de tipo de contexto o el no VoIP de tipo de contexto sobre los valores de confianza de la voz de VoIP, voz de no VoIP, ruido de VoIP y ruido de no VoIP.

Los tipos de contenidos de la voz de VoIP, voz de no VoIP, ruido de VoIP y ruido de no VoIP pueden identificarse con las técnicas existentes según se indica en la Parte 6, sección 1.2 y sección 7.1.

Como alternativa, el clasificador de contenido de audio 202A puede tener una estructura jerárquica según se ilustra en la Figura 35. Es decir, se tendrá la ventaja de los resultados procedentes de un clasificador de voz/ruido 2025 para clasificar primero el segmento a corto plazo en voz o ruido/sonido de fondo.

Sobre la base de la forma de realización utiliza simplemente un clasificador de voz de VoIP 2026, si se determina un segmento a corto plazo como voz por el clasificador de voz/ruido 2025 (en tal situación, es solamente un clasificador de voz), entonces, el clasificador de voz de VoIP 2026 sigue clasificando si es la voz de VoIP o la voz de no VoIP, y calcula el resultado de la clasificación binaria; de no ser así, puede considerarse que el valor de confianza de voz de VoIP es de nivel bajo o la decisión sobre la voz de VoIP es incierta.

Sobre la base de la forma de realización que utiliza simplemente el clasificador de ruido de VoIP 2028, si se determina el segmento a corto plazo como un ruido, por el clasificador de voz/ruido 2025 (en tal situación, es solamente un clasificador de ruido (de fondo)), entonces, el clasificador de ruido de VoIP 2028 sigue clasificando el ruido de VoIP o ruido de no VoIP, y calculando el resultado de la clasificación binaria. De no ser así, puede considerarse que el valor de confianza del ruido de VoIP es de nivel bajo, o la decisión del ruido de VoIP es incierta.

En este caso, puesto que la voz suele ser de un tipo de contenido informativo y el ruido/sonido de fondo es un tipo de contenido interferente, aun cuando si un segmento a corto plazo no es un ruido, en la forma de realización en el apartado anterior no se puede determinar, por supuesto, que el segmento a corto plazo no sea del VoIP de tipo de contexto. Mientras que si un segmento a corto plazo no es una voz, en la forma de realización simplemente utilizando el clasificador de voz de VoIP 2026 es probablemente no la VoIP de tipo de contexto. Por lo tanto, por lo general, la forma de realización que utiliza simplemente un clasificador de voz de VoIP 2026 puede realizarse de forma independiente, mientras que la otra forma de realización que utiliza simplemente un clasificador de ruido de VoIP 2028 puede utilizarse como una forma de realización suplementaria que coopera con, a modo de ejemplo, la forma de realización que utiliza el clasificador de voz de VoIP 2026.

Es decir, el clasificador de voz de VoIP 2026 y el clasificador de ruido de VoIP 2028 pueden utilizarse a este respecto. Si un segmento a corto plazo se determina como de voz por el clasificador de voz/ruido 2025, entonces, el clasificador de voz de VoIP 2026 sigue clasificando si es voz de VoIP o voz de no VoIP, y calcula el resultado de la clasificación binaria. Si el segmento a corto plazo se determina como ruido por el clasificador de voz/ruido 2025, entonces, el clasificador de

ruido de VoIP 2028 sigue clasificándolo en ruido de VoIP o ruido de no VoIP, y calcula el resultado de la clasificación binario. De no ser así, puede considerarse que el segmento a corto plazo puede clasificarse como no VoIP.

En la puesta en práctica del clasificador de voz/ruido 2025, el clasificador de voz de VoIP 2026 y el clasificador de ruido de VoIP 2028 pueden adoptarse cualesquiera técnicas existentes y pueden ser el clasificador de contenido de audio 202 según se describe en las Partes 1 a 6.

Si el clasificador de contenido de audio 202A puesto en práctica de conformidad con la descripción anterior clasifica finalmente un segmento a corto plazo en ninguno de voz, ruido y sonido de fondo, o ninguno de voz de VoIP, voz de no VoIP, ruido de VoIP y ruido de no VoIP, ello significa que todos los valores de confianza pertinentes son de nivel bajo, y entonces, el clasificador de contenido de audio 202A (y el clasificador de contexto de audio 204A) puede clasificar el segmento a corto plazo como de no VoIP.

Para clasificar el segmento a corto plazo en los tipos de contextos de VoIP o no VoIP sobre la base del so resultados del clasificador de voz de VoIP 2026 y el clasificador de ruido de VoIP 2028, el clasificador de contexto de audio 204A puede adoptar técnicas basadas en el aprendizaje de máquina según se describe en la sección 7.1 y como una modificación, pueden utilizarse más características, incluyendo las características a corto plazo directamente extraídas desde el segmento a corto plazo y/o resultados de otros clasificadores de contenido de audio orientados a otros tipos de contenidos que los tipos de contenidos relacionados con VoIP, según se describió con anterioridad en la sección 7.1.

Además de las técnicas basadas en el aprendizaje de máquina anteriormente descritas, un método alternativo para la clasificación de VoIP/no VoIP puede ser una regla heurística que obtiene ventaja del conocimiento del dominio y que utiliza los resultados de la clasificación en relación con la voz de VoIP y el ruido de VoIP. Un ejemplo de dichas reglas heurísticas se ilustrará a continuación.

En el segmento a corto plazo actual de tiempo *t* se determina como voz de VoIP o voz de no VoIP, el resultado de la clasificación se toma directamente como el resultado de la clasificación de VoIP/no VoIP puesto que la clasificación de voz de VoIP/no VoIP es sólida según se indicó con anterioridad. Es decir, si el segmento a corto plazo se determina como voz de VoIP, entonces, es el VoIP de tipo de contexto; si el segmento a corto plazo se determina como voz de no VoIP, entonces, se trata de no VoIP de tipo de contexto.

Cuando el clasificador de voz de VoIP 2026 toma una decisión binaria con respecto a la voz de VoIP/voz de no VoIP con respecto a la voz determinada por el clasificador de voz/ruido 2025 según se describió con anterioridad, los valores de confianza de voz de VoIP y de voz de no VoIP podrían ser complementarios, es decir, su suma es 1 (si 0 representa un 100 % no y 1 representa un 100 % sí) y los valores umbrales del valor de confianza para diferenciar voz de VoIP y voz de no VoIP pueden indicar realmente el mismo punto. Si el clasificador de voz de VoIP 2026 no es un clasificador binario, los valores de confianza de voz de VoIP y de voz de no VoIP podrían no ser complementarios y los valores umbrales del valor de confianza para diferenciar la voz de VoIP y la voz de no VoIP pueden no indicar necesariamente el mismo punto.

Sin embargo, en el caso en donde la voz de VoIP o la voz de no VoIP tiene una confianza próxima y fluctúa alrededor del valor umbral, los resultados de clasificación de VoIP/no VoIP es posible que se conmuten con demasiada frecuencia. Para evitar dicha fluctuación, un sistema de memorización intermedia puede proveerse a este respecto: ambos valores umbral para voz de VoIP y voz de no VoIP pueden establecerse de mayor magnitud, de modo que no sea fácil la conmutación desde el tipo de contenido actual al otro tipo de contenido. Para facilidad de descripción, se puede convertir el valor de confianza para la voz de no VoIP al valor de confianza de la voz de VoIP. Es decir, si el valor de confianza es alto, el segmento a corto plazo se considera como más próximo a la voz de VoIP y si el valor de confianza es bajo, el segmento a corto plazo se considera como más próximo a la voz de no VoIP. Aunque para un clasificador no binario, según se describió con anterioridad, un valor de confianza alto de voz de no VoIP no significa necesariamente un valor de confianza bajo de la voz de VoIP, dicha simplificación puede reflejar bien la esencia de la solución y las reivindicaciones pertinentes descritas con el lenguaje de clasificadores binarios deberá interpretarse como cubriendo las soluciones equivalentes para clasificadores no binarios.

El sistema de memorización intermedia se ilustra en la Figura 36. Existe un área de memoria intermedia entre dos valores umbrales Th1 y Th2 (Th1>=Th2). Cuando el valor de confianza v(t) de la voz de VoIP cae dentro del área, no cambiará la clasificación de contexto, según se ilustra por las flechas en los lados izquierdo y derecho en la Figura 36. Solamente cuando el valor de confianza v(t) es mayor que el valor umbral Th1 de mayor magnitud, se clasificará el segmento a corto plazo como de VoIP (según se ilustra por la flecha en la parte inferior de la Figura 36); y solamente cuando el valor de confianza no es mayor que el umbral Th2 más pequeño, se clasificará el segmento a corto plazo como de no VoIP (según se ilustra por la flecha en la parte superior de la Figura 36).

Si el clasificador de ruido de VoIP 2028 se utiliza en cambio, la situación es similar. Para obtener la solución más sólida, el clasificador de voz de VoIP 2026 y el clasificador de ruido de VoIP 2028 pueden utilizarse conjuntamente. A continuación, el clasificador de contexto de audio 204A puede configurarse para: clasificar el segmento a corto plazo como el VoIP de tipo de contexto si el valor de confianza de la voz de VoIP es mayor que un primer valor umbral o si el valor de confianza del ruido de VoIP es mayor que un tercer valor umbral; clasificar el segmento a corto plazo como el no VoIP de tipo de contexto si el valor de confianza de la voz de VoIP no es mayor que un segundo valor umbral, en donde

el segundo valor umbral no es mayor que el primer valor umbral, o si el valor de confianza del ruido de VoIP no es mayor que un cuarto valor umbral, en donde el cuarto valor umbral no es mayor que el tercer valor umbral; de no ser así, clasificar el segmento a corto plazo como el tipo de contexto para el último segmento a corto plazo.

En este caso, un primer umbral puede ser igual al segundo umbral, y el tercer umbral puede ser igual al cuarto umbral, en particular, sin limitación, para el clasificador de voz de VoIP binario y el clasificador de ruido de VoIP binario. Sin embargo, puesto que, en general, el resultado de la clasificación de ruido de VoIP no es tan sólido, sería mejor si los tercero y cuarto umbrales no sean iguales entre sí y que estuvieran separados en 0.5 (0 indica una alta confianza para ser ruido de no VoIP y 1 indica una alta confianza para ser ruido de VoIP).

7.3 Alisado de la fluctuación

Para evitar una fluctuación rápida, otra solución es el alisado del valor de confianza según se determina por el clasificador de contenido de audio. Por lo tanto, según se ilustra en la Figura 37, una unidad de alisado de tipo 203A puede incluirse en el clasificador de audio 200A. Para el valor de confianza de cada uno de los 4 tipos de contenidos relacionados con VoIP, según se describió con anterioridad, los sistema de alisado descritos en la sección 1.3 pueden adoptarse a este respecto.

Como alternativa, de modo similar a la sección 7.2, la voz de VoIP y la voz de no VoIP pueden considerarse como un par que tienen valores de confianza complementarios; y el ruido de VoIP y el ruido de no VoIP pueden considerarse también como un par que tiene valores de confianza complementarios. En dicha situación, solamente uno de cada par necesita ser objeto de alisado y los sistemas de alisado descritos en la sección 1.3 pueden adoptarse a este respecto.

Tomando a modo de ejemplo el valor de confianza de *voz de VoIP*, puede expresarse de nuevo la fórmula (3) como:

$$v(t) = \beta \cdot v(t-1) + (1-\beta) \cdot voipSpeechConf\,(t) \qquad (3")$$

en donde v(t) es el valor de confianza de voz de VoIP objeto de alisado en el *tiempo t, v(t-1)* es el valor de confianza de *voz de VoIP* alisado en la última vez y *voipSpeechConf* es el valor de confianza de *voz de VoIP* en un tiempo actual *t* antes del alisado, siendo α un coeficiente de ponderación.

En una variante, si existe un clasificador de voz/ruido 2025 según se describió con anterioridad, si el valor de confianza de la voz para un segmento corto es bajo, entonces, el segmento a corto plazo no puede clasificarse como voz de VoIP de forma firme y se puede establecer directamente *voipSpeechConf (t)=v(t-1)* sin hacer que funcione realmente el clasificador de voz de VoIP 2026.

Como alternativa, en la situación descrita con anterioridad, se podría establecer *voipSpeechConf (t)* = 0.5 (u otro valor no superior a 0.5, tal como 0.4-0.5) que indica un caso incierto (en este caso, confianza = 1 indica una alta confianza de que sea VoIP y confianza = 0 indica una alta confianza de que no sea un VoIP).

Por lo tanto, en conformidad con una variante, según se ilustra en la Figura 37, el clasificador de contenido de audio 200A puede comprender, además, un clasificador de voz/ruido 2025 para identificar el tipo de contenido de voz del segmento a corto plazo, y la unidad de alisado de tipos 203A puede configurarse para establecer el valor de confianza de voz de VoIP para el presente segmento a corto plazo antes del alisado como un valor de confianza predefinido (tal como 0.5 u otro valor, tal como 0.4-0.5) o el valor de confianza alisado del último segmento a corto plazo, en donde el valor de confianza para la voz del tipo de contenido se clasifique por el clasificador de voz/ruido como más bajo que un quinto valor umbral. En dicha situación, el clasificador de voz de VoIP 2026 puede funcionar o no hacerlo. Como alternativa, el establecimiento del valor de confianza puede realizarse por el clasificador de voz de VoIP 2026, lo que es equivalente a la solución en donde se realiza el trabajo por la unidad de alisado de tipos 203A, y la reivindicación deberá interpretarse como que cubre ambas situaciones. Además, en este caso, se utiliza la expresión “el valor de confianza para la voz de tipo de contenido, según se clasifica por el clasificador de voz/ruido, es más baja que un quinto valor umbral” pero el alcance de protección no está limitado a este respecto y es equivalente a la situación en donde el segmento a corto plazo se clasifica en otros tipos de contenidos que el de la voz.

Para el valor de confianza del ruido de VoIP, la situación es similar y se omite aquí una descripción detallada.

Para evitar una fluctuación rápida, todavía otra solución es el alisado del valor de confianza según se determina por el clasificador de contexto de audio 204A, y los sistemas de alisado descritos en la sección 1.3 pueden adoptarse a este respecto.

Para evitar una fluctuación rápida, todavía otra solución es retardar la transición del tipo de contexto entre VoIP y no VoIP, y el mismo sistema que se describe en la sección 1.6 puede utilizarse en este caso. Según se describe en la sección 1.6, el temporizador 916 puede estar fuera del clasificador de audio o dentro del clasificador de audio como una parte del mismo. Por lo tanto, según se ilustra en la Figura 38, el clasificador de audio 200A puede comprender, además, el temporizador 916. Y el clasificador de audio se configura para continuar proporcionando, a la salida el tipo de contexto

actual hasta que la longitud del tiempo de duración de un nuevo tipo de contexto alcance el valor de un sexto valor umbral (el tipo de contexto es una instancia del tipo de contenido). Haciendo referencia a la sección 1.6, puede omitirse aquí una descripción detallada.

Como alternativa o de forma adicional, como otro sistema para retrasar la transición entre VoIP y no VoIP, el primero y/o segundo valor umbral según se describió con anterioridad para la clasificación de VoIP/no VoIP, puede ser diferente dependiendo del tipo de contexto del último segmento a corto plazo. Es decir, el primero y/o segundo valor umbral se hace mayor cuando el tipo de contexto del nuevo segmento a corto plazo es diferente del tipo de contexto del último segmento a corto plazo, mientras que se hace más pequeño cuando el tipo de contexto del nuevo segmento a corto plazo es el mismo que el tipo de contexto del último segmento. De este modo, el tipo de contexto tiende a mantenerse en el tipo de contexto actual y en consecuencia, una fluctuación brusca del tipo de contexto puede suprimirse en alguna medida.

7.4 Combinación de formas de realización y escenarios de aplicación

De forma similar a la Parte 1, todas las formas de realización y variantes anteriormente descritas pueden ponerse en práctica en cualquiera de sus combinaciones y cualesquiera componentes mencionados en diferentes partes/formas de realización, pero teniendo las mismas o funciones similares puede ponerse en práctica como los mismos o componentes separados.

A modo de ejemplo, cualesquiera dos o más soluciones descritas en las secciones 7.1 a 7.3 pueden combinarse entre sí. Y cualquiera de las combinaciones puede combinarse, además, con cualquier forma de realización descrita o implícita en las Partes 1 a 6. En particular, las formas de realización descritas en esta parte y cualquiera de sus combinaciones pueden combinarse con las formas de realización del aparato/método de procesamiento de audio o el controlador/método de control del nivelador de volumen descrito en la Parte 4.

7.5 Método de clasificación de VoIP

De modo similar a la Parte 1, en el proceso de describir el clasificador de audio en las formas de realización anteriormente descritas, se da a conocer evidentemente que existen también algunos procesos o métodos aplicables. A continuación, se proporciona un resumen de estos métodos sin repetir algunos de los detalles ya descritos con anterioridad.

En una forma de realización según se ilustra en la Figura 39, un método de clasificación de audio incluye la identificación de un tipo de contenido de un segmento a corto plazo de una señal de audio (operación 4004), identificando luego un tipo de contexto del segmento a corto plazo al menos en parte sobre la base del tipo de contenido que se identifica (operación 4008).

Para identificar el tipo de contexto de una señal de audio de forma dinámica y rápida, el método de clasificación de audio en esta parte es de utilidad particular en la identificación del VoIP y no VoIP del tipo de contexto. En tal situación, el segmento a corto plazo puede clasificarse primero en la voz de VoIP de tipo de contenido o la voz de no VoIP del tipo de contenido y la operación de identificar el tipo de contexto está configurada para clasificar el segmento a corto plazo en el VoIP de tipo de contexto o no VoIP de tipo de contexto sobre la base de los valores de confianza de la voz de VoIP y la voz de no VoIP.

Como alternativa, el segmento a corto plazo puede clasificarse primero en el ruido de VoIP de tipo de contenido o el ruido no vip de tipo de contenido, y la operación de identificar el tipo de contexto puede configurarse para clasificar el segmento a corto plazo en VoIP de tipo de contexto o de no VoIP de tipo de contexto sobre la base de los valores de confianza del ruido de VoIP y ruido de no VoIP.

La voz y el ruido pueden considerarse conjuntamente. En dicha situación, la operación de identificar el tipo de contexto puede configurarse para clasificar el segmento a corto plazo en el VoIP de tipo de contexto o el no VoIP de tipo de contexto sobre la base de los valores de confianza de la voz de VoIP, la voz de no VoIP, ruido de VoIP y ruido de no VoIP.

Para identificar el tipo de contexto del segmento a corto plazo, puede utilizarse un modelo de aprendizaje de máquina, considerando los valores de confianza de los tipos de contenidos del segmento a corto plazo y otras características extraídas a partir del segmento a corto plazo como características.

