(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 629 172**

(51) Int. CI.:

G10L 25/81 (2013.01)
G10L 25/18 (2013.01)
G10L 25/78 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **26.09.2013 PCT/CN2013/084252**

(87) Fecha y número de publicación internacional: **12.02.2015 WO15018121**

(96) Fecha de presentación y número de la solicitud europea: **26.09.2013 E 13891232 (4)**

(97) Fecha y número de publicación de la concesión europea: **10.05.2017 EP 3029673**

(54) Título: **Procedimiento y dispositivo de clasificación de señales de audio**

(30) Prioridad:

06.08.2013 CN 201310339218

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
07.08.2017

(73) Titular/es:

HUAWEI TECHNOLOGIES CO., LTD. (100.0%)
Huawei Administration Building, Bantian Longgang District , Shenzhen, Guangdong 518129, CN

(72) Inventor/es:

WANG, ZHE

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

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

DESCRIPCIÓN

Procedimiento y dispositivo de clasificación de señales de audio

Campo técnico

La presente invención se refiere al campo de las tecnologías de procesamiento de señales digitales y, en particular, a un procedimiento y aparato de clasificación de señales de audio.

Antecedentes

Para reducir los recursos ocupados por una señal de vídeo durante su almacenamiento o transmisión, una señal de audio se comprime en un extremo de transmisión y después se transmite a un extremo de recepción, y el extremo de recepción restaura la señal de audio mediante descompresión.

En una aplicación de procesamiento de audio, la clasificación de señales de audio es una tecnología importante que se aplica de manera generalizada. Por ejemplo, en una aplicación de codificación/descodificación de audio, un códec relativamente popular es un tipo de híbrido de codificación y descodificación simultánea. Este códec incluye generalmente un codificador (por ejemplo, CELP) basado en un modelo de generación de voz, y un codificador basado en conversión (por ejemplo, un codificador basado en MDCT). A una velocidad binaria intermedia o baja, el codificador basado en un modelo de generación de voz puede obtener una calidad de codificación de voz relativamente buena, pero tiene una calidad de codificación de música relativamente mala, mientras que el codificador basado en conversión puede obtener una calidad de codificación de música relativamente buena pero tiene una calidad de codificación de voz relativamente mala. Por lo tanto, el códec híbrido codifica una señal de voz usando el codificador basado en un modelo de generación de voz y codifica una señal de música usando el codificador basado en conversión, obteniéndose así un efecto de codificación óptimo en general. En el presente documento, la tecnología principal es la clasificación de señales de audio, o la selección del modo de codificación, en lo que respecta específicamente a esta solicitud.

El códec híbrido necesita obtener información precisa del tipo de señal antes de que el códec híbrido pueda obtener una selección óptima de un modo de codificación. En el presente documento, un clasificador de señales de audio puede considerarse también, en términos generales, como un clasificador de voz/música. La velocidad de reconocimiento de voz y la velocidad de reconocimiento de música son indicadores importantes para medir el rendimiento del clasificador de voz/música. En lo que se refiere en particular a una señal de música, debido a la diversidad/complejidad de sus características de señal, el reconocimiento de la señal de música es generalmente más difícil que el de una señal de voz. Además, el retardo de reconocimiento es también un indicador muy importante. Debido a la imprecisión de las características de la voz/música en un breve espacio de tiempo, generalmente se necesita un espacio de tiempo relativamente largo antes de que la voz/música pueda reconocerse de manera relativamente precisa. Generalmente, en una sección intermedia de un mismo tipo de señales, un retardo de reconocimiento más largo indica un reconocimiento más preciso. Sin embargo, en una sección de transición de dos tipos de señales, un retardo de reconocimiento más largo indica una menor precisión del reconocimiento, lo que se acentúa especialmente en una situación en la que se introduce una señal híbrida (por ejemplo, voz que tiene música de fondo). Por lo tanto, el que haya una alta velocidad de reconocimiento y un bajo retardo de reconocimiento es un atributo necesario de un reconocedor de voz/música de alto rendimiento. Además, la estabilidad de la clasificación es también un atributo importante que afecta a la calidad de la codificación de un codificador híbrido. Generalmente, cuando el codificador híbrido conmuta entre diferentes tipos de codificadores, la calidad puede deteriorarse. Si en un clasificador se produce una conmutación de tipos frecuente en un mismo tipo de señales, la calidad de la codificación se ve afectada de manera relativamente importante; por lo tanto, es necesario que el resultado de clasificación de salida del clasificador sea preciso y uniforme. Además, en algunas aplicaciones, tal como un algoritmo de clasificación en un sistema de comunicaciones, también se requiere que la complejidad del cálculo y las sobrecargas de almacenamiento del algoritmo de clasificación sean lo más bajas posible para satisfacer los requisitos comerciales.

La norma G.720.1 de la ITU-T incluye un clasificador de voz/música. Este clasificador usa un parámetro principal: una varianza de fluctuación de espectro de frecuencia, var_flux, como base principal para la clasificación de señales, y usa dos parámetros de máximos de espectro de diferente frecuencia, p1 y p2, como base auxiliar. La clasificación de una señal de entrada según var_flux se completa en una memoria intermedia FIFO de var_flux según estadísticas locales de var_flux. A continuación se resume un proceso específico: Primero, una fluctuación de espectro de frecuencia, flux, se extrae de cada trama de audio de entrada y se almacena en una primera memoria intermedia, y, en este caso, la fluctuación flux se calcula en las cuatro últimas tramas, que incluyen una trama de entrada actual, o puede calcularse usando otro procedimiento. Después se calcula una varianza de flux de las N últimas tramas, que incluyen la trama de entrada actual, para obtener var_flux de la trama de entrada actual, y var_flux se almacena en una segunda memoria intermedia. Después se cuenta una cantidad K de tramas cuya var_flux es mayor que un primer umbral entre las M últimas tramas, que incluyen la trama de entrada actual, de la segunda memoria intermedia. Si una relación de K respecto a M es mayor que un segundo umbral, se determina que la trama de entrada actual es una trama de voz; en caso contrario, la trama de entrada actual es una trama de

música. Los parámetros auxiliares p1 y p2 se utilizan principalmente para modificar la clasificación y también se calculan para cada trama de audio de entrada. Cuando p1 y/o p2 es mayor que un tercer umbral y/o un cuarto umbral, se determina directamente que la trama de audio de entrada actual es una trama de música.

Las desventajas de este clasificador de voz/música son las siguientes: por un lado, sigue siendo necesario mejorar la velocidad absoluta de reconocimiento de música y, por otro lado, puesto que las aplicaciones objetivo del clasificador no son específicas de un escenario de aplicación de una señal híbrida, sigue siendo necesario mejorar el rendimiento de reconocimiento de una señal híbrida.

Muchos clasificadores de voz/música existentes están diseñados en función de un principio de reconocimiento de modo. Este tipo de clasificador extrae generalmente múltiples (desde una docena a varias docenas de) parámetros de características a partir de una trama de audio de entrada e introduce estos parámetros en un clasificador basándose en un modelo híbrido gaussiano, o una red neuronal u otro procedimiento de clasificación clásico para llevar a cabo la clasificación.

Este tipo de clasificadores tiene una base teórica relativamente sólida, pero generalmente tienen una complejidad de cálculo o almacenamiento relativamente alta y, por lo tanto, los costes de implementación son relativamente altos.

El documento EP 2 339 575 A1 describe un procedimiento y dispositivo de clasificación de señales.

El documento de EDITOR G GSAD, "*Draft new ITU-T Recommendation G.720.1 (ex G.GSAD) Generic sound activity detector (for Consent)*", 3GPP DRAFT; COM16-LS121-ATT.1-TD-PLEN-0186, PROYECTO DE ASOCIACIÓN DE TERCERA GENERACIÓN (3GPP), CENTRO DE COMPETENCIAS MÓVILES; 650, ROUTE DES LUCIOLES ; F-06921 SOPHIA-ANTIPOLIS CEDEX ; FRANCIA, (20091107), XP050638609, describe un detector de actividad de sonido genérico.

El documento CN 102 446 504 A describe un procedimiento y un equipo de identificación de voz/música.

Resumen

Un objetivo de las formas de realización de la presente invención es proporcionar un procedimiento y aparato de clasificación de señales de audio para reducir la complejidad en la clasificación de señales, a la vez que se garantiza la velocidad de reconocimiento de clasificación de una señal de audio híbrida.

Según un primer aspecto, se proporciona un procedimiento de clasificación de señales de audio, donde el procedimiento incluye:

determinar, según una actividad de voz de una trama de audio actual, si hay que obtener una fluctuación de espectro de frecuencia de la trama de audio actual y almacenar la fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia, donde la fluctuación de espectro de frecuencia denota una fluctuación de energía de un espectro de frecuencia de una señal de audio;
actualizar, dependiendo de si la trama de audio es música percutante, las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia; y
clasificar la trama de audio actual como una trama de voz o una trama de música según las estadísticas de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia.

En una primera manera de implementación posible, determinar, según la actividad de voz de una trama de audio actual, si hay que obtener una fluctuación de espectro de frecuencia de la trama de audio actual y almacenar la fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia incluye:

si la trama de audio actual es una trama activa, almacenar la fluctuación de espectro de frecuencia de la trama de audio actual en la memoria de fluctuaciones de espectro de frecuencia.

En una segunda manera de implementación posible, determinar, según la actividad de voz de una trama de audio actual, si hay que obtener una fluctuación de espectro de frecuencia de la trama de audio actual y almacenar la fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia incluye:

si la trama de audio actual es una trama activa y la trama de audio actual no pertenece a un ataque de energía, almacenar la fluctuación de espectro de frecuencia de la trama de audio actual en la memoria de fluctuaciones de espectro de frecuencia.