La operación de identificar el tipo de contexto puede realizarse también sobre la base de reglas heurísticas. Cuando solamente están implicadas la voz de VoIP y la voz de no VoIP, la regla heurística es como sigue: clasificar el segmento a corto plazo como el VoIP de tipo de contexto si el valor de confianza de la voz de VoIP es mayor que un primer valor umbral; clasificar el segmento a corto plazo como de no VoIP de tipo de contexto si el valor de confianza de la voz de VoIP no es mayor que un segundo valor umbral, en donde el segundo valor umbral no es mayor que el primer valor umbral; de no ser así, clasificar el segmento a corto plazo como el tipo de contexto para el último segmento a corto plazo.

La regla heurística para la situación en donde solamente se implica el ruido de VoIP y el ruido de no VoIP es similar.

Cuando la voz y el ruido están implicados, la regla heurística es como sigue: clasificar el segmento a corto plazo como VoIP de tipo de contexto si el valor de confianza de la voz de VoIP es mayor que un primer valor umbral o si el valor de confianza del ruido de VoIP es mayor que un tercer valor umbral; clasificar el segmento a corto plazo como de no VoIP de tipo de contexto si el valor de confianza de la voz de VoIP no es mayor que un segundo valor umbral, en donde el segundo valor umbral no es mayor que el primer valor umbral, o si el valor de confianza del ruido de VoIP no es mayor que un cuarto valor umbral, en donde el cuarto valor umbral no es mayor que un tercer valor umbral; de no ser así, clasificar el segmento a corto plazo como el tipo de contexto para el último segmento a corto plazo.

El sistema de alisado descrito en la sección 1.3 y sección 1.8 puede adoptarse aquí y se omite una descripción detallada. Como una modificación al sistema de alisado descrito en la sección 1.3 antes de la operación de alisado 4106, el método puede comprender, además, la identificación de la voz del tipo de contenido a partir del segmento a corto plazo (operación 40040 en la Figura 40), en donde el valor de confianza de la voz de VoIP para el segmento a corto plazo actual antes del alisado se establece como un valor de confianza predeterminado o el valor de confianza alisado del último segmento a corto plazo (operación 40044 en la Figura 40), en donde el valor de confianza para la voz de tipo de contenido es menor que un quinto umbral ("N" en la operación 40041).

Si, de no ser así, la operación de identificar la voz de tipo de contenido determina en firme el segmento a corto plazo como voz ("Y" en la operación 40041), entonces, el segmento a corto plazo se clasifica, además, en voz de VoIP o voz de no VoIP (operación 40042), antes de la operación de alisado 4106.

De hecho, incluso sin utilizar el sistema de alisado, el método puede identificar también la voz de tipo de contenido y/o el ruido en primer lugar, cuando el segmento a corto plazo se clasifica como voz o ruido, una clasificación adicional se realiza para clasificar el segmento a corto plazo en uno de voz de VoIP y voz de no VoIP y uno de ruido de VoIP y ruido de no VoIP. A continuación, se realiza la operación de identificar el tipo de contexto.

Según se indicó en la sección 1.6 y la sección 1.8 el sistema de transición aquí descrito puede tomarse como una parte del método de clasificación de audio también aquí descrito, y se omite los detalles. En resumen, el método puede comprender, además, la medida del tiempo de duración durante el cual se realiza la operación de identificar el tipo de contexto continuamente proporcionando el mismo tipo de contexto, en donde el método de clasificación de audio está configurado para continuar proporcionando el tipo de contexto actual hasta que la longitud del tiempo de duración de un nuevo tipo de contexto alcance un sexto valor umbral.

De modo similar, pueden establecerse seis umbrales diferentes para diferentes pares de transición desde un tipo de contexto a otro tipo de contexto. Además, el sexto umbral puede estar en correlación negativa con el valor de confianza del nuevo tipo de contexto.

Como una modificación al sistema de transición en el método de clasificación de audio especialmente orientado a la clasificación de VoIP/no VoIP, cualquier o más del primero al cuarto valor umbral para el presente segmento a corto plazo puede establecerse de forma diferente dependiendo del tipo de contexto del último segmento a corto plazo.

De modo similar a las formas de realización del aparato de procesamiento de audio, cualquier combinación de las formas de realización del método de procesamiento de audio y sus variantes son prácticas por un lado; y por otro lado, cualquier aspecto de las formas de realización del método de procesamiento de audio y sus variantes pueden ser soluciones separadas. Además, cualesquiera dos o más soluciones descritas en esta sección pueden combinarse entre sí y estas combinaciones pueden combinarse, además, con cualquier forma de realización descrita o implícita en las otras partes de esta idea inventiva. Más concretamente, el método de clasificación de audio aquí descrito puede utilizarse en el método de procesamiento de audio también aquí descrito, en particular, el método de control del nivelador de volumen.

Según se describe al principio de la descripción detallada de la presente invención, la forma de realización de la aplicación puede realizarse en hardware o en software o en ambos a la vez. La Figura 41 es un diagrama de bloques que ilustra un sistema ejemplo para poner en práctica los aspectos de la presente invención.

En la Figura 41, una unidad central de procesamiento (CPU) 4201 realiza varios procesos en conformidad con un programa memorizado en una memoria de solamente lectura (ROM) 4202 o un programa cargado desde una sección de memorización 4208 a una memoria de acceso auditorio (RAM) 4203. En la memoria RAM 4203, los datos adquiridos cuando la unidad CPU 4201 realiza los diversos procesos o similares, se memoriza también cuando se requiere.

La unidad CPU 4201, la memoria ROM 4202 y la memoria RAM 4203 están conectadas entre sí mediante un bus de conexión 4204. Una interfaz de entrada/salida 4205 está también conectada al bus 4204.

Los siguientes componentes están conectados a la interfaz de entrada/salida 4205: una sección de entrada 4206 que incluye un teclado, un ratón o similar; una sección de salida 4207 que incluye un monitor tal como un tubo de rayos catódicos (CRT), una pantalla de cristal líquido (LCD) o similar y un altavoz o similar; la sección de memorización 4208 que incluye un disco duro o similar y una sección de comunicaciones 4209 que incluye una tarjeta de interfaz de red tal

como una tarjeta de red LAN, un módem o similar. La sección de comunicaciones 4209 realiza un proceso de comunicaciones por intermedio de la red tal como Internet.

Una unidad 4210 está también conectada a la interfaz de entrada/salida 4205 cuando se requiere. Un soporte extraíble 4211, tal como un disco magnético, un disco óptico, un disco magneto-óptico, una memoria de semiconductores o similar, está montado en la unidad 4210 cuando se requiere, de modo que un programa informático allí leído sea instalado en la sección de memorización 4208 cuando se requiera.

En el caso en donde los componentes anteriormente descritos sean puestos en práctica por el software, el programa que constituye el software está instalado desde la red tal como Internet o el soporte de memorización tal como el soporte extraíble 4211.

Conviene señalar que la terminología aquí utilizada es para los fines de describir formas de realización particulares solamente y no está prevista para ser limitadora de la idea inventiva. Tal como aquí se utilizan las formas singulares "un", "una" y "el" están previstos para incluir las formas del plural también, a no ser que el contexto lo indique claramente de otro modo. Además, se entenderá que los términos "comprende" y/o "comprendiendo", cuando se utilizan en esta especificación, sirven para especificar la presencia de características, números enteros, operaciones, etapas, elementos y/o componentes establecidos, pero no excluyen la presencia o adición de una o más otras características, números enteros, operaciones, etapas, elementos, componentes y/o sus grupos.

Las correspondientes estructuras, materiales, actos y equivalentes de todos los medios o elementos de operación y de función en las reivindicaciones siguientes están previstas para intentar incluir cualquier estructura, material o acto para realizar la función en combinación con otros elementos reivindicados según se reivindica de forma específica. La descripción de la presente invención ha sido presentada para fines de ilustración y descripción, pero no está prevista para ser exhaustiva ni está limitada a la solicitud en la forma dada a conocer. Numerosas modificaciones y variantes serán evidentes para un experto en esta técnica sin desviarse por ello del alcance de protección de la solicitud de patente. La forma de realización fue elegida y descrita con el fin de explicar mejor los principios de la idea inventiva y la aplicación práctica y, para permitir a otros expertos en esta técnica entender la aplicación para varias formas de realización con diversas modificaciones que sean adecuadas para el uso particular previsto.

REIVINDICACIONES

1. Un controlador de ecualizador (2200) que comprende:
un clasificador de audio (200) para identificar continuamente el tipo de audio de una señal de audio que comprende un contenido de audio a identificarse; y
una unidad de ajuste (300) para ajustar un ecualizador (408) de una manera continua sobre la base de un valor de confianza del tipo de audio que se identifica, en donde el clasificador de audio está configurado para clasificar la señal de audio en múltiples tipos de audio con valores de confianza correspondientes, y la unidad de ajuste está configurada para considerar al menos algunos de los múltiples tipos de audio mediante la ponderación de los valores de confianza de los múltiples tipos de audio sobre la base de la importancia de los múltiples tipos de audio, representando el valor de confianza el nivel de coincidencia entre el contenido de audio a identificarse y un tipo de audio objetivo.

2. Un controlador de ecualizador (2200) que comprende:
un clasificador de audio (200) para identificar continuamente el tipo de audio de una señal de audio que comprende el contenido de audio a identificarse; y
una unidad de ajuste (300) para ajustar un ecualizador (408) de una manera continua sobre la base de un valor de confianza del tipo de audio que se identifica, en donde el clasificador de audio está configurador para clasificar la señal de audio en múltiples tipos de audio con valores de confianza correspondientes, y la unidad de ajuste está configurada para considerar al menos algunos de los múltiples tipos de audio mediante la ponderación de los efectos de los múltiples tipos de audio sobre la base de los valores de confianza, representando el valor de confianza el nivel de coincidencia entre el contenido de audio a identificarse y un tipo de audio objetivo

3. El controlador de ecualizador según la reivindicación 2, en donde la unidad de ajuste está configurada para considerar al menos un tipo de audio dominante sobre la base de los valores de confianza.

4. El controlador de ecualizador según la reivindicación 1, que comprende, además, una unidad de suavizado de parámetros para, con respecto a un parámetro del ecualizador ajustado por la unidad de ajuste, suavizar el valor de parámetro determinado por la unidad de ajuste en el momento actual sobre la base de los valores de los parámetros en el pasado.

5. El controlador de ecualizador según la reivindicación 4, en donde la unidad de suavizado de parámetros está configurada para determinar un valor de parámetro suavizado en el presente calculando una suma ponderada del valor de parámetro determinado por la unidad de ajuste en el momento actual y un valor de parámetro suavizado de la última vez.

6. El controlador de ecualizador según la reivindicación 5, en donde los pesos de ponderación para calcular la suma ponderada se cambian, de forma adaptativa, sobre la base de uno entre: el tipo de audio de la señal de audio; diferentes pares de transición desde un tipo de audio a otro tipo de audio; o un aumento o disminución de la tendencia del valor de parámetro determinado por la unidad de ajuste.

7. El controlador de ecualizador según cualquiera de las reivindicaciones 1 a 6, en donde
el clasificador de audio comprende un clasificador de contenido de audio para identificar el tipo de contenido de la señal de audio; y
la unidad de ajuste está configurada para correlacionar positivamente un nivel de ecualización con un valor de confianza de la música a corto plazo y/o correlacionar negativamente el nivel de ecualización con un valor de confianza de la voz.

8. El controlador de ecualizador según cualquiera de las reivindicaciones 1 a 6, en donde,
el clasificador de audio comprende un clasificador de contexto de audio para identificar el tipo de contexto de la señal de audio; y
la unidad de ajuste está configurada para correlacionar positivamente un nivel de ecualización con un valor de confianza de música a largo plazo y/o correlacionar negativamente el nivel de ecualización con un valor de confianza de los medios de comunicación de tipo película cinematográfica y/o del juego.

9. El controlador de ecualizador según cualquiera de las reivindicaciones 1 a 6, en donde
el clasificador de audio comprende un clasificador de contenido de audio para identificar el tipo de contenido de la señal de audio; y
la unidad de ajuste está configurada para correlacionar positivamente un nivel de ecualización con un valor de confianza de una música a corto plazo sin fuentes dominantes y/o correlacionar negativamente el nivel de ecualización con un valor de confianza de la música a corto plazo con fuentes dominantes.

10. El controlador de ecualizador según la reivindicación 7 u 8, en donde la unidad de ajuste está configurada para correlacionar positivamente el nivel de ecualización con un valor de confianza de música a corto plazo sin fuentes dominantes y/o correlacionar negativamente el nivel de ecualización con un valor de confianza de música a corto plazo con fuentes dominantes; y
de modo opcional, en donde la unidad de ajuste está configurada para considerar la música a corto plazo sin/con fuentes dominantes cuando el valor de confianza para la música a corto plazo es superior a un umbral.

11. El controlador de ecualizador según cualquiera de las reivindicaciones 1 a 6, en donde,
el clasificador de audio comprende un clasificador de contenido de audio para identificar el tipo de contenido de la señal de audio; y
la unidad de ajuste está configurada para correlacionar positivamente un nivel de ecualización con un valor de confianza del sonido de fondo y/o correlacionar negativamente el nivel de ecualización con un valor de confianza de ruido.

12. El controlador de ecualizador según cualquiera de las reivindicaciones 1 a 6, en donde la unidad de ajuste está configurada para asignar un nivel de ecualización y/o perfil de ecualización y/o equilibrio espectral predeterminado para cada tipo de audio.

13. El controlador de ecualizador según la reivindicación 12, en donde el clasificador de audio comprende un clasificador de contenido de audio para clasificar la señal de audio en un tipo de contenido a corto plazo que comprende al menos uno de entre la música a corto plazo, la voz, el sonido de fondo y el ruido, y
de modo opcional, en donde la música a corto plazo comprende al menos un agrupamiento musical, y
de modo opcional, en donde el al menos un agrupamiento musical comprende un agrupamiento basado en el género y/o un agrupamiento basado en el instrumento y/o
un agrupamiento musical clasificado sobre la base de ritmo, tempo, timbre de música y/o cualquier otro atributo musical.

14. El controlador de ecualizador según la reivindicación 12, en donde el clasificador de audio comprende un clasificador de contexto de audio para clasificar la señal de audio en un tipo de contexto a largo plazo que comprende al menos uno de entre medios de difusión de tipo película cinematográfica, música a largo plazo, VoIP y juego.

15. Un método de control del ecualizador que comprende:
identificar el tipo de audio de una señal de audio que comprende el contenido de audio a identificarse en tiempo real; y
ajustar un ecualizador (408) de una manera continua sobre la base de un valor de confianza del tipo de audio que se identifica, en donde la señal de audio se clasifica en múltiples tipos de audio con valores de confianza correspondientes, y la operación de ajuste está configurada para considerar al menos algunos de los múltiples tipos de audio mediante la ponderación de los valores de confianza de los múltiples tipos de audio sobre la base de la importancia de dichos múltiples tipos de audio, representando el valor de confianza el nivel de coincidencia entre el contenido de audio a identificarse y un tipo de audio objetivo.

16. Un método de control de ecualizador que comprende:
identificar el tipo de audio de una señal de audio que comprende el contenido de audio a identificarse en tiempo real; y
ajustar un ecualizador (408) de una manera continua sobre la base de un valor de confianza del tipo de audio que se identifica, en donde la señal de audio se clasifica en múltiples tipos de audio con valores de confianza correspondientes, y la operación de ajuste está configurada para considerar al menos algunos de los múltiples tipos de audio mediante la ponderación de los efectos de los múltiples tipos de audio sobre la base de los valores de confianza, representando el valor de confianza el nivel de coincidencia entre el contenido de audio a identificarse y un tipo de audio objetivo.

17. Un soporte legible por ordenador que tiene instrucciones de programas informáticos registradas que, cuando se ejecutan por un procesador, permiten al procesador ejecutar el método de control del ecualizador según la reivindicación 15 o la reivindicación 16.

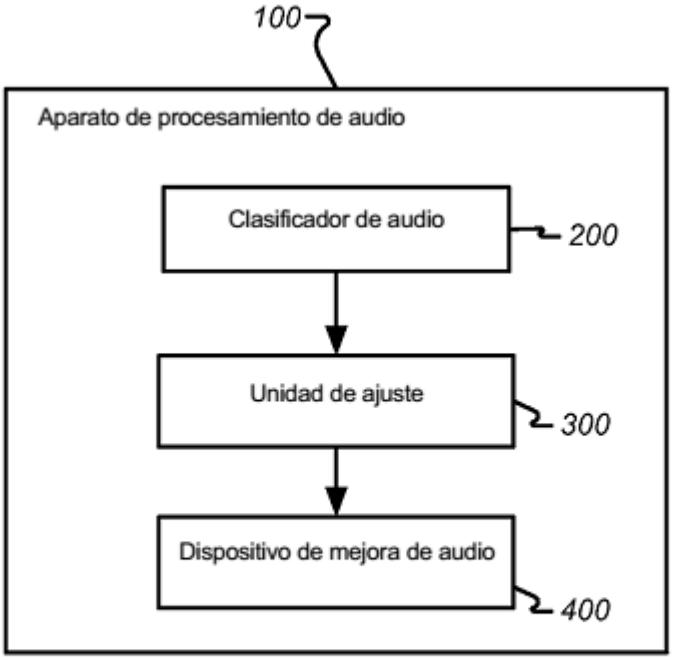
100
Aparato de procesamiento de audio
Clasificador de audio
200
Unidad de ajuste
300
Dispositivo de mejora de audio
400

FIG. 1

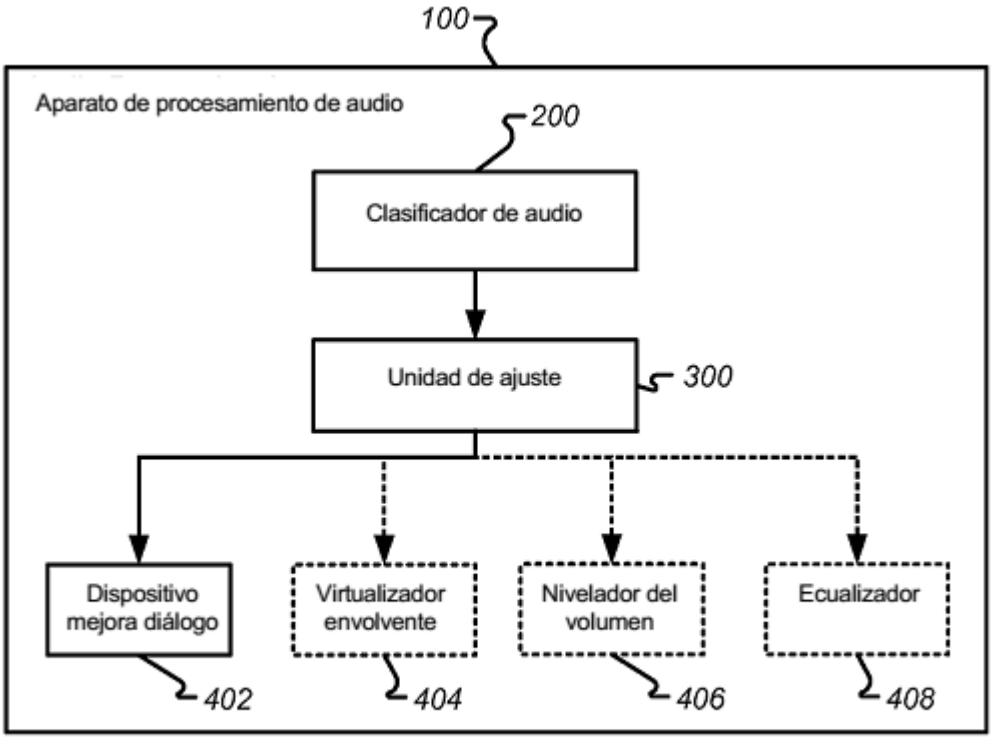
100
Aparato de procesamiento de audio
200
Clasificador de audio
Unidad de ajuste
300
Dispositivo mejora diálogo
Virtualizador envolvente
Nivelador del volumen
Ecualizador
402
404
406
408