En una tercera manera de implementación posible, determinar, según la actividad de voz de una trama de audio actual, si hay que obtener una fluctuación de espectro de frecuencia de la trama de audio actual y almacenar la fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia incluye:

si la trama de audio actual es una trama activa y ninguna de múltiples tramas consecutivas, que incluyen la trama de audio actual y una trama histórica de la trama de audio actual, pertenece a un ataque de energía, almacenar la fluctuación de espectro de frecuencia de la trama de audio actual en la memoria de fluctuaciones de espectro de frecuencia.

Con referencia al primer aspecto o la primera manera de implementación posible del primer aspecto o la segunda manera de implementación posible del primer aspecto o la tercera manera de implementación posible del primer aspecto, en una cuarta manera de implementación posible, actualizar, dependiendo de si la trama de audio actual es música percutante, las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia incluye:

si la trama de audio actual pertenece a música percutante, modificar los valores de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia.

Con referencia al primer aspecto o la primera manera de implementación posible del primer aspecto o la segunda manera de implementación posible del primer aspecto o la tercera manera de implementación posible del primer aspecto o la cuarta manera de implementación posible del primer aspecto, en una quinta manera de implementación posible, clasificar la trama de audio actual como una trama de voz o una trama de música según las estadísticas de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia incluye:

obtener un valor promedio de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia; y
cuando el valor promedio obtenido de los datos eficaces de las fluctuaciones de espectro de frecuencia satisface una condición de clasificación de música, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz.

Con referencia al primer aspecto o la primera manera de implementación posible del primer aspecto o la segunda manera de implementación posible del primer aspecto o la tercera manera de implementación posible del primer aspecto o la cuarta manera de implementación posible del primer aspecto, en una sexta manera de implementación posible, el procedimiento de clasificación de señales de audio incluye además:

obtener un máximo de banda de alta frecuencia de espectro de frecuencia, un grado de correlación de espectro de frecuencia y una pendiente de energía residual de predicción lineal de la trama de audio actual, donde el máximo de banda de alta frecuencia de espectro de frecuencia denota un máximo o un pico de energía, en una banda de alta frecuencia, de un espectro de frecuencia de la trama de audio actual; el grado de correlación de espectro de frecuencia denota estabilidad, entre tramas adyacentes, de una estructura armónica de señal de la trama de audio actual, y la pendiente de energía residual de predicción lineal denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio a medida que aumenta el orden de predicción lineal; y
determinar, según la actividad de voz de la trama de audio actual, si hay que almacenar en memorias el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal,
donde clasificar la trama de audio según las estadísticas de una parte de o todos los datos de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia incluye:
obtener por separado un valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, un valor promedio de datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, un valor promedio de datos eficaces de grados de correlación de espectro de frecuencia almacenados y una varianza de datos eficaces de pendientes de energía residual de predicción lineal almacenadas; y
cuando se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz: el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

Según un segundo aspecto, se proporciona un aparato de clasificación de señales de audio, donde el aparato está configurado para clasificar una señal de audio de entrada, e incluye:

una unidad de determinación de almacenamiento, configurada para determinar, según la actividad de voz de la trama de audio actual, si hay que obtener y almacenar una fluctuación de espectro de frecuencia de la trama de audio actual, donde la fluctuación de espectro de frecuencia denota una fluctuación de energía de un espectro de frecuencia de una señal de audio;

una memoria, configurada para almacenar la fluctuación de espectro de frecuencia cuando la unidad de determinación de almacenamiento proporciona un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia;
una unidad de actualización, configurada para actualizar, dependiendo de si la trama de audio es música percutante, las fluctuaciones de espectro de frecuencia almacenadas en la memoria; y
una unidad de clasificación, configurada para clasificar la trama de audio actual como una trama de voz o una trama de música según las estadísticas de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria.

En una primera manera de implementación posible, la unidad de determinación de almacenamiento está configurada específicamente para: cuando se determina que la trama de audio actual es una trama activa, proporcionar un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia de la trama de audio actual.

En una segunda manera de implementación posible, la unidad de determinación de almacenamiento está configurada específicamente para: cuando se determina que la trama de audio actual es una trama activa y la trama de audio actual no pertenece a un ataque de energía, proporcionar un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia de la trama de audio actual.

En una tercera manera de implementación posible, la unidad de determinación de almacenamiento está configurada específicamente para: cuando se determina que la trama de audio actual es una trama activa y ninguna de múltiples tramas consecutivas, que incluyen la trama de audio actual y una trama histórica de la trama de audio actual, pertenece a un ataque de energía, proporcionar un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia de la trama de audio actual.

Con referencia al segundo aspecto o la primera manera de implementación posible del segundo aspecto o la segunda manera de implementación posible del segundo aspecto o la tercera manera de implementación posible del segundo aspecto, en una cuarta manera de implementación posible, la unidad de actualización está configurada específicamente para: si la trama de audio actual pertenece a música percutante, modificar valores de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia.

Con referencia al segundo aspecto o la primera manera de implementación posible del segundo aspecto o la segunda manera de implementación posible del segundo aspecto o la tercera manera de implementación posible del segundo aspecto o la cuarta manera de implementación posible del segundo aspecto, en una quinta manera de implementación posible, la unidad de clasificación incluye:

una unidad de cálculo, configurada para obtener un valor promedio de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria; y
una unidad de determinación, configurada para comparar el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia con una condición de clasificación de música; y cuando el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia satisface la condición de clasificación de música, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz.