FIG. 2

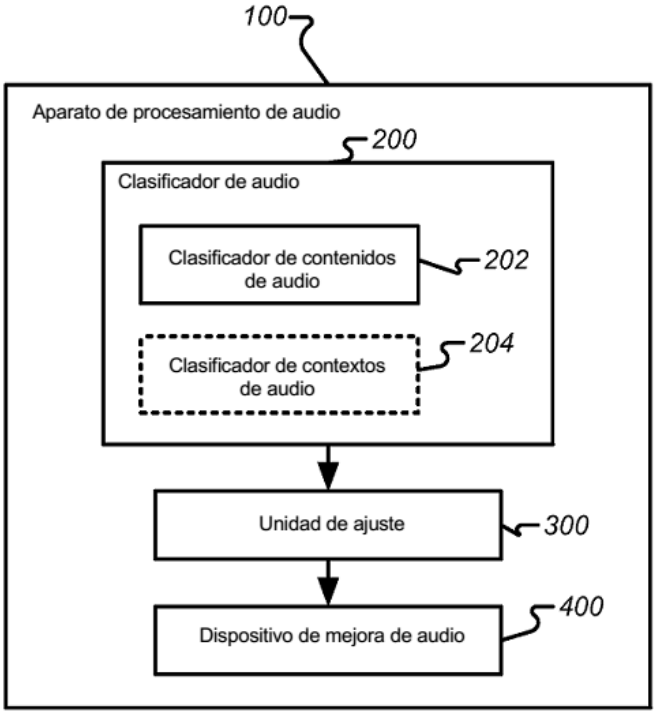
100
Aparato de procesamiento de audio
200
Clasificador de audio
Clasificador de contenidos de audio
202
Clasificador de contextos de audio
204
Unidad de ajuste
300
Dispositivo de mejora de audio
400

FIG. 3

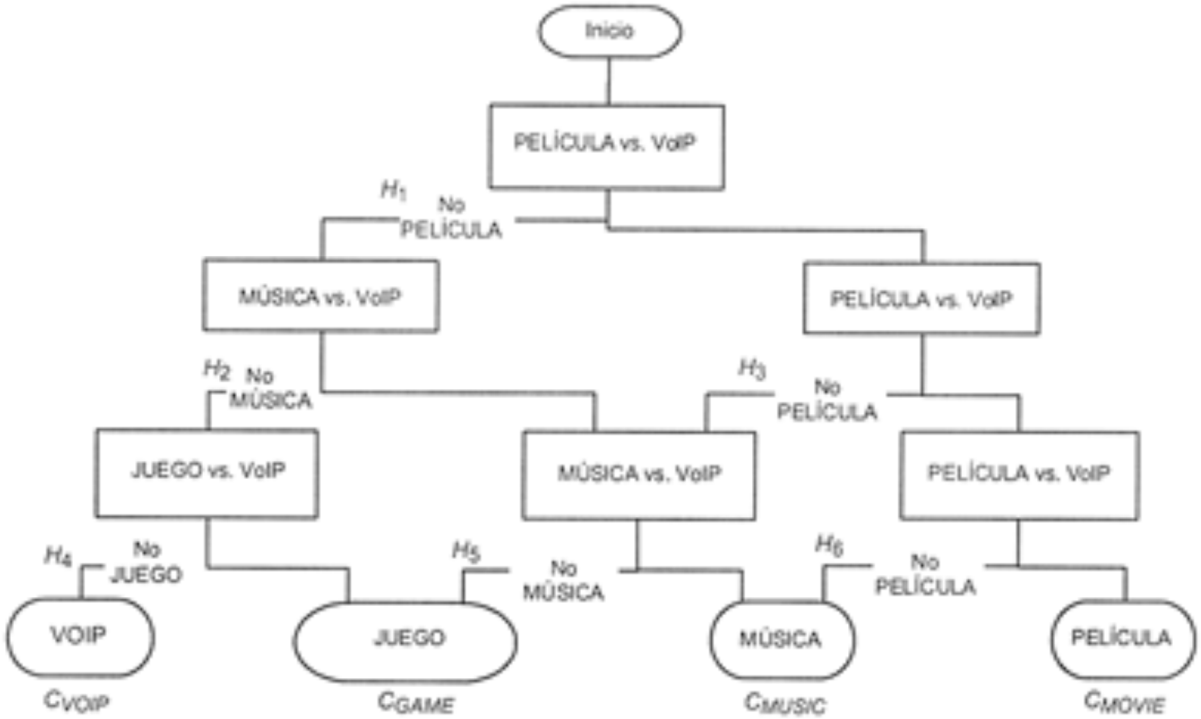
Inicio
PELÍCULA vs. VoIP
H1 No PELÍCULA
MÚSICA vs. VoIP
PELÍCULA vs. VoIP
H2 No MÚSICA
H3 No PELÍCULA
JUEGO vs. VoIP
MÚSICA vs. VoIP
PELÍCULA vs. VoIP
H4 No JUEGO
H5 No MÚSICA
H6 No PELÍCULA
VOIP
JUEGO
MÚSICA
PELÍCULA
CVOIP
CGAME
CMUSIC
CMOVIE

FIG. 4

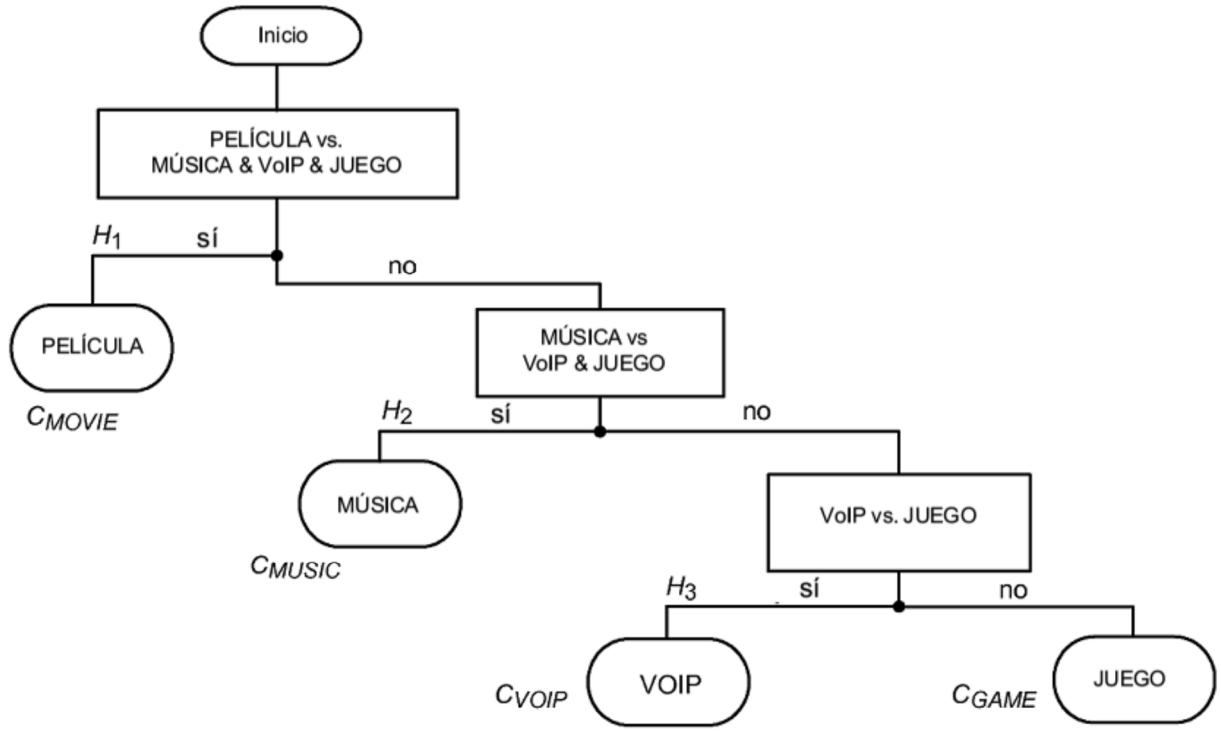
Inicio
PELÍCULA vs.
MÚSICA & VoIP & JUEGO
H_1
sí
no
PELÍCULA
C_{MOVIE}
MÚSICA vs
VoIP & JUEGO
H_2
sí
no
MÚSICA
C_{MUSIC}
VoIP vs. JUEGO
H_3
sí
no
C_{VOIP}
VOIP
C_{GAME}
JUEGO

FIG. 5

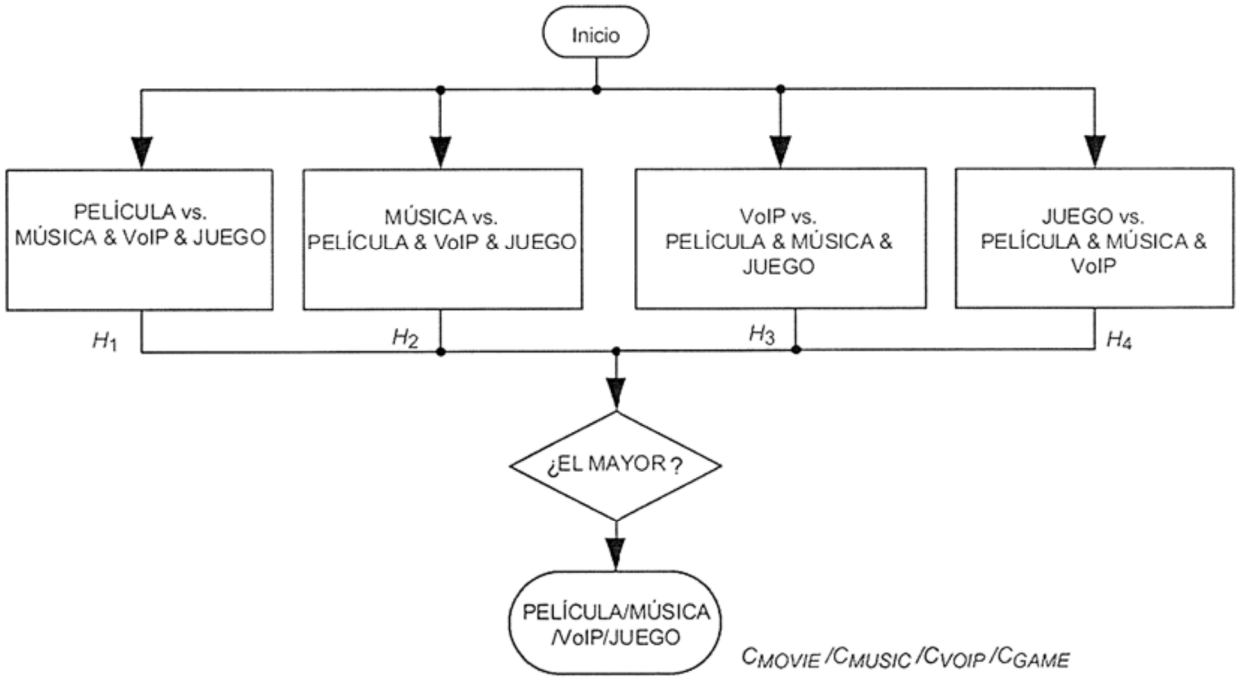
Inicio
PELÍCULA vs.
MÚSICA & VoIP & JUEGO
MÚSICA vs.
PELÍCULA & VoIP & JUEGO
VoIP vs.
PELÍCULA & MÚSICA &
JUEGO
JUEGO vs.
PELÍCULA & MÚSICA &
VoIP
H_1
H_2
H_3
H_4
¿EL MAYOR?
PELÍCULA/MÚSICA
/VoIP/JUEGO
C_{MOVIE} /C_{MUSIC} /C_{VOIP} /C_{GAME}

FIG. 6

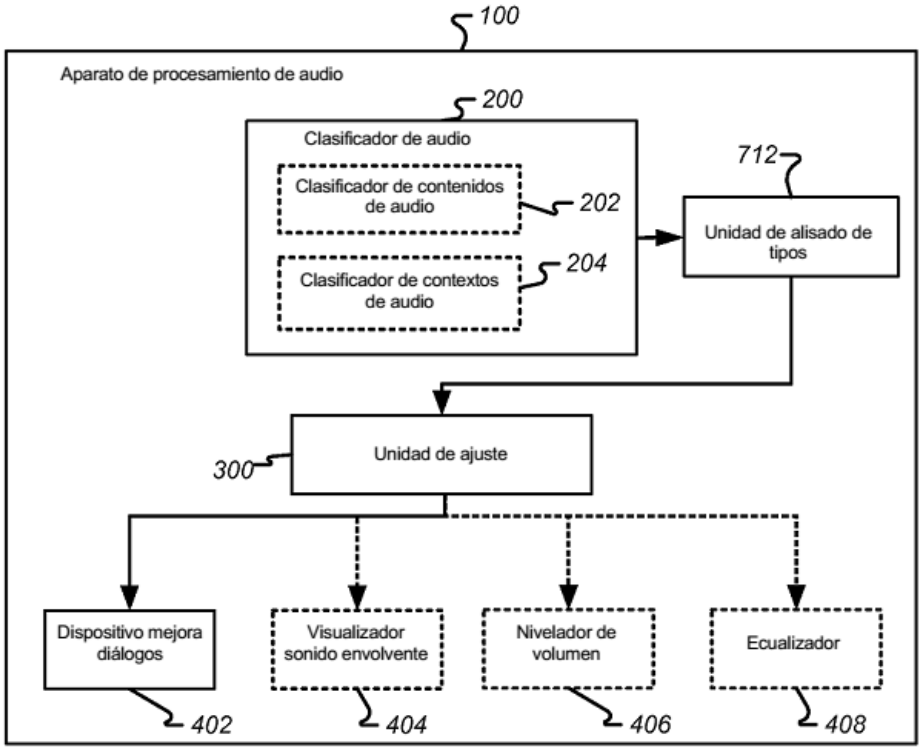
100
Aparato de procesamiento de audio
200
Clasificador de audio
Clasificador de contenidos de audio
202
Clasificador de contextos de audio
204
712
Unidad de alisado de tipos
300
Unidad de ajuste
Dispositivo mejora diálogos
Visualizador sonido envolvente
Nivelador de volumen
Ecualizador
402
404
406
408

FIG. 7

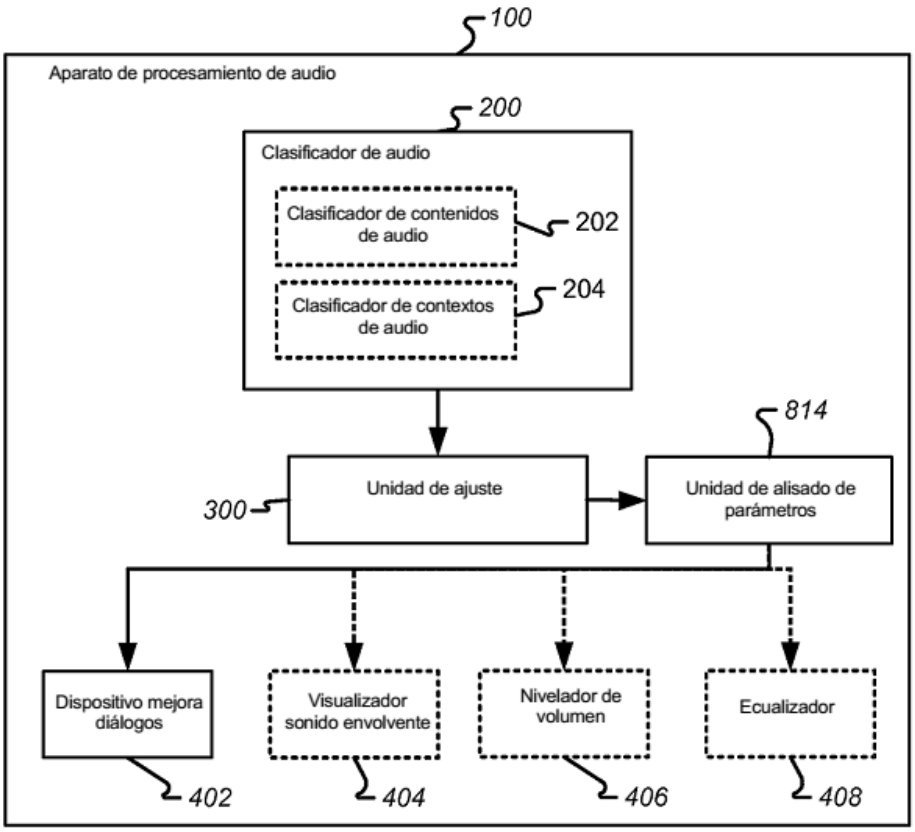
100
Aparato de procesamiento de audio
200
Clasificador de audio
Clasificador de contenidos de audio
202
Clasificador de contextos de audio
204
814
300
Unidad de ajuste
Unidad de alisado de parámetros
Dispositivo mejora diálogos
Visualizador sonido envolvente
Nivelador de volumen
Ecualizador
402
404
406
408

FIG. 8

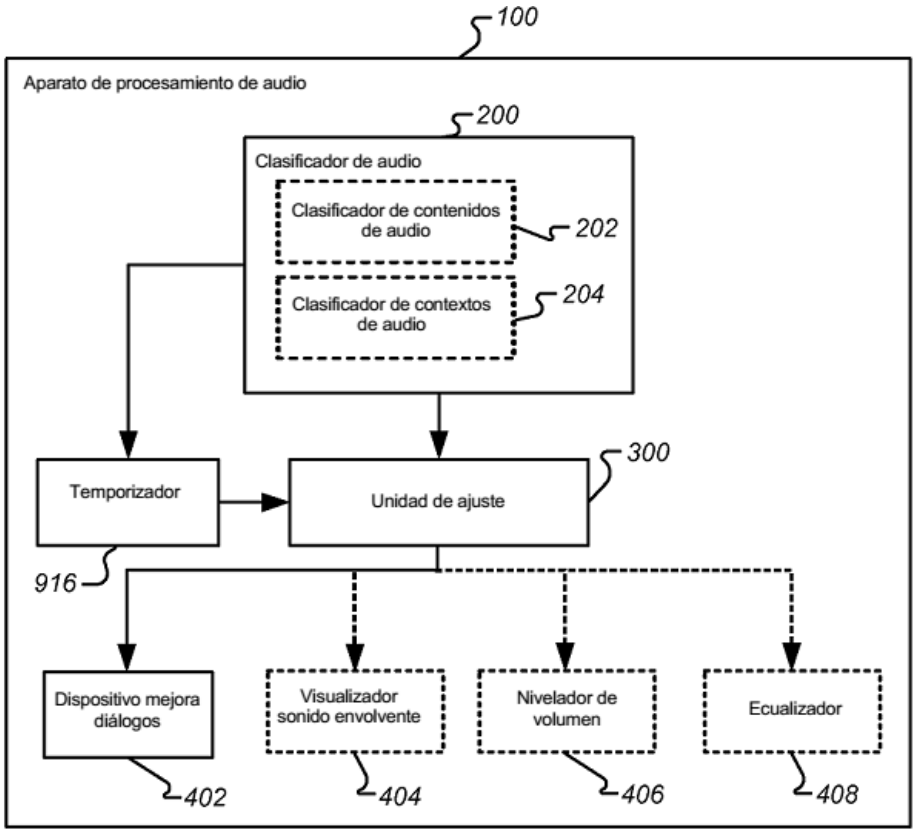
100
Aparato de procesamiento de audio
200
Clasificador de audio
Clasificador de contenidos de audio
202
Clasificador de contextos de audio
204
Temporizador
Unidad de ajuste
300
916
Dispositivo mejora diálogos
Visualizador sonido envolvente
Nivelador de volumen
Ecualizador
402
404
406
408

FIG. 9

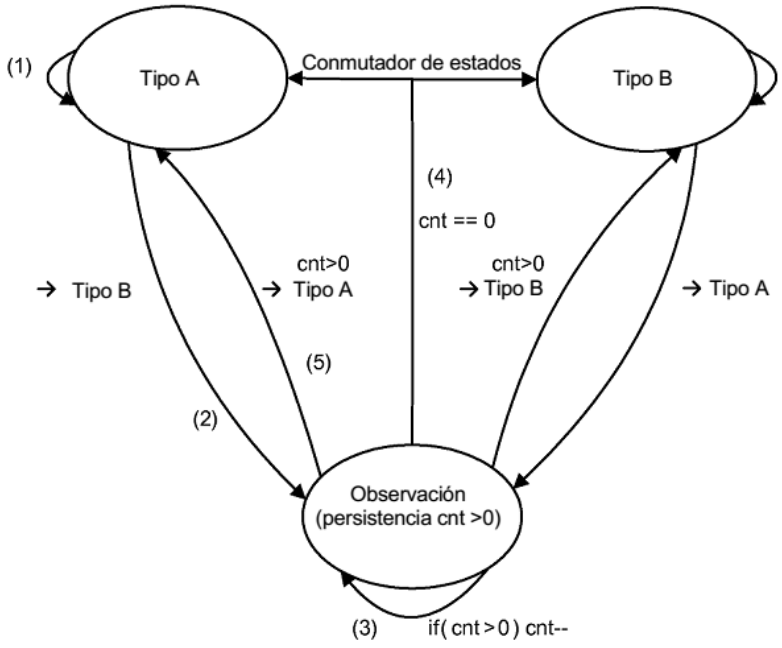
(1)
Tipo A
Conmutador de estados
Tipo B
(4)
cnt == 0
cnt>0
→ Tipo B
cnt>0
→ Tipo A
cnt>0
→ Tipo B
→ Tipo A
(5)
(2)
Observación
(persistencia cnt >0)
(3)
if(cnt > 0) cnt--

FIG. 10

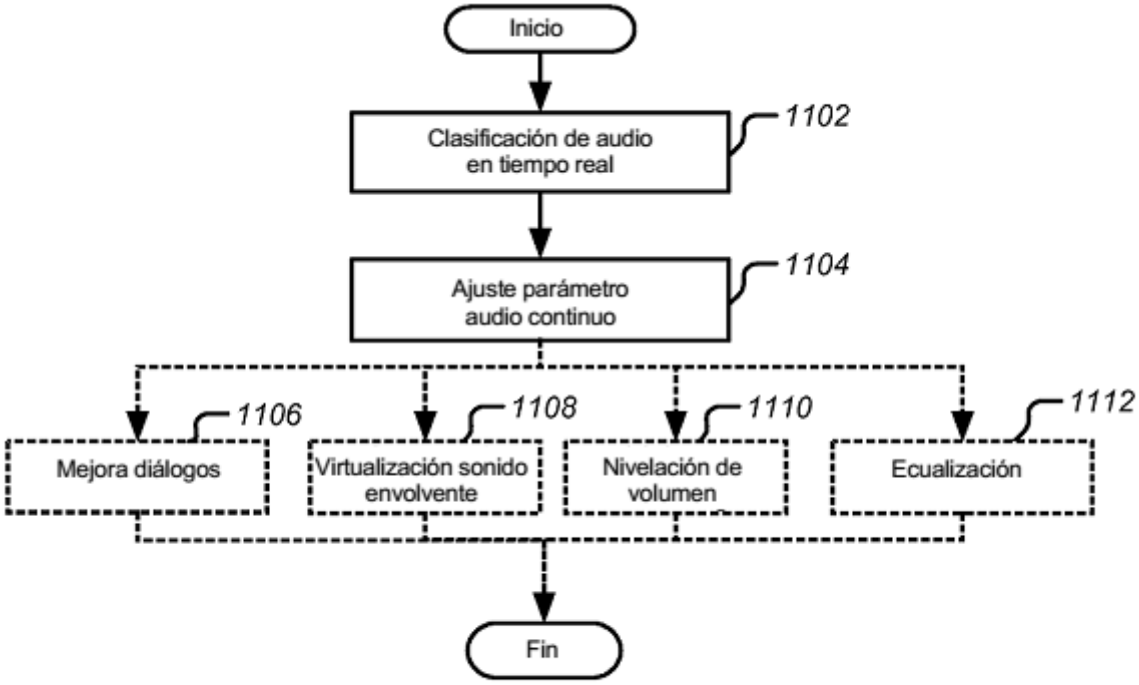
Inicio
Clasificación de audio en tiempo real
1102
Ajuste parámetro audio continuo
1104
Mejora diálogos
1106
Virtualización sonido envolvente
1108
Nivelación de volumen
1110
Ecualización
1112
Fin

FIG. 11

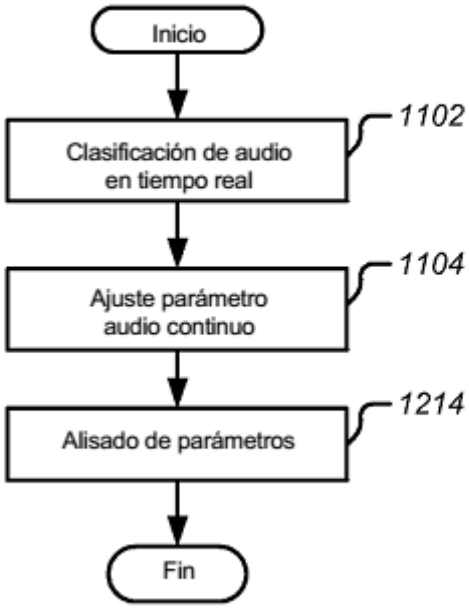
Inicio
Clasificación de audio en tiempo real
1102
Ajuste parámetro audio continuo
1104
Alisado de parámetros
1214
Fin

FIG. 12

v

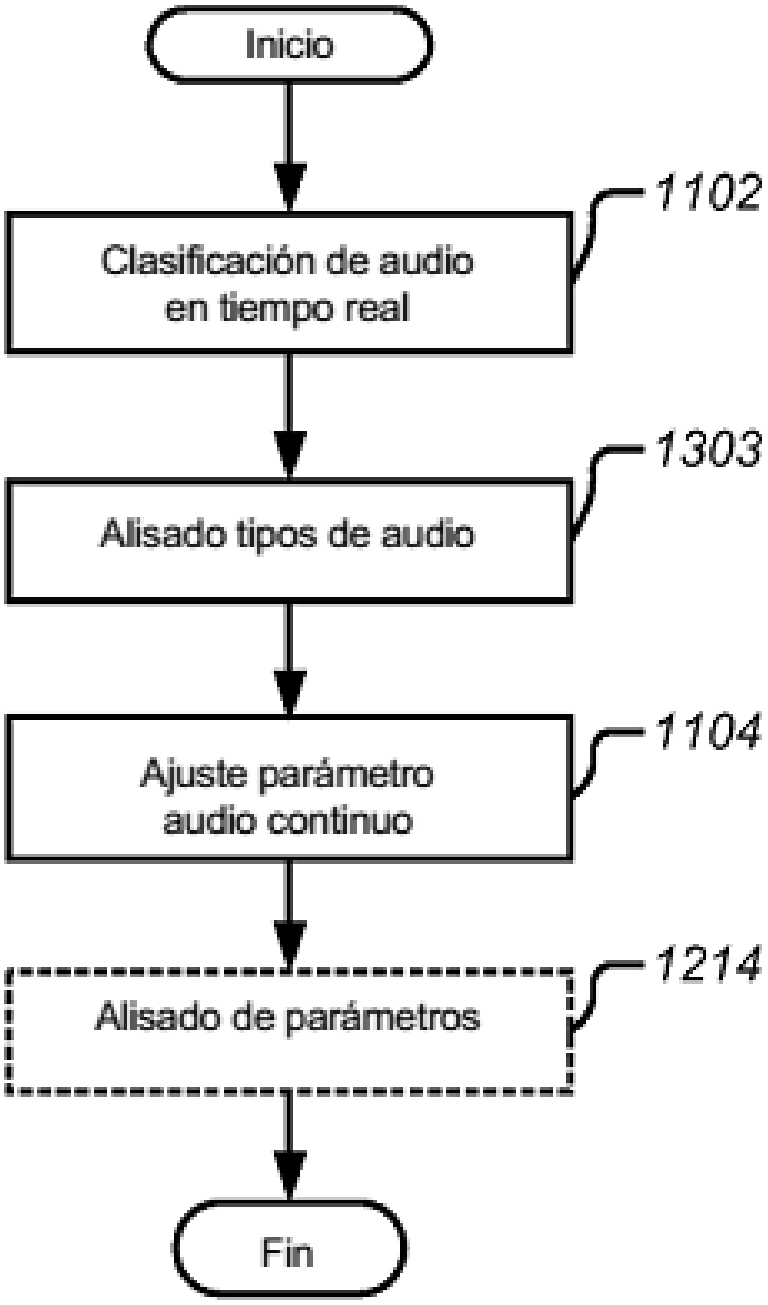
Inicio
1102
Clasificación de audio en tiempo real
1303
Alisado tipos de audio
1104
Ajuste parámetro audio continuo
1214
Alisado de parámetros
Fin

FIG. 13

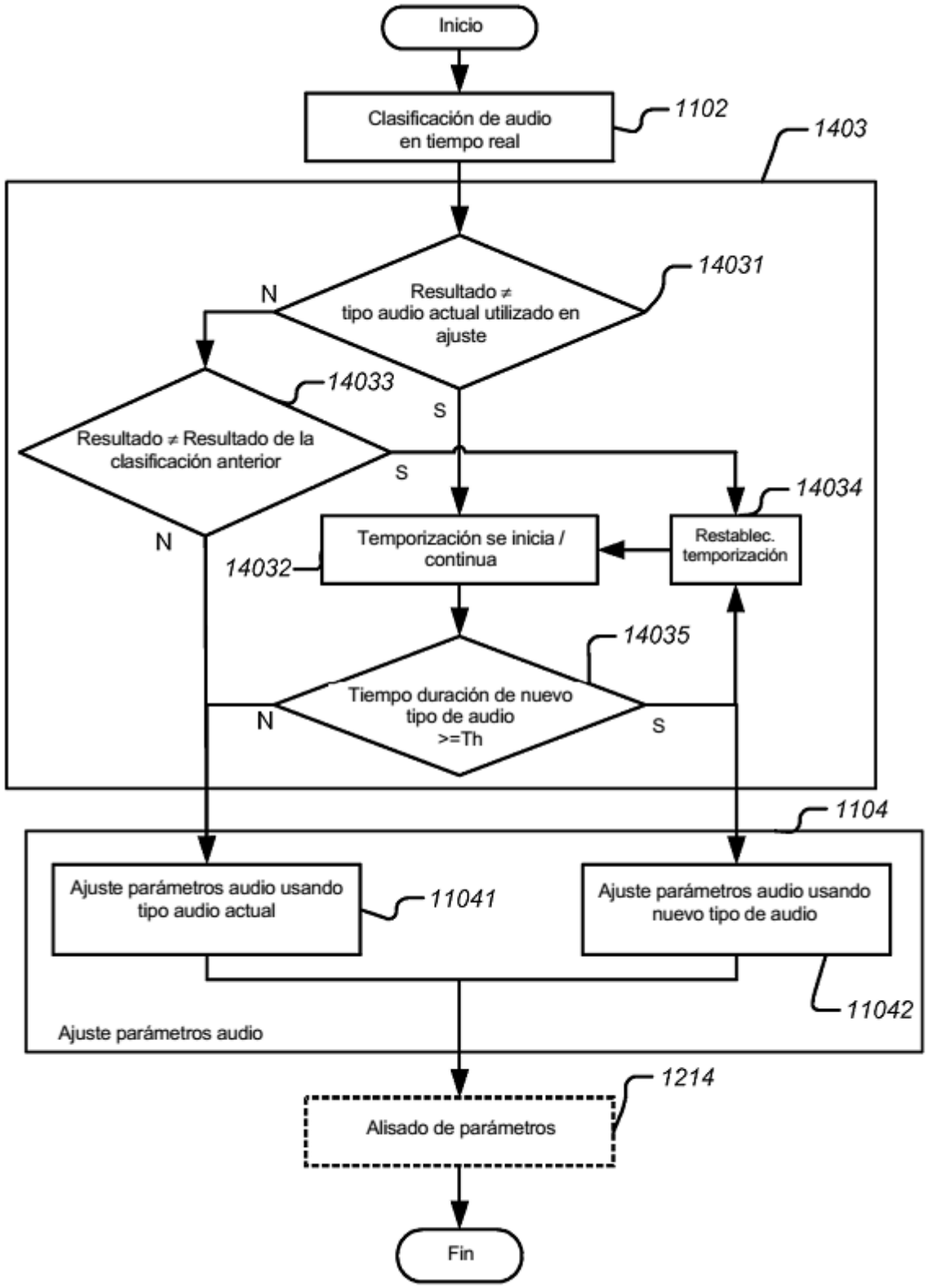
Inicio
Clasificación de audio en tiempo real
1102
1403
14031
Resultado ≠ tipo audio actual utilizado en ajuste
N
S
14033
Resultado ≠ Resultado de la clasificación anterior
S
N
14034
Restablec. temporización
Temporización se inicia / continua
14032
14035
Tiempo duración de nuevo tipo de audio >=Th
N
S
1104
Ajuste parámetros audio usando tipo audio actual
11041
Ajuste parámetros audio usando nuevo tipo de audio
11042
Ajuste parámetros audio
1214
Alisado de parámetros
Fin

FIG. 14

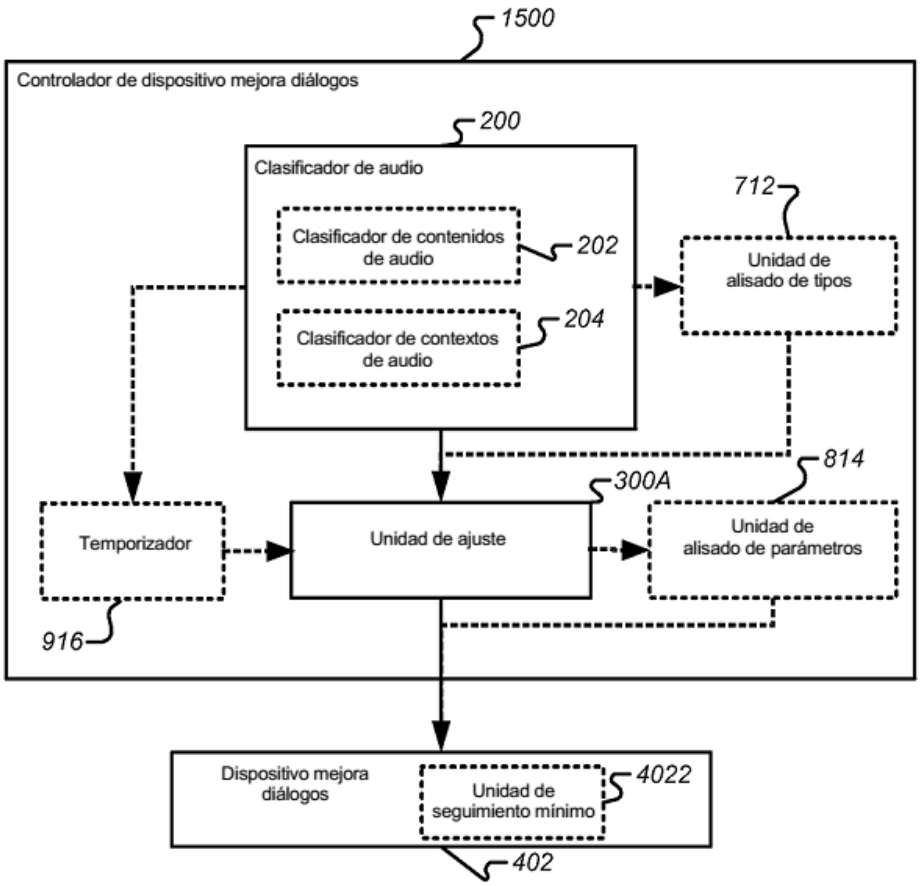
1500
Controlador de dispositivo mejora diálogos
200
Clasificador de audio
Clasificador de contenidos de audio
202
Clasificador de contextos de audio
204
712
Unidad de alisado de tipos
814
300A
Temporizador
Unidad de ajuste
Unidad de alisado de parámetros
916
Dispositivo mejora diálogos
Unidad de seguimiento mínimo
4022
402

FIG. 15

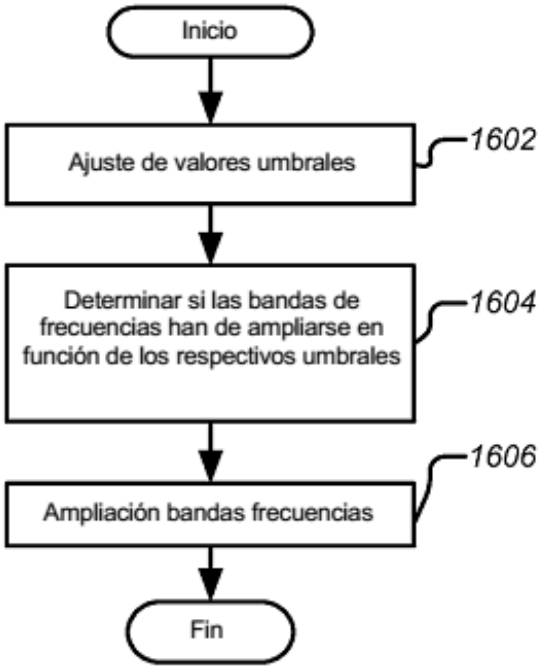
Inicio
Ajuste de valores umbrales
1602
Determinar si las bandas de frecuencias han de ampliarse en función de los respectivos umbrales
1604
Ampliación bandas frecuencias
1606
Fin