Con referencia al segundo aspecto o la primera manera de implementación posible del segundo aspecto o la segunda manera de implementación posible del segundo aspecto o la tercera manera de implementación posible del segundo aspecto o la cuarta manera de implementación posible del segundo aspecto, en una sexta manera de implementación posible, el aparato de clasificación de señales de audio incluye además:

una unidad de obtención de parámetros, configurada para obtener un máximo de banda de alta frecuencia de espectro de frecuencia, un grado de correlación de espectro de frecuencia, un parámetro de sonoridad y una pendiente de energía residual de predicción lineal de la trama de audio actual, donde el máximo de banda de alta frecuencia de espectro de frecuencia denota un máximo o un pico de energía, en una banda de alta frecuencia, de un espectro de frecuencia de la trama de audio actual; el grado de correlación de espectro de frecuencia denota estabilidad, entre tramas adyacentes, de una estructura armónica de señal de la trama de audio actual; el parámetro de sonoridad denota un grado de correlación de dominio de tiempo entre la trama de audio actual y una señal antes de un periodo de tono; y la pendiente de energía residual de predicción lineal denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio a medida que aumenta un orden de predicción lineal; donde
la unidad de determinación de almacenamiento está configurada además para determinar, según la actividad de voz de la trama de audio actual, si hay que almacenar en memorias el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal;
la memoria está configurada además para: cuando la unidad de determinación de almacenamiento proporciona un resultado que indica que es necesario almacenar el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual

de predicción lineal, almacenar el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal; y
la unidad de clasificación está configurada específicamente para obtener estadísticas de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, estadísticas de datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, estadísticas de datos eficaces de grados de correlación de espectro de frecuencia almacenados, y estadísticas de datos eficaces de pendientes de energía residual de predicción lineal almacenadas, y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de los datos eficaces.

Con referencia a la sexta manera de implementación posible del segundo aspecto, en una séptima manera de implementación posible, la unidad de clasificación incluye:

una unidad de cálculo, configurada para obtener por separado un valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, un valor promedio de los datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, un valor promedio de los datos eficaces de grados de correlación de espectro de frecuencia almacenados y una varianza de los datos eficaces de pendientes de energía residual de predicción lineal almacenadas; y
una unidad de determinación, configurada para: cuando se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz: el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

En las formas de realización de la presente invención, una señal de audio se clasifica según estadísticas a largo plazo de fluctuaciones de espectro de frecuencia; por lo tanto, hay un número relativamente bajo de parámetros, la velocidad de reconocimiento es relativamente alta y la complejidad es relativamente baja. Además, las fluctuaciones de espectro de frecuencia se ajustan teniendo en cuenta factores tales como la actividad de voz y la música percutante; por lo tanto, la presente invención tiene una mayor velocidad de reconocimiento para una señal de música, y es adecuada para la clasificación de señales de audio híbridas.

Breve descripción de los dibujos

Para describir con mayor claridad las soluciones técnicas de las formas de realización de la presente invención o de la técnica anterior, a continuación se introducen brevemente los dibujos adjuntos requeridos para describir las formas de realización o la técnica anterior. Evidentemente, los dibujos adjuntos de la siguiente descripción muestran simplemente algunas formas de realización de la presente invención, y los expertos en la técnica pueden obtener otros dibujos a partir de estos dibujos adjuntos sin realizar investigaciones adicionales.

La FIG. 1 es un diagrama esquemático para dividir una señal de audio en tramas.
La FIG. 2 es un diagrama de flujo esquemático de una forma de realización de un procedimiento de clasificación de señales de audio según la presente invención.
La FIG. 3 es un diagrama de flujo esquemático de una forma de realización para obtener una fluctuación de espectro de frecuencia.
La FIG. 4 es un diagrama de flujo esquemático de otra forma de realización de un procedimiento de clasificación de señales de audio.
La FIG. 5 es un diagrama de flujo esquemático de otra forma de realización de un procedimiento de clasificación de señales de audio.
La FIG. 6 es un diagrama de flujo esquemático de otra forma de realización de un procedimiento de clasificación de señales de audio.
Las FIG. 7 a 10 son diagramas de flujo específicos de clasificación de señales de audio.
La FIG. 11 es un diagrama de flujo esquemático de otra forma de realización de un procedimiento de clasificación de señales de audio.
La FIG. 12 es un diagrama de flujo específico de clasificación de señales de audio.
La FIG. 13 es un diagrama estructural esquemático de una forma de realización de un aparato de clasificación de señales de audio según la presente invención.
La FIG. 14 es un diagrama estructural esquemático de una forma de realización de una unidad de clasificación.
La FIG. 15 es un diagrama estructural esquemático de otra forma de realización de un aparato de clasificación de señales de audio.
La FIG. 16 es un diagrama estructural esquemático de otra forma de realización de un aparato de clasificación de señales de audio.
La FIG. 17 es un diagrama estructural esquemático de una forma de realización de una unidad de clasificación.

La FIG. 18 es un diagrama estructural esquemático de otra forma de realización de un aparato de clasificación de señales de audio.
La FIG. 19 es un diagrama estructural esquemático de otra forma de realización de un aparato de clasificación de señales de audio.

Descripción de formas de realización

A continuación se describe de manera clara y completa las soluciones técnicas de las formas de realización, donde no todas ellas pertenecen a la invención, con referencia a los dibujos adjuntos.

En el campo del procesamiento de señales digitales, los códecs de audio y los códecs de vídeo se aplican de manera generalizada en varios dispositivos electrónicos, por ejemplo un teléfono móvil, un aparato inalámbrico, un asistente digital personal (PDA), un ordenador manual o portátil, un receptor/navegador GPS, una cámara, un reproductor de audio/vídeo, una cámara de vídeo, una grabadora de vídeo y un dispositivo de supervisión. Normalmente, este tipo de dispositivo electrónico incluye un codificador de audio o un descodificador de audio, donde el codificador o descodificador de audio pueden implementarse directamente mediante un circuito o chip digital, por ejemplo un DSP (procesador de señales digitales) o implementarse mediante un código de software que hace que un procesador ejecute un proceso del código de software. En un codificador de audio, en primer lugar se clasifica una señal de audio, diferentes tipos de señales de audio se codifican en diferentes modos de codificación y después un flujo de bits obtenido tras la codificación se transmite a un descodificador.

Generalmente, una señal de audio se procesa mediante una división en tramas, y cada trama de señal representa una señal de audio de una duración específica. Con referencia a la FIG. 1, una trama de audio que se introduce en un momento dado y que tiene que clasificarse puede denominarse trama de audio actual, y cualquier trama de audio anterior a la trama de audio actual puede denominarse trama de audio histórica. Según una secuencia de tiempo desde la trama de audio actual hasta las tramas de audio históricas, las tramas de audio históricas pueden denotarse secuencialmente como una trama de audio anterior, una segunda trama de audio anterior, una tercera trama de audio anterior y una enésima trama de audio anterior, donde N es mayor o igual a cuatro.

En esta forma de realización, una señal de audio de entrada es una señal de audio de banda ancha muestreada a 16 kHz, y la señal de audio de entrada se divide en tramas usando 20 ms como una trama, es decir, cada trama tiene 320 puntos de muestreo de dominio de tiempo. Antes de extraer un parámetro de característica, una trama de señal de audio de entrada se muestrea primero de manera descendente a una frecuencia de muestreo de 12,8 kHz, es decir, hay 256 puntos de muestreo en cada trama. En lo que sigue, cada trama de señal de audio de entrada se refiere a una trama de señal de audio obtenida después del muestreo descendente.

Con referencia a la FIG. 2, una forma de realización de un procedimiento de clasificación de señales de audio incluye:

S101: Llevar a cabo un procesamiento de división en tramas en una señal de audio de entrada y determinar, según la actividad de voz de una trama de audio actual, si hay que obtener una fluctuación de espectro de frecuencia de la trama de audio actual y almacenar la fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia, donde la fluctuación de espectro de frecuencia denota una fluctuación de energía de un espectro de frecuencia de una señal de audio.

La clasificación de señales de audio se lleva a cabo generalmente en cada trama, y un parámetro se extrae de cada trama de señal de audio para realizar la clasificación, determinar si la trama de señal de audio pertenece a una trama de voz o una trama de música, y realizar una codificación en un modo de codificación correspondiente. En una forma de realización, una fluctuación de espectro de frecuencia de una trama de audio actual puede obtenerse después de llevarse a cabo un proceso de división en tramas en una señal de audio, y después se determina, según la actividad de voz de la trama de audio actual, si hay que almacenar la fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia. En otra forma de realización, tras llevar a cabo el procesamiento de división en tramas en una señal de audio, puede determinarse, según la actividad de voz de una trama de audio actual, si hay que almacenar una fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia, y cuando es necesario almacenar la fluctuación de espectro de frecuencia, la fluctuación de espectro de frecuencia se obtiene y se almacena.

La fluctuación de espectro de frecuencia, flux, denota una fluctuación de energía de corta duración o de larga duración de un espectro de frecuencia de una señal, y es un valor promedio de valores absolutos de diferencias de energía logarítmica entre frecuencias correspondientes de una trama de audio actual y una trama histórica en un espectro de banda baja y media, donde la trama histórica se refiere a cualquier trama anterior a la trama de audio actual. En una forma de realización, una fluctuación de espectro de frecuencia es un valor promedio de valores absolutos de diferencias de energía logarítmica entre frecuencias correspondientes de una trama de audio actual y una trama histórica de la trama de audio actual en un espectro de banda baja y media. En otra forma de realización, una fluctuación de espectro de frecuencia es un valor promedio de valores absolutos de diferencias de energía

logarítmica entre valores pico de espectro de frecuencia correspondientes de una trama de audio actual y una trama histórica en un espectro de banda baja y media.