FIG. 16

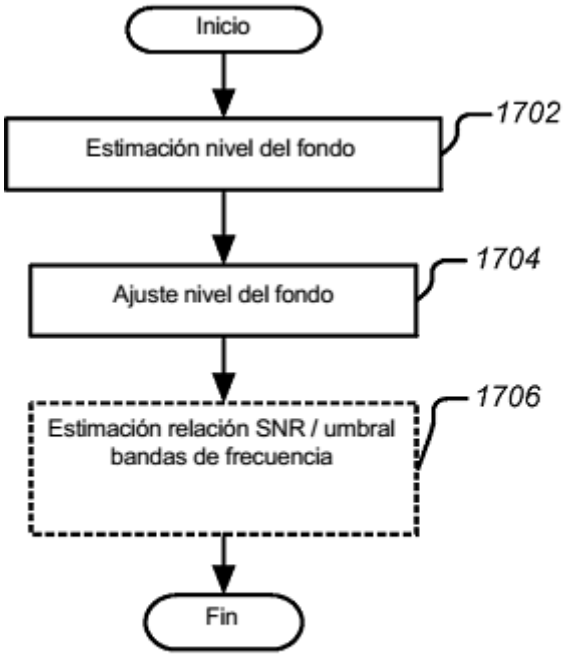
Inicio
Estimación nivel del fondo
1702
Ajuste nivel del fondo
1704
Estimación relación SNR / umbral bandas de frecuencia
1706
Fin

FIG. 17

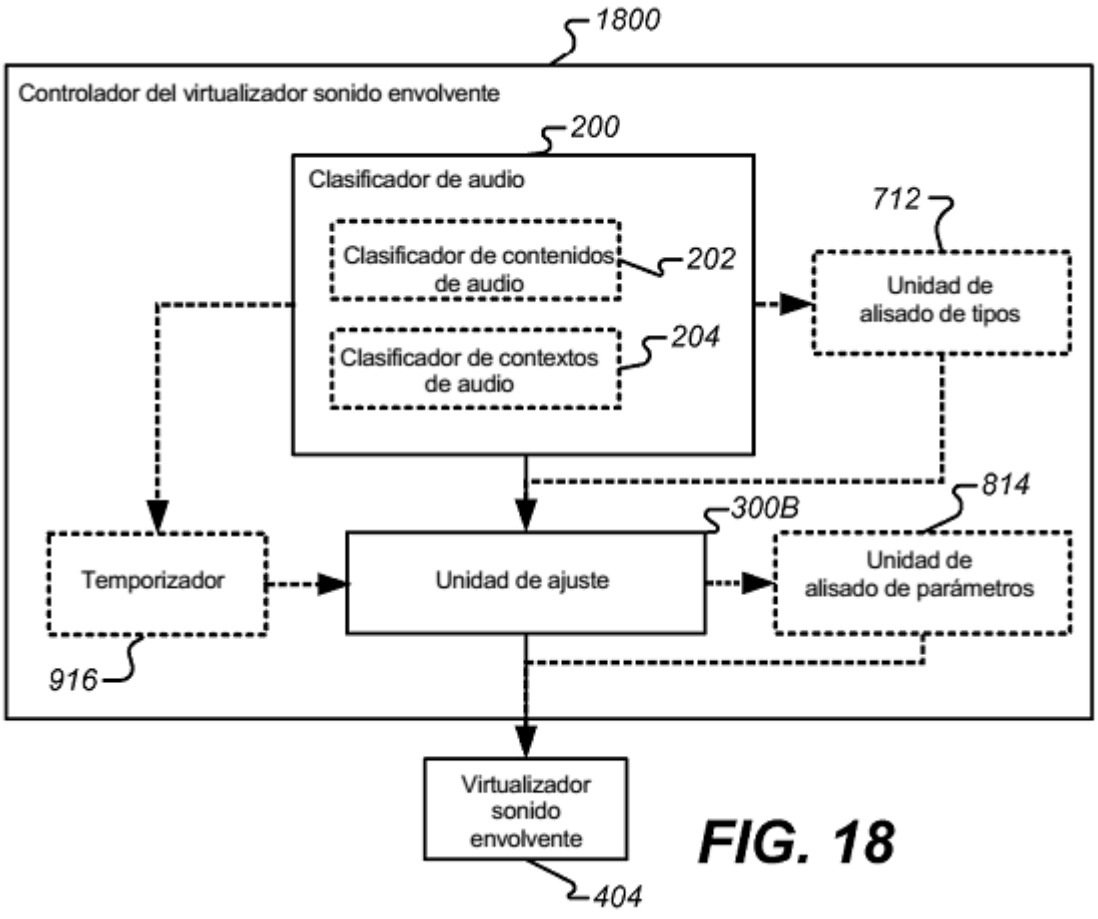
1800
Controlador del virtualizador sonido envolvente
200
Clasificador de audio
Clasificador de contenidos de audio
202
Clasificador de contextos de audio
204
712
Unidad de alisado de tipos
814
300B
Temporizador
Unidad de ajuste
Unidad de alisado de parámetros
916
Virtualizador sonido envolvente
404

FIG. 18

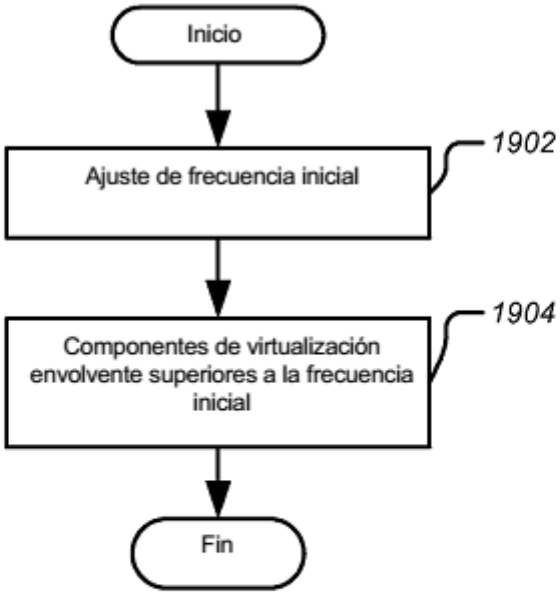
Inicio
Ajuste de frecuencia inicial
1902
Componentes de virtualización envolvente superiores a la frecuencia inicial
1904
Fin

FIG. 19

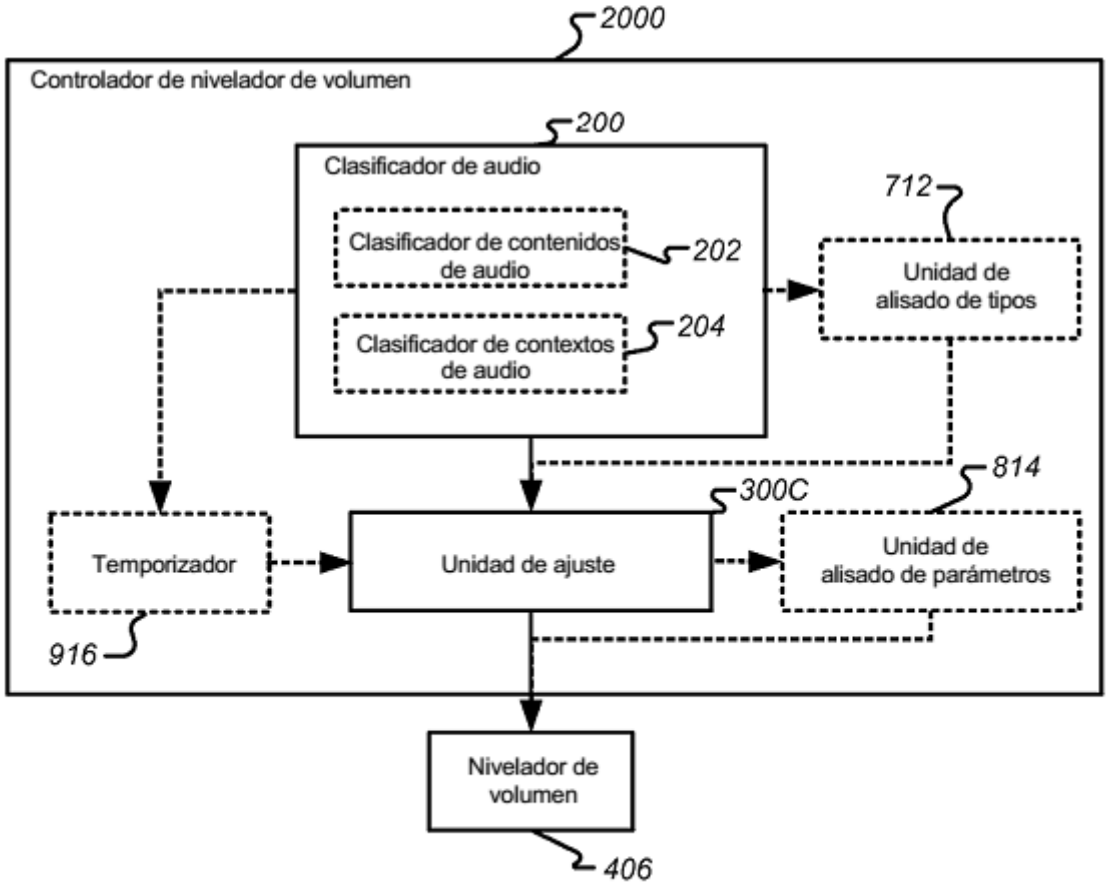
2000
Controlador de nivelador de volumen
200
Clasificador de audio
Clasificador de contenidos de audio
202
Clasificador de contextos de audio
204
712
Unidad de alisado de tipos
814
300C
Temporizador
Unidad de ajuste
Unidad de alisado de parámetros
916
Nivelador de volumen
406

FIG. 20

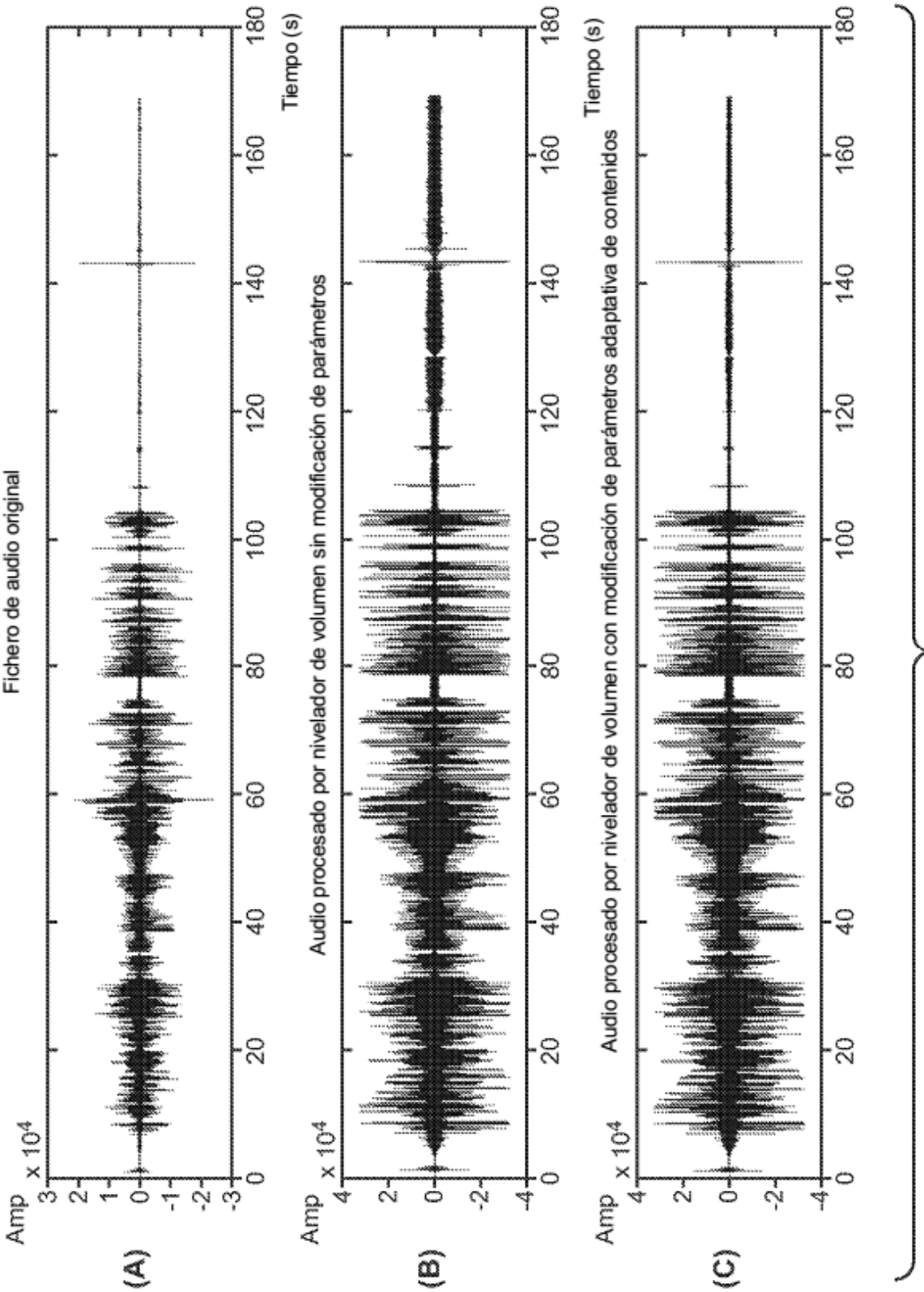
Fichero de audio original
Amp
x 10^4
(A)
Tiempo (s)
Audio procesado por nivelador de volumen sin modificación de parámetros
Amp
x 10^4
(B)
Audio procesado por nivelador de volumen con modificación de parámetros adaptativa de contenidos
Tiempo (s)
Amp
x 10^4
(C)

FIG. 21

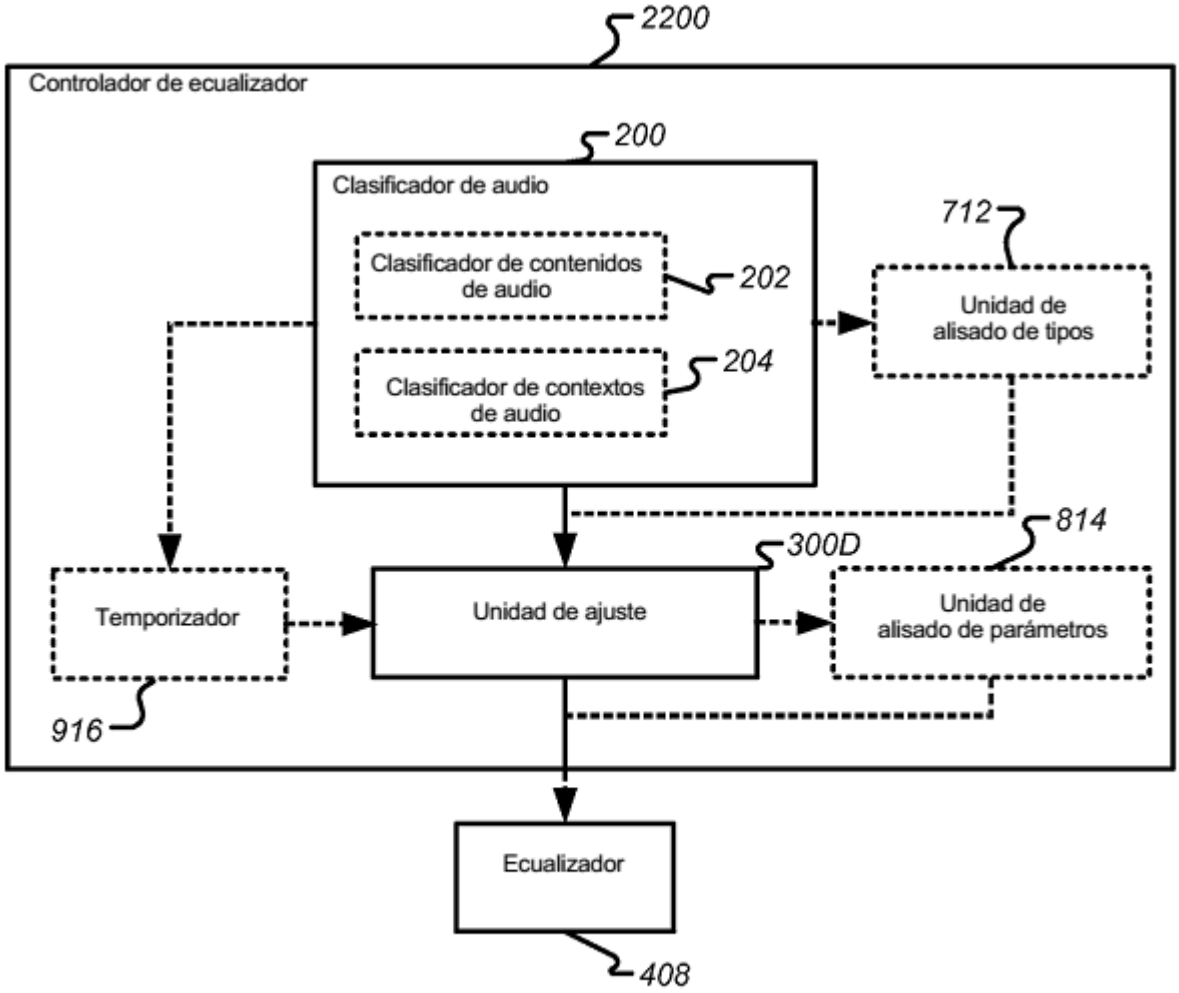
2200
Controlador de ecualizador
200
Clasificador de audio
Clasificador de contenidos de audio
202
Clasificador de contextos de audio
204
712
Unidad de alisado de tipos
814
300D
Temporizador
Unidad de ajuste
Unidad de alisado de parámetros
916
Ecualizador
408