Con referencia a la FIG. 3, una forma de realización para obtener una fluctuación de espectro de frecuencia incluye las siguientes etapas:

S1011: Obtener un espectro de frecuencia de una trama de audio actual.

En una forma de realización, un espectro de frecuencia de una trama de audio puede obtenerse directamente; en otra forma de realización se obtienen espectros de frecuencia, es decir, espectros de energía, de dos subtramas cualesquiera de una trama de audio actual, y un espectro de frecuencia de la trama de audio actual se obtiene usando un valor promedio de los espectros de frecuencia de las dos subtramas.

S1012: Obtener un espectro de frecuencia de una trama histórica de la trama de audio actual.

La trama histórica se refiere a cualquier trama de audio anterior a la trama de audio actual, y puede ser la tercera trama de audio anterior a la trama de audio actual en una forma de realización.

S1013: Calcular un valor promedio de valores absolutos de diferencias de energía logarítmica entre frecuencias correspondientes de la trama de audio actual y la trama histórica en un espectro de banda baja y media, para usar el valor promedio como una fluctuación de espectro de frecuencia de la trama de audio actual.

En una forma de realización puede calcularse un valor promedio de valores absolutos de diferencias entre la energía logarítmica de todas las celdas (*bins*) de frecuencia de una trama de audio actual en un espectro de banda baja y media y la energía logarítmica de celdas de frecuencia correspondientes de una trama histórica en el espectro de banda baja y media.

En otra forma de realización puede calcularse un valor promedio de valores absolutos de diferencias entre la energía logarítmica de valores pico de espectro de frecuencia de una trama de audio actual en un espectro de banda baja y media y la energía logarítmica de valores pico de espectro de frecuencia correspondientes de una trama histórica en el espectro de banda baja y media.

El espectro de banda baja y media es, por ejemplo, un intervalo de espectro de frecuencia comprendido entre 0 y fs/4 o entre 0 y fs/3.

Se usa un ejemplo en el que una señal de audio de entrada es una señal de audio de banda ancha muestreada a 16 kHz y la señal de audio de entrada usa 20 ms como una trama, una primera FFT de 256 puntos y una FFT posterior de 256 puntos se llevan a cabo en una trama de audio actual cada 20 ms, dos ventanas FFT están solapadas en un 50% y espectros de frecuencia (espectros de energía) de dos subtramas de la trama de audio actual se obtienen y se denotan respectivamente como $C^0(i)$ y $C^1(i)$, i = 0, 1, ..., 127, donde $C^x(i)$ denota un espectro de frecuencia de una x-ésima subtrama. Los datos de una segunda subtrama de una trama anterior tienen que usarse en la FFT de una primera subtrama de la trama de audio actual, donde

$$C^x(i) = rel^2(i) + img^2(i),$$

donde rel(i) e img(i) denotan una parte real y una parte imaginaria de un coeficiente FFT de la i-ésima celda de frecuencia, respectivamente. El espectro de frecuencia C(i) de la trama de audio actual se obtiene calculando el promedio de los espectros de frecuencia de las dos subtramas, donde

$$C(i) = \frac{1}{2}(C^0(i) + C^1(i)).$$

La fluctuación de espectro de frecuencia, flux, de la trama de audio actual es un valor promedio de valores absolutos de diferencias de energía logarítmica entre frecuencias correspondientes de la trama de audio actual y una trama ubicada 60 ms por delante de la trama de audio actual en un espectro de banda baja y media en una forma de realización, y el intervalo puede no ser de 60 ms en otra forma de realización, donde

$$flux = \frac{1}{42}\sum_{i=0}^{42}\left[10\log(C(i)) - 10\log(C_{-3}(i))\right],$$

donde $C_{-3}(i)$ denota un espectro de frecuencia de la tercera trama histórica anterior a la trama de audio actual, es decir, una trama histórica ubicada 60 ms por delante de la trama de audio actual cuando una longitud de trama es de 20 ms en esta forma de realización. Cada forma similar a $X_{-n}()$ en esta memoria descriptiva denota un parámetro X

de la enésima trama histórica de la trama de audio actual, y un subíndice 0 puede omitirse para la trama de audio actual. log(.) denota un algoritmo de base 10.

En otra forma de realización, la fluctuación de espectro de frecuencia, flux, de la trama de audio actual también puede obtenerse usando el siguiente procedimiento, es decir, la fluctuación de espectro de frecuencia, flux, es un valor promedio de valores absolutos de diferencias de energía logarítmica entre valores pico de espectro de frecuencia correspondientes de la trama de audio actual y una trama ubicada 60 ms por delante de la trama de audio actual en un espectro de banda baja y media, donde

$$flux = \frac{1}{K}\sum_{i=0}^{K}\left[10\log(P(i)) - 10\log(P_{-3}(i))\right],$$

donde P(i) denota la energía del i-ésimo valor pico local del espectro de frecuencia de la trama de audio actual, una celda de frecuencia en la que un valor pico local está localizado es una celda de frecuencia, en el espectro de frecuencia, cuya energía es mayor que la energía de una celda de frecuencia superior adyacente y la energía de una celda de frecuencia inferior adyacente, y K denota una cantidad de valores pico locales en el espectro de banda baja y media.

El determinar, según la actividad de voz de una trama de audio actual, si hay que almacenar una fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia puede implementarse de varias maneras.

En una forma de realización, si un parámetro de actividad de voz de la trama de audio denota que la trama de audio es una trama activa, la fluctuación de espectro de frecuencia de la trama de audio se almacena en la memoria de fluctuaciones de espectro de frecuencia; en caso contrario, la fluctuación de espectro de frecuencia no se almacena.

En otra forma de realización se determina, según la actividad de voz de la trama de audio y si la trama de audio es un ataque de energía, si hay que almacenar la fluctuación de espectro de frecuencia en la memoria. Si un parámetro de actividad de voz de la trama de audio denota que la trama de audio es una trama activa, y un parámetro que denota si la trama de audio es un ataque de energía denota que la trama de audio no pertenece a un ataque de energía, la fluctuación de espectro de frecuencia de la trama de audio se almacena en la memoria de fluctuaciones de espectro de frecuencia; en caso contrario, la fluctuación de espectro de frecuencia no se almacena. En otra forma de realización, si la trama de audio actual es una trama activa y ninguna de múltiples tramas consecutivas, que incluyen la trama de audio actual y una trama histórica de la trama de audio actual, pertenece a un ataque de energía, la fluctuación de espectro de frecuencia de la trama de audio se almacena en la memoria de fluctuaciones de espectro de frecuencia; en caso contrario, la fluctuación de espectro de frecuencia no se almacena. Por ejemplo, si la trama de audio actual es una trama activa y ninguna de entre la trama de audio actual, una trama de audio anterior y una segunda trama de audio anterior pertenece a un ataque de energía, la fluctuación de espectro de frecuencia de la trama de audio se almacena en la memoria de fluctuaciones de espectro de frecuencia; en caso contrario, la fluctuación de espectro de frecuencia no se almacena.

Un indicador de actividad de voz, ind_vad, denota si una señal de entrada actual es una señal activa en primer plano (voz, música o similar) o una señal silenciosa en segundo plano (tal como ruido de fondo o silencio) de una señal en primer plano, y se obtiene mediante un detector de actividad de voz VAD. Ind_vad = 1 denota que la trama de señal de entrada es una trama activa, es decir, una trama de señal en primer plano; en caso contrario, ind_vad = 0 denota una trama de señal en segundo plano. Puesto que el VAD no pertenece al contenido inventivo de la presente invención, un algoritmo específico del VAD no se describe en detalle en el presente documento.

Un indicador de ataque de voz, ind_ataque, denota si la trama de audio actual pertenece a un ataque de energía relativo a la música. Cuando varias tramas históricas anteriores a la trama de audio actual son principalmente tramas de música, si la energía de trama de la trama de audio actual aumenta de manera relativamente considerable con respecto a la de una primera trama histórica anterior a la trama de audio actual, y aumenta de manera relativamente considerable con respecto a la energía promedio de tramas de audio que están dentro de un periodo de tiempo por delante de la trama de audio actual, y una envolvente de dominio de tiempo de la trama de audio actual también aumenta de manera relativamente considerable con respecto a una envolvente promedio de tramas de audio que están dentro de un periodo de tiempo por delante de la trama de audio actual, se considera que la trama de audio actual pertenece a un ataque de energía relativo a la música.

Según la actividad de voz de la trama de audio actual, la fluctuación de espectro de frecuencia de la trama de audio actual se almacena solamente cuando la trama de audio actual es una trama activa, lo que puede reducir la tasa de interpretaciones erróneas de una trama inactiva y mejorar la velocidad de reconocimiento de una clasificación de audio.

Cuando se satisfacen las siguientes condiciones, ind_ataque se fija a 1, es decir, denota que la trama de audio actual es un ataque de energía de un fragmento de música:

$$\begin{cases} etot - etot_{-1} > 6 \\ etot - lp\_voz > 5 \\ \mathrm{mod}\, e\_mov > 0.9 \\ \log\_\max\_spl - mov\_\log\_\max\_spl > 5 \end{cases},$$

donde etot denota la energía de trama logarítmica de la trama de audio actual; $etot_{-1}$ denota la energía de trama logarítmica de una trama de audio anterior; lp_voz denota un promedio móvil a largo plazo de la energía de trama logarítmica etot; log_max_spl y mov_log_max_spl denotan, respectivamente, una amplitud máxima de puntos de muestreo logarítmicos de dominio de tiempo de la trama de audio actual y un promedio móvil a largo plazo de la amplitud máxima de puntos de muestreo logarítmicos de dominio de tiempo; y mode_mov denota un promedio móvil a largo plazo de resultados históricos de clasificación final de la clasificación de señales.