FIG. 22

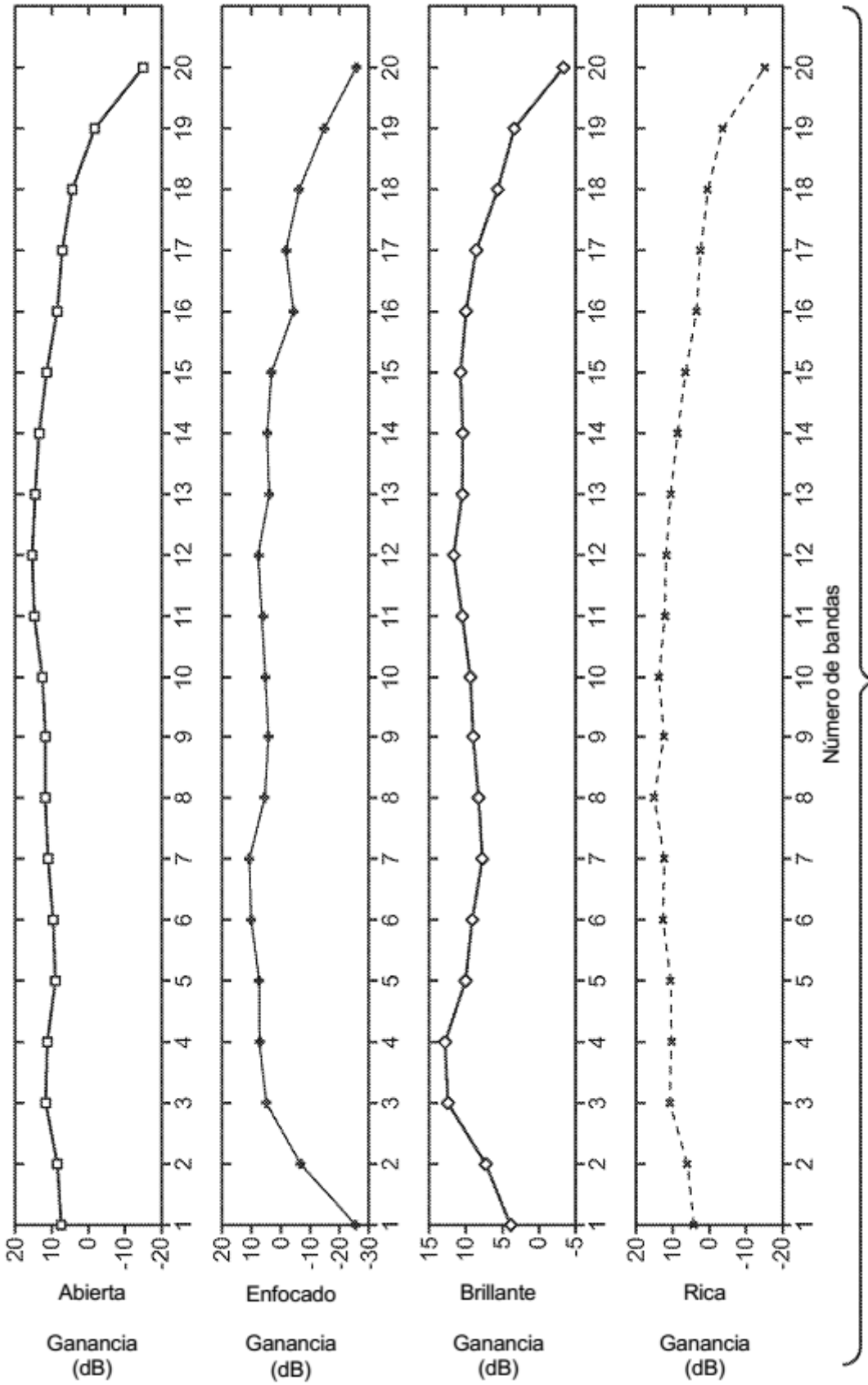
Abierta
Ganancia (dB)
Enfocado
Ganancia (dB)
Brillante
Ganancia (dB)
Rica
Ganancia (dB)
Número de bandas

FIG. 23

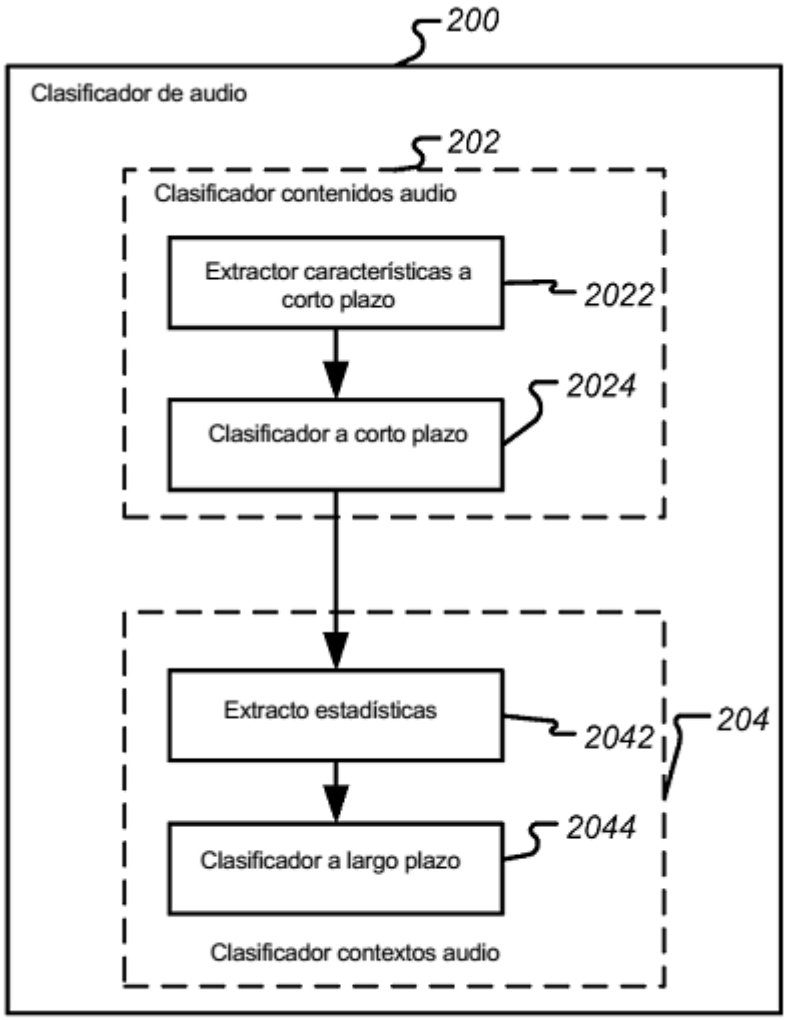
200
Clasificador de audio
202
Clasificador contenidos audio
Extractor características a corto plazo
2022
2024
Clasificador a corto plazo
Extracto estadísticas
2042
204
2044
Clasificador a largo plazo
Clasificador contextos audio

FIG. 24

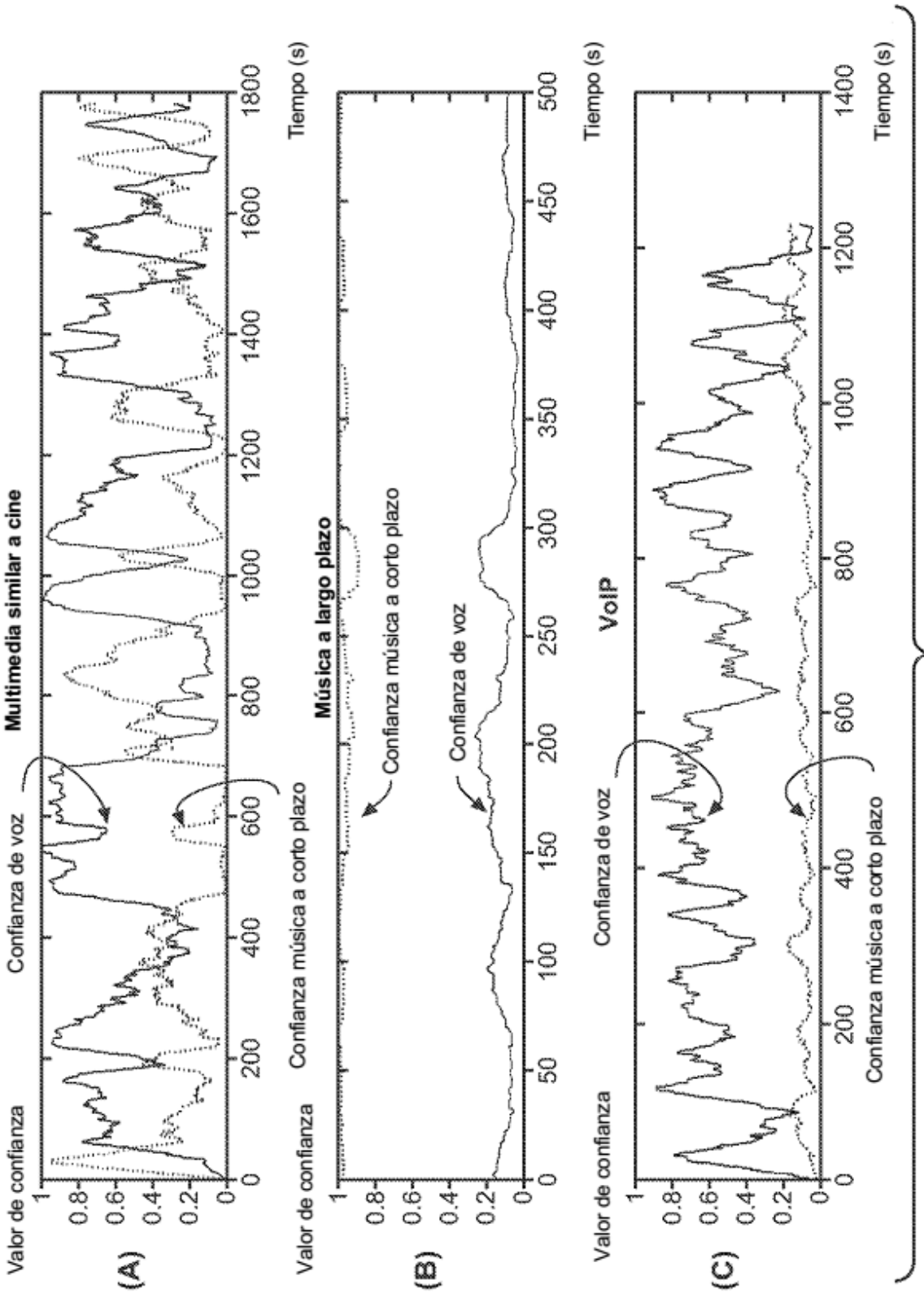
Multimedia similar a cine
Valor de confianza
Confianza de voz
Tiempo (s)
(A)
Música a largo plazo
Confianza música a corto plazo
Confianza de voz
(B)
VoIP
(C)

FIG. 25

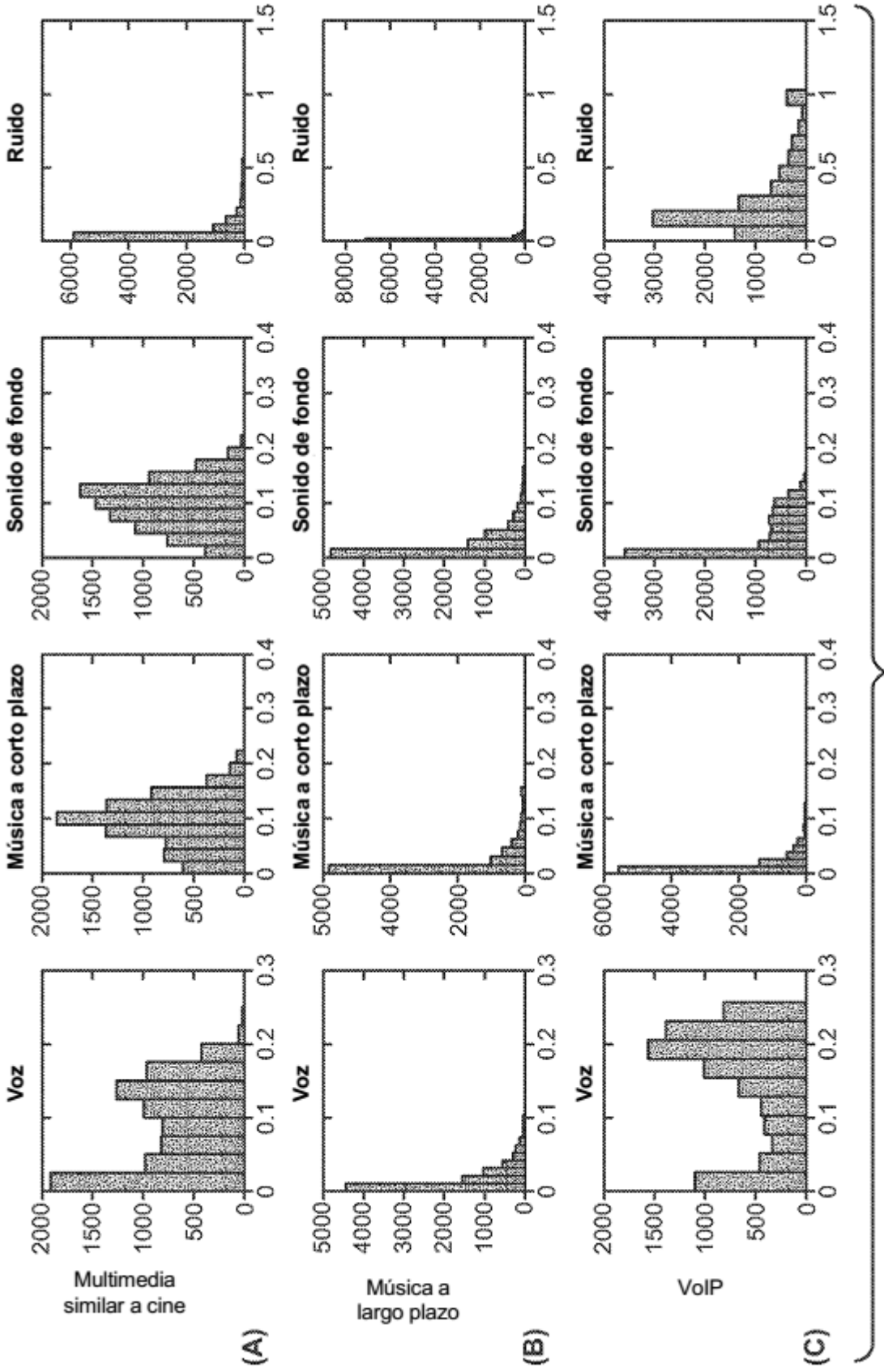
Voz
Música a corto plazo
Sonido de fondo
Ruido
Multimedia similar a cine
(A)
Música a largo plazo
(B)
VoIP
(C)

FIG. 26

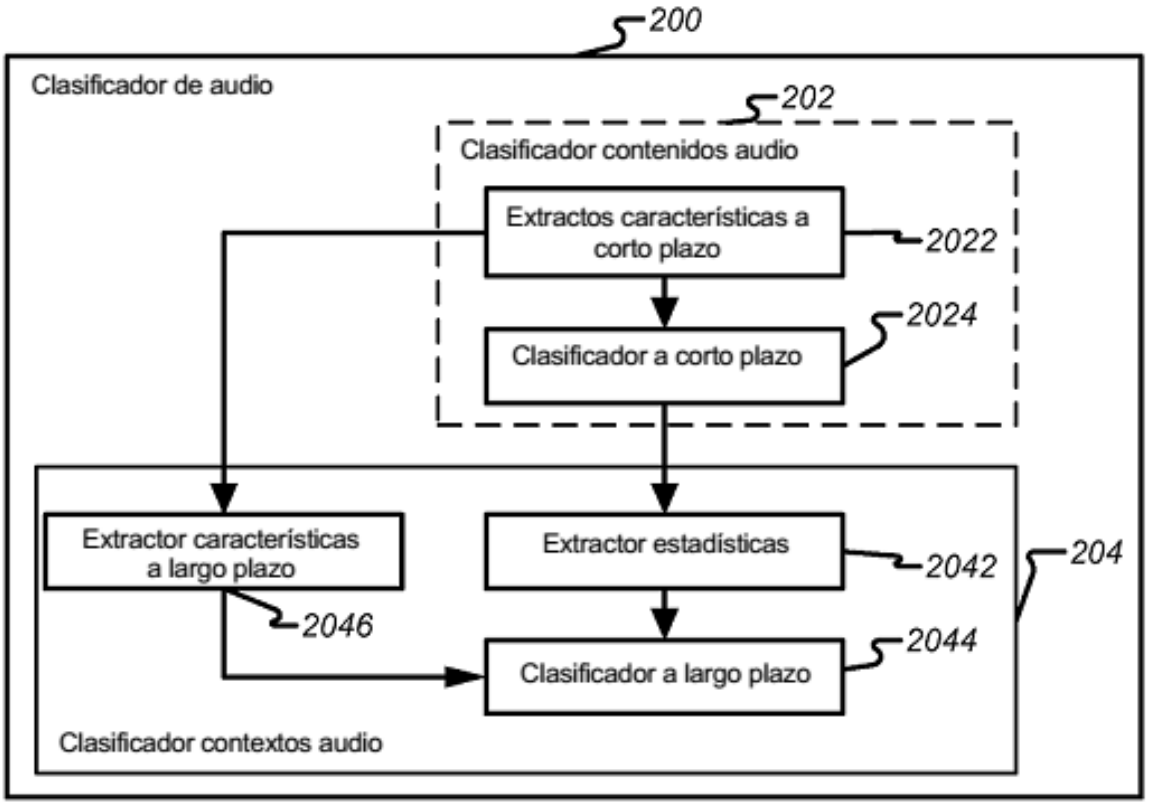
200
Clasificador de audio
202
Clasificador contenidos audio
Extractos características a corto plazo
2022
2024
Clasificador a corto plazo
Extractor características a largo plazo
Extractor estadísticas
2042
204
2046
2044
Clasificador a largo plazo
Clasificador contextos audio

FIG. 27

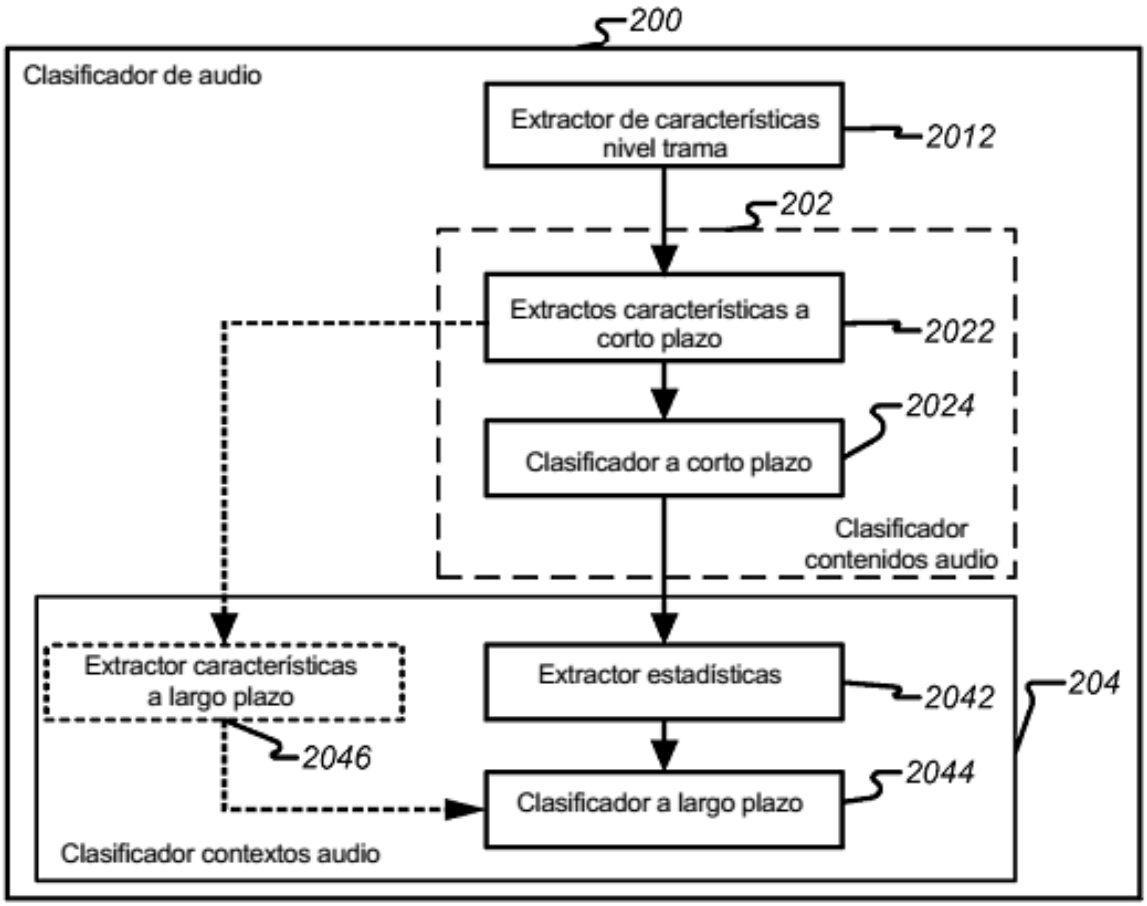
200
Clasificador de audio
Extractor de características nivel trama
2012
202
Extractos características a corto plazo
2022
2024
Clasificador a corto plazo
Clasificador contenidos audio
Extractor características a largo plazo
2046
Extractor estadísticas
2042
204
2044
Clasificador a largo plazo
Clasificador contextos audio

FIG. 28

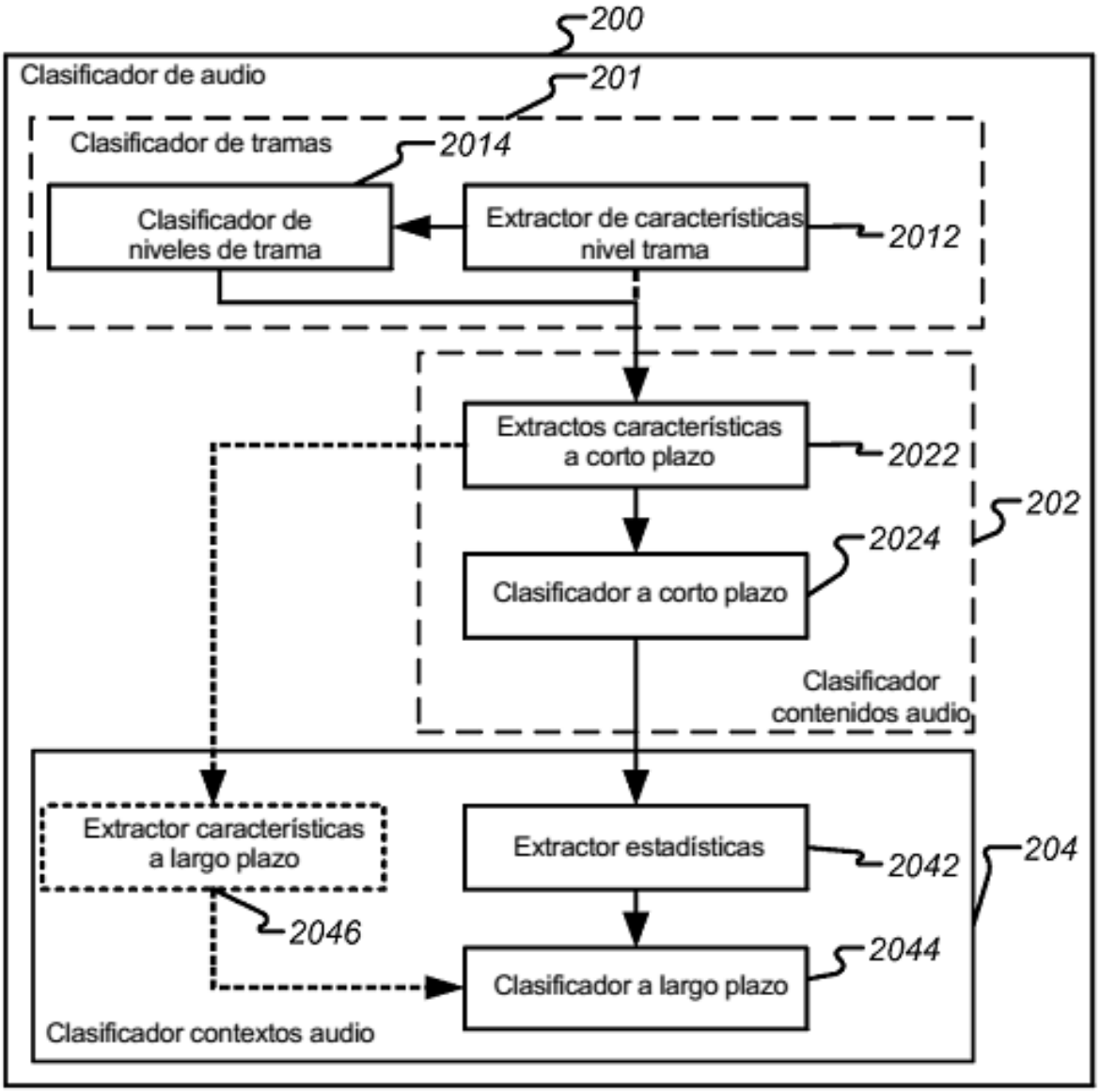
200
Clasificador de audio
201
Clasificador de tramas
2014
Clasificador de niveles de trama
Extractor de características nivel trama
2012
Extractos características a corto plazo
2022
202
2024
Clasificador a corto plazo
Clasificador contenidos audio
Extractor características a largo plazo
Extractor estadísticas
2042
204
2046
2044
Clasificador a largo plazo
Clasificador contextos audio

FIG. 29

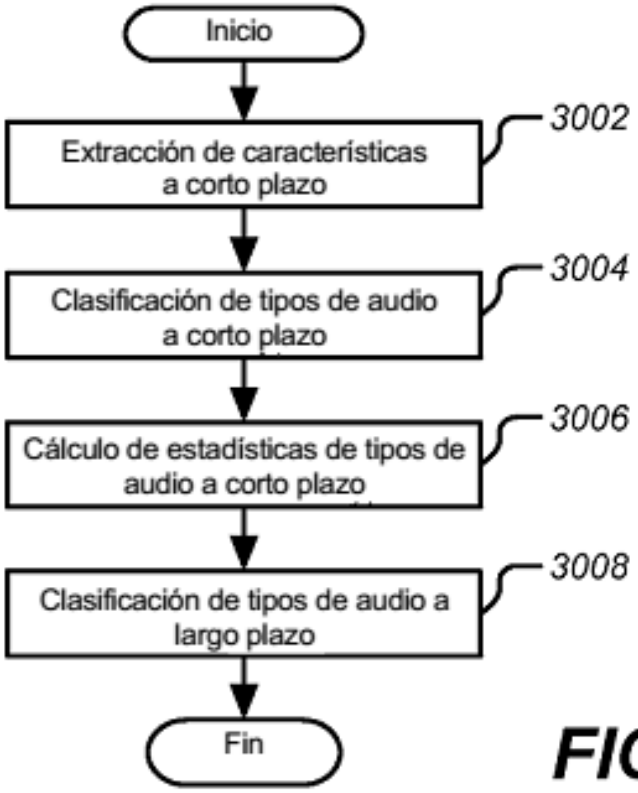
Inicio
Extracción de características a corto plazo
3002
Clasificación de tipos de audio a corto plazo
3004
Cálculo de estadísticas de tipos de audio a corto plazo
3006
Clasificación de tipos de audio a largo plazo
3008
Fin

FIG. 30

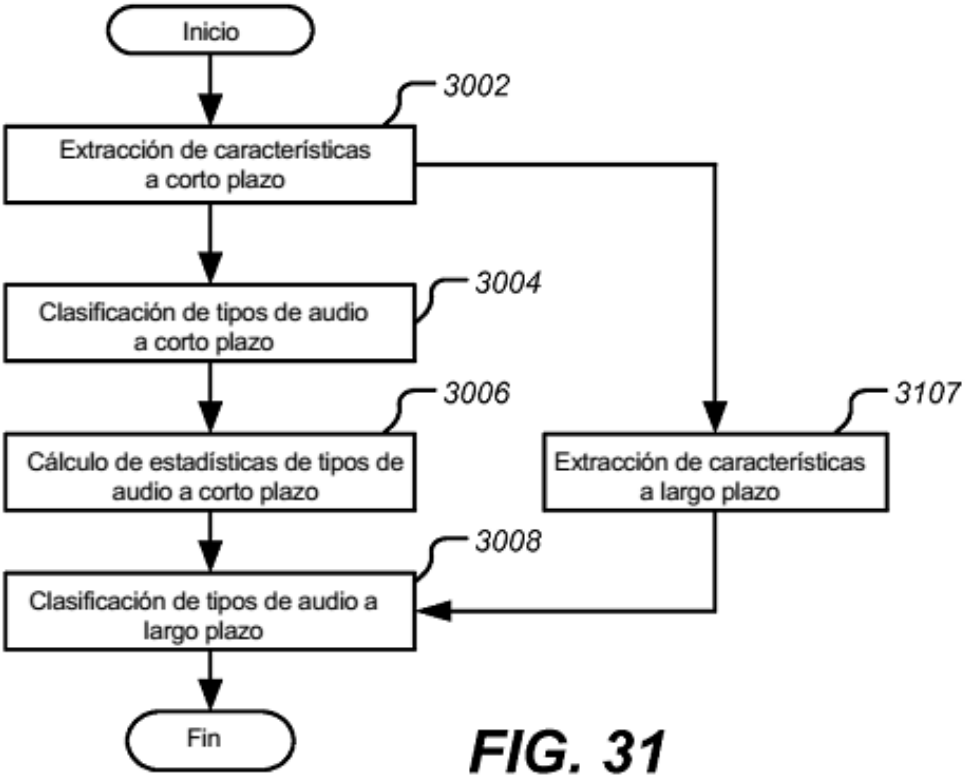
Inicio
Extracción de características a corto plazo
3002
Clasificación de tipos de audio a corto plazo
3004
Cálculo de estadísticas de tipos de audio a corto plazo
3006
Extracción de características a largo plazo
3107
Clasificación de tipos de audio a largo plazo
3008
Fin

FIG. 31

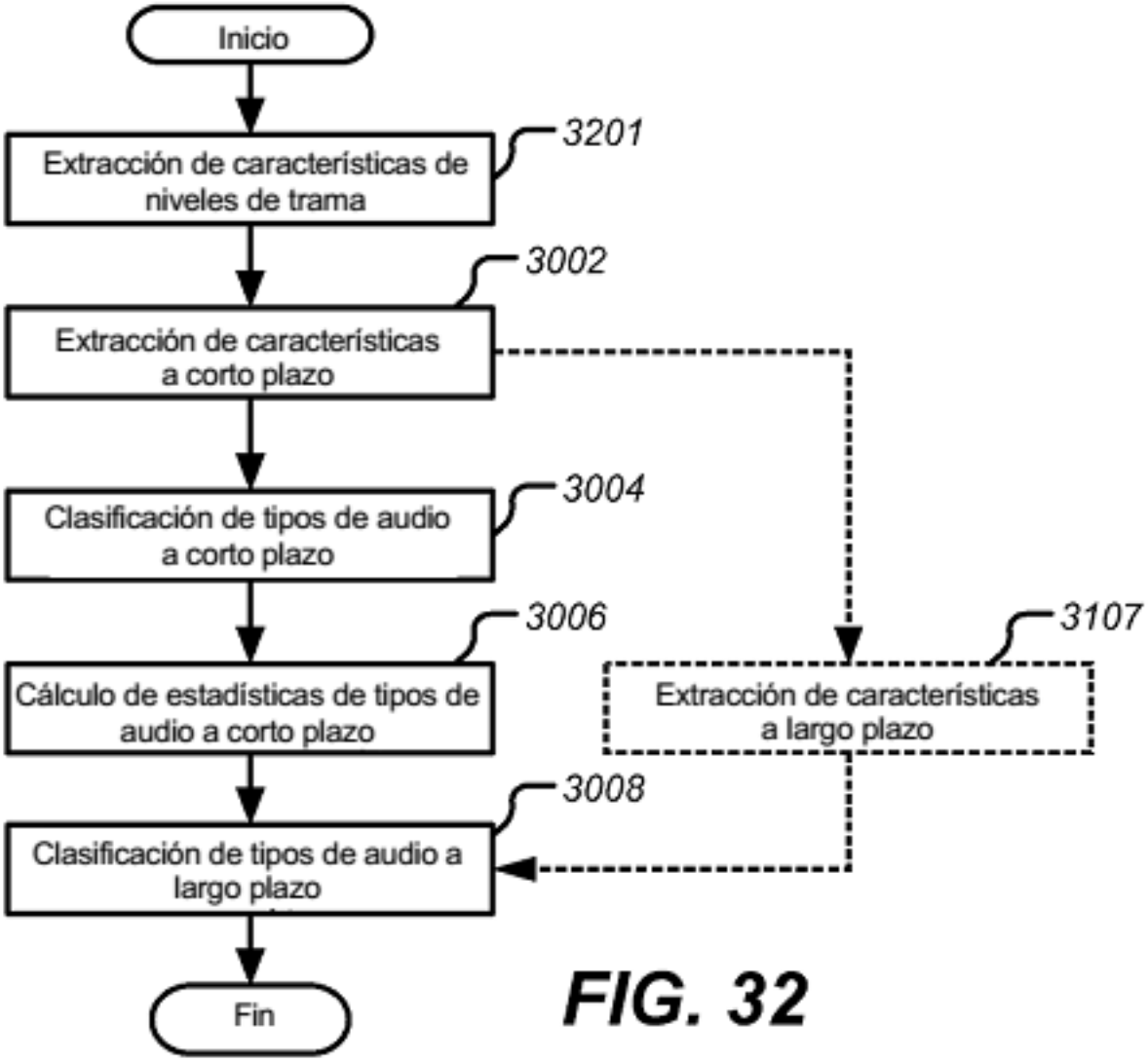
Inicio
Extracción de características de niveles de trama
3201
Extracción de características a corto plazo
3002
Clasificación de tipos de audio a corto plazo
3004
Cálculo de estadísticas de tipos de audio a corto plazo
3006
Extracción de características a largo plazo
3107
Clasificación de tipos de audio a largo plazo
3008
Fin

FIG. 32

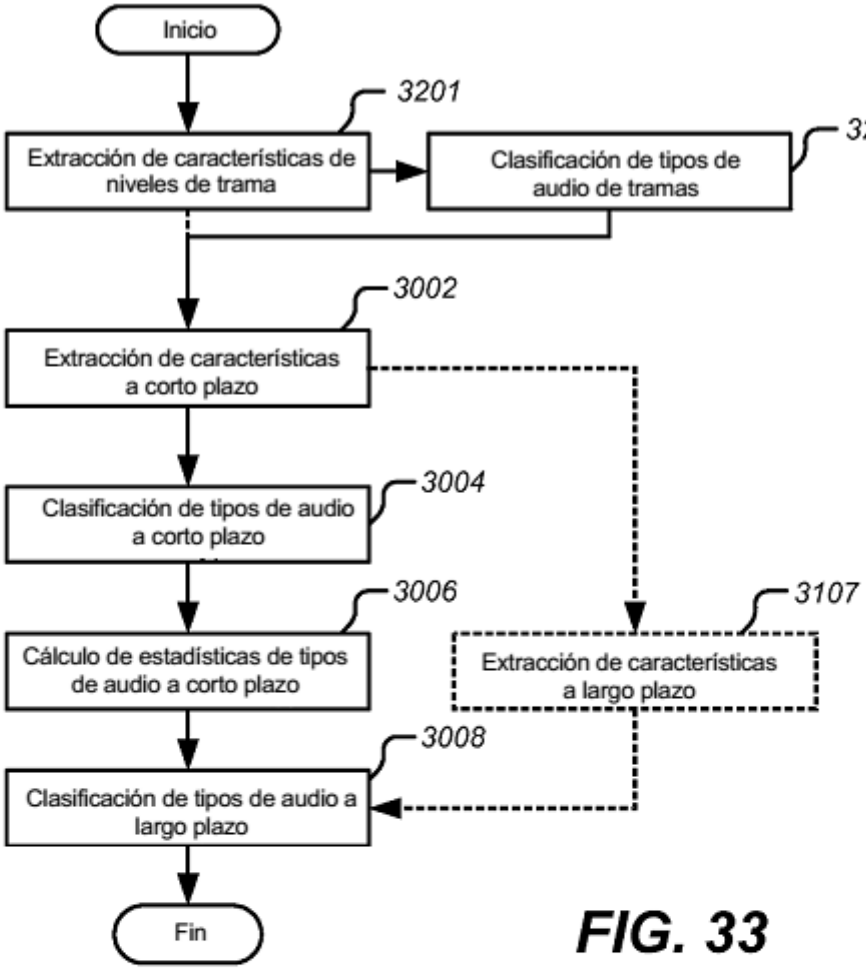
Inicio
3201
Extracción de características de niveles de trama
Clasificación de tipos de audio de tramas
32011
3002
Extracción de características a corto plazo
3004
Clasificación de tipos de audio a corto plazo
3006
Cálculo de estadísticas de tipos de audio a corto plazo
3107
Extracción de características a largo plazo
3008
Clasificación de tipos de audio a largo plazo
Fin

FIG. 33

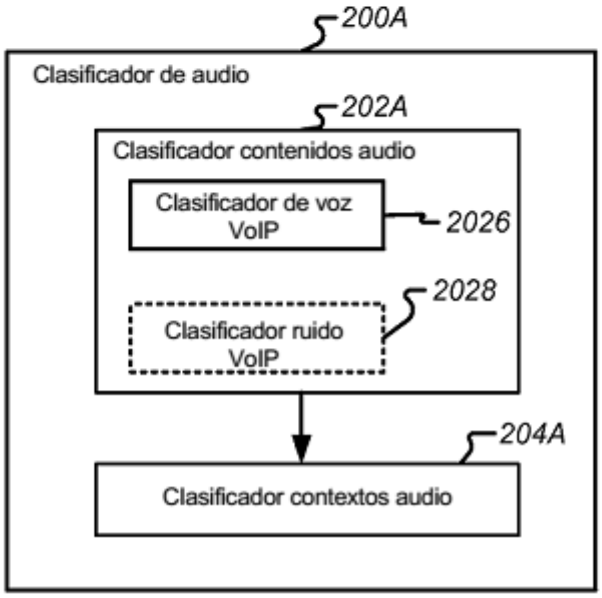
200A
Clasificador de audio
202A
Clasificador contenidos audio
Clasificador de voz VoIP
2026
2028
Clasificador ruido VoIP
204A
Clasificador contextos audio

FIG. 34

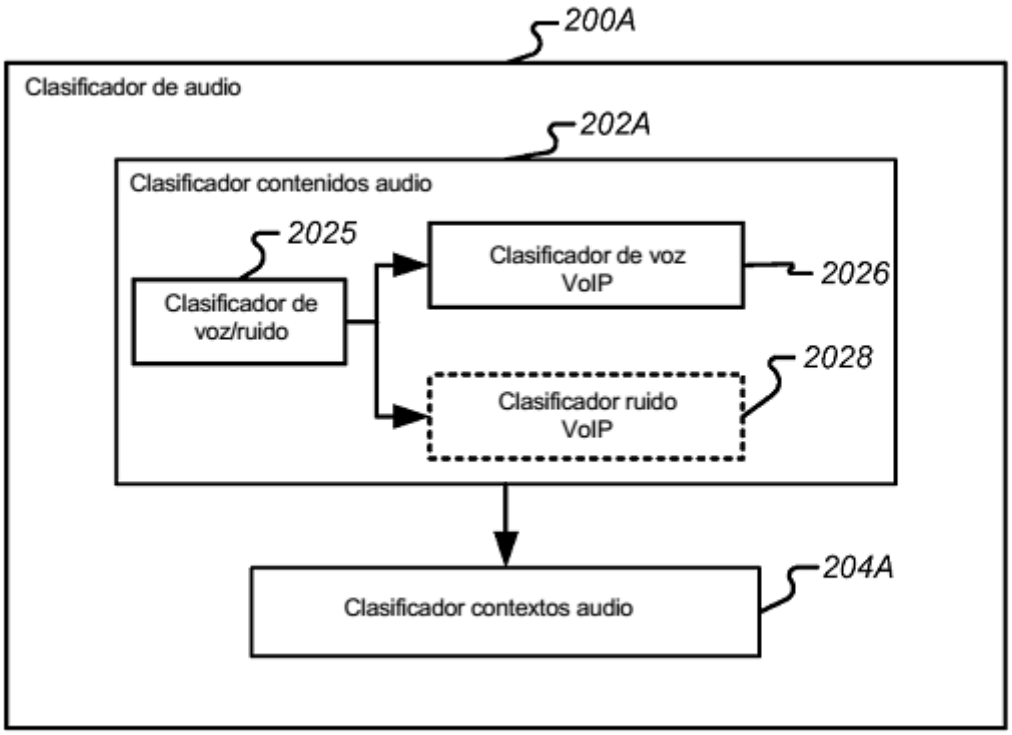
200A
Clasificador de audio
202A
Clasificador contenidos audio
2025
Clasificador de voz/ruido
Clasificador de voz VoIP
2026
2028
Clasificador ruido VoIP
204A
Clasificador contextos audio

FIG. 35

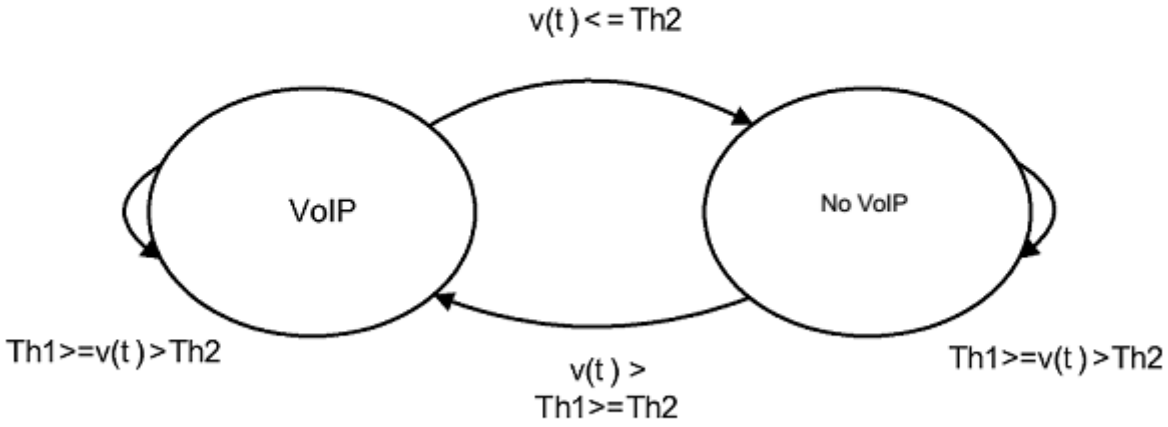
v(t) < = Th2
VoIP
No VoIP
Th1>=v(t)>Th2
v(t) > Th1>=Th2
Th1>=v(t)>Th2

FIG. 36

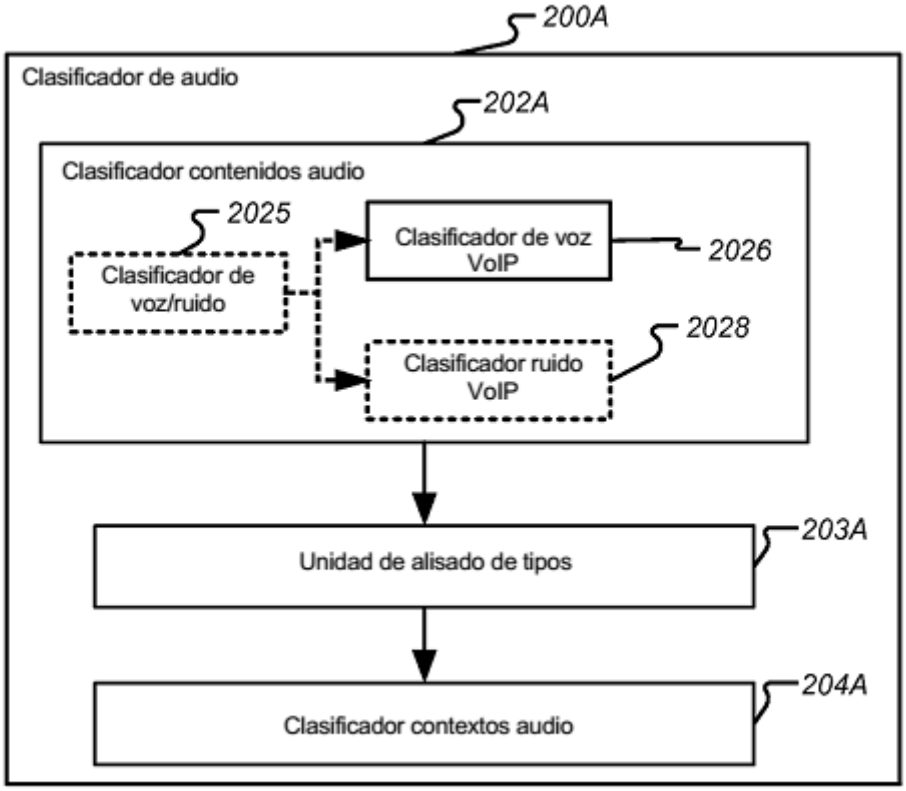
200A
Clasificador de audio
202A
Clasificador contenidos audio
2025
Clasificador de voz/ruido
Clasificador de voz VoIP
2026
2028
Clasificador ruido VoIP
Unidad de alisado de tipos
203A
Clasificador contextos audio
204A

FIG. 37

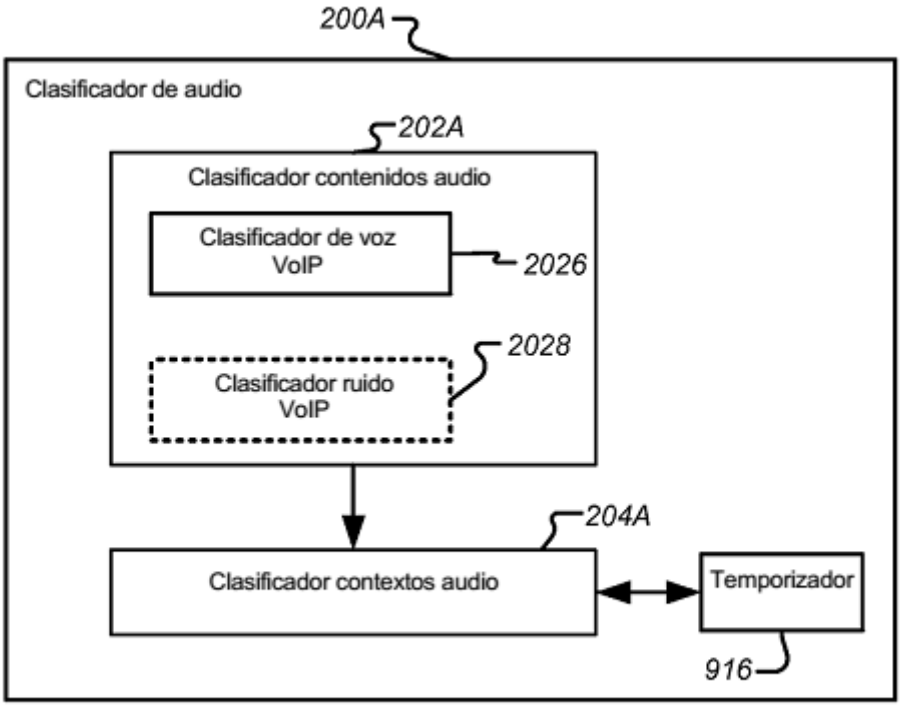
200A
Clasificador de audio
202A
Clasificador contenidos audio
Clasificador de voz VoIP
2026
2028
Clasificador ruido VoIP
204A
Clasificador contextos audio
Temporizador
916

FIG. 38

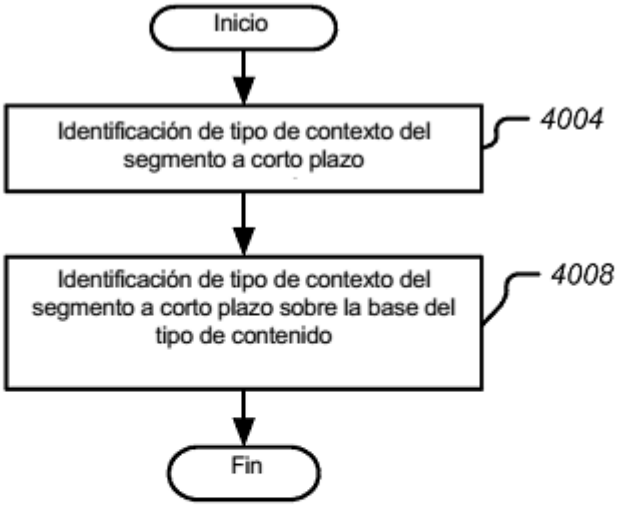
Inicio
Identificación de tipo de contexto del segmento a corto plazo
4004
Identificación de tipo de contexto del segmento a corto plazo sobre la base del tipo de contenido
4008
Fin

FIG. 39

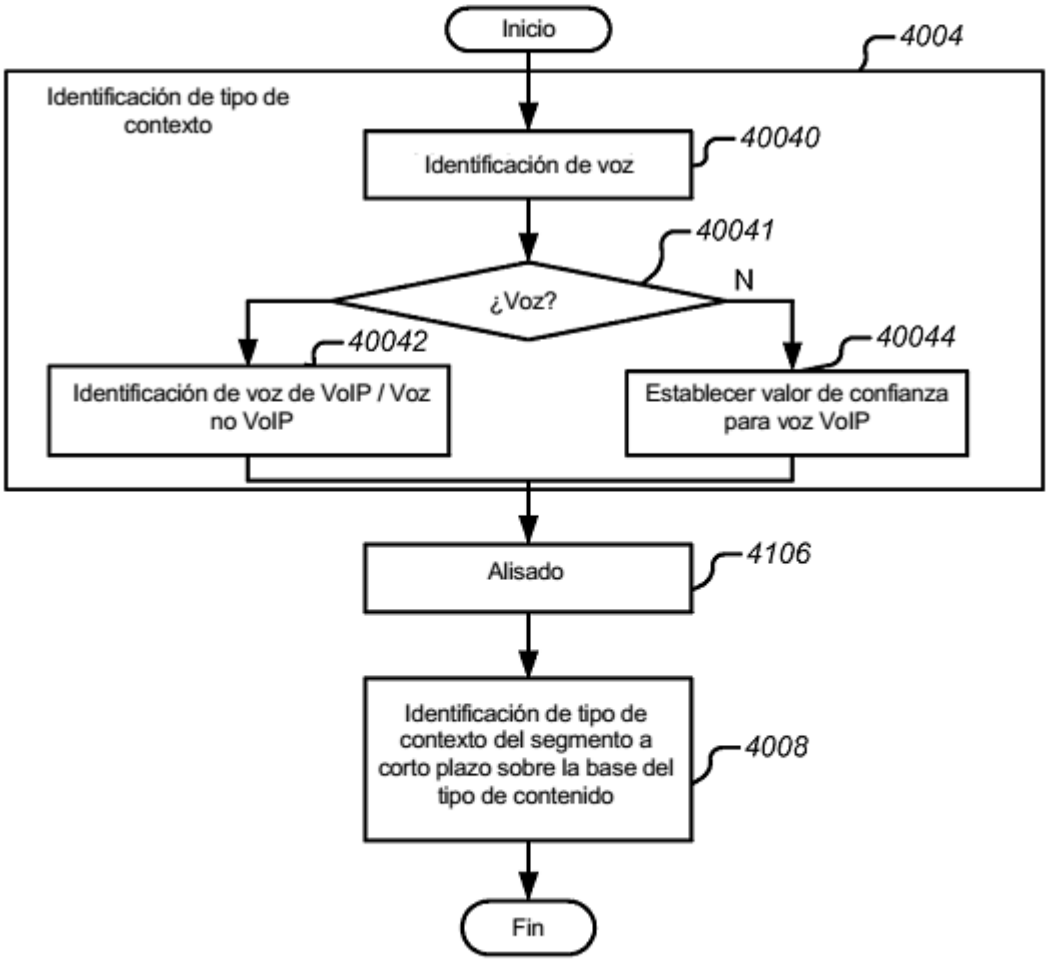
Inicio
4004
Identificación de tipo de contexto
Identificación de voz
40040
40041
¿Voz?
N
40042
Identificación de voz de VoIP / Voz no VoIP
40044
Establecer valor de confianza para voz VoIP
Alisado
4106
Identificación de tipo de contexto del segmento a corto plazo sobre la base del tipo de contenido
4008
Fin

FIG. 40

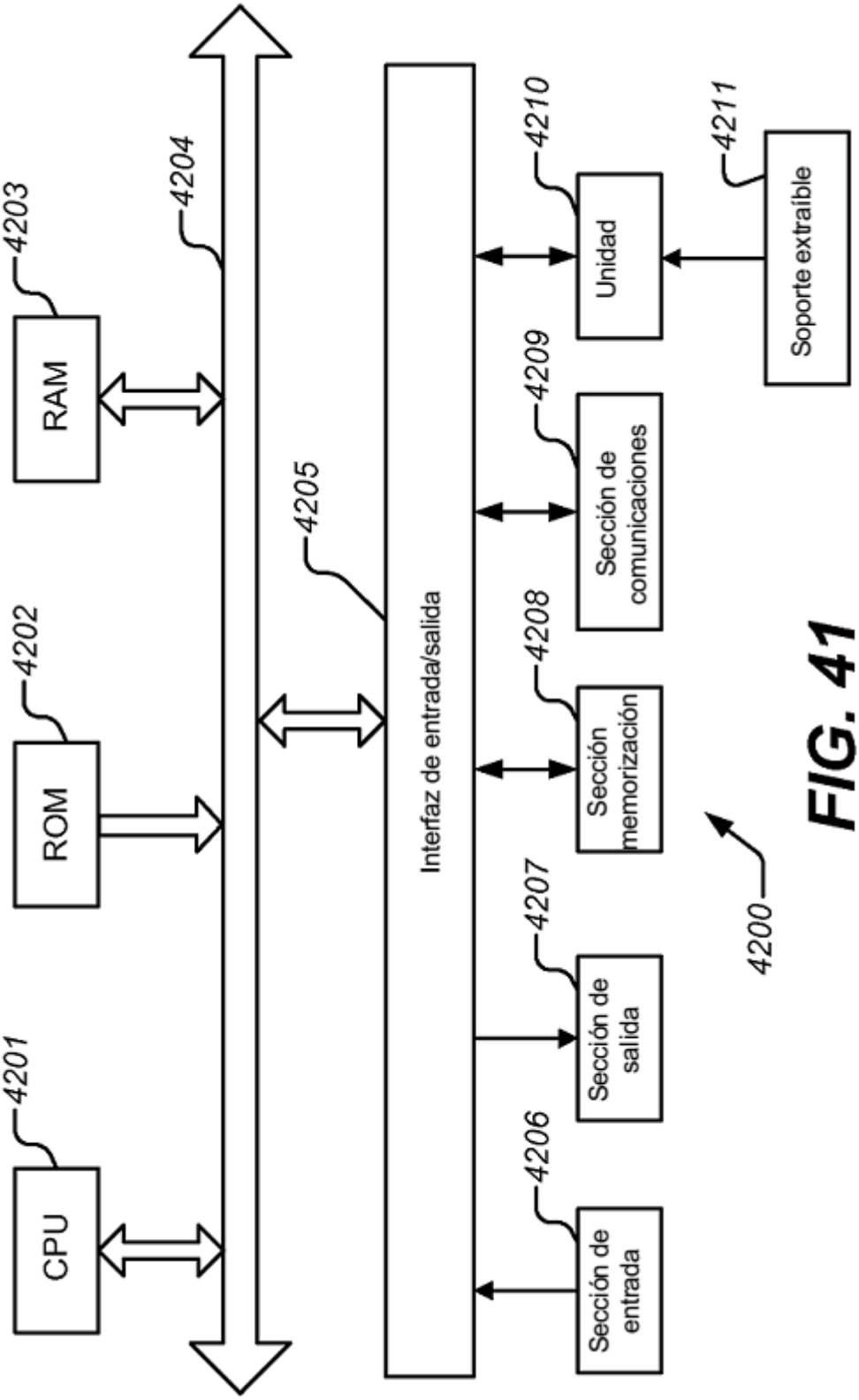
4201
CPU
4202
ROM
4203
RAM
4204
4205
Interfaz de entrada/salida
4206
Sección de entrada
4207
Sección de salida
4208
Sección memorización
4209
Sección de comunicaciones
4210
Unidad
4211
Soporte extraíble
4200

FIG. 41

a