El significado de la anterior fórmula es el siguiente: cuando varias tramas históricas anteriores a la trama de audio actual son principalmente tramas de música, si la energía de trama de la trama de audio actual aumenta de manera relativamente considerable con respecto a la de una primera trama histórica anterior a la trama de audio actual, y aumenta de manera relativamente considerable con respecto a la energía promedio de tramas de audio que están dentro de un periodo de tiempo por delante de la trama de audio actual, y una envolvente de dominio de tiempo de la trama de audio actual también aumenta de manera relativamente considerable con respecto a una envolvente promedio de tramas de audio que están dentro de un periodo de tiempo por delante de la trama de audio actual, se considera que la trama de audio actual pertenece a un ataque de energía relativo a la música.

La energía de trama logarítmica etot se denota mediante energía de subbanda total logarítmica de una trama de audio de entrada:

$$etot = 10\log(\sum_{j=0}^{19}\left[\frac{1}{hb(i)-lb(i)+1}\cdot\sum_{i=lb(j)}^{hb(j)}C(i)\right]),$$

donde hb(j) y lb(j) denotan, respectivamente, un límite de alta frecuencia y un límite de baja frecuencia de la j-ésima subbanda en un espectro de frecuencia de la trama de audio de entrada; y C(i) denota el espectro de frecuencia de la trama de audio de entrada.

El promedio móvil a largo plazo, mov_log_max_spl, de la amplitud máxima de puntos de muestreo logarítmicos de dominio de tiempo de la trama de audio actual solo se actualiza en una trama de voz activa:

$$mov\_\log\_\max\_spl = \begin{cases} 0.95\cdot mov\_\log\_\max\_spl_{-1} + 0.05\cdot\log\_\max\_spl & \log\_\max\_spl > mov\_\log\_\max\_spl_{-1} \\ 0.995\cdot mov\_\log\_\max\_spl_{-1} + 0.005\cdot\log\_\max\_spl & \log\_\max\_spl \le mov\_\log\_\max\_spl_{-1} \end{cases}.$$

En una forma de realización, la fluctuación de espectro de frecuencia, flux, de la trama de audio actual se almacena en una memoria intermedia FIFO de datos históricos de flux. En esta forma de realización, la longitud de la memoria intermedia de datos históricos de flux es de 60 (60 tramas). Se determina la actividad de voz de la trama de audio actual y si la trama de audio es un ataque de energía, y cuando la trama de audio actual es una trama de señal en primer plano y ninguna de entre la trama de audio actual y dos tramas anteriores a la trama de audio actual pertenece a un ataque de energía de música, la fluctuación de espectro de frecuencia, flux, de la trama de audio actual se almacena en la memoria.

Antes de almacenar la fluctuación flux de la trama de audio actual, se comprueba si se satisfacen las siguientes condiciones:

$$\begin{cases} ind\_vad \neq 0 \\ ind\_ataque \neq 1 \\ ind\_ataque_{-1} \neq 1 \\ ind\_ataque_{-2} \neq 1 \end{cases};$$

si se satisfacen las condiciones, se almacena la fluctuación flux; en caso contrario, no se almacena la fluctuación flux.

Ind_vad denota si la señal de entrada actual es una señal activa en primer plano o una señal silenciosa en segundo plano de una señal en primer plano, e ind_vad = 0 denota una trama de señal en segundo plano; e ind_ataque denota si la trama de audio actual pertenece a un ataque de energía en música, e ind_ataque = 1 denota que la trama de audio actual es un ataque de energía en un fragmento de música.

El significado de la fórmula anterior es el siguiente: la trama de audio actual es una trama activa y ninguna de entre la trama de audio actual, la trama de audio anterior y la segunda trama de audio anterior pertenece a un ataque de energía.

S102: Actualizar, dependiendo de si la trama de audio es música percutante, las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia.

En una forma de realización según la invención, si un parámetro que denota si la trama de audio pertenece a música percutante denota que la trama de audio actual pertenece a música percutante, los valores de las fluctuaciones de espectro de frecuencia almacenados en la memoria de fluctuaciones de espectro de frecuencia se modifican, y valores válidos de fluctuación de espectro de frecuencia de la memoria de fluctuaciones de espectro de frecuencia se modifican pasando a ser un valor inferior o igual a un umbral de música, donde si una fluctuación de espectro de frecuencia de una trama de audio es menor que el umbral de música, el audio se clasifica como una trama de música. En una forma de realización, los valores válidos de fluctuación de espectro de frecuencia se fijan a 5. Es decir, cuando un indicador de sonido percutante percus_flag está fijado a 1, todos los datos válidos de la memoria intermedia de datos históricos de flux se fijan 5. En el presente documento, los datos válidos de memoria intermedia son equivalentes a un valor válido de fluctuación de espectro de frecuencia. Generalmente, un valor de fluctuación de espectro de frecuencia de una trama de música es relativamente pequeño, mientras que un valor de fluctuación de espectro de frecuencia de una trama de voz es relativamente grande. Cuando la trama de audio pertenece a música percutante, los valores válidos de fluctuación de espectro de frecuencia se modifican pasando a ser un valor inferior o igual al umbral de música, lo que puede mejorar la probabilidad de que la trama de audio se clasifique como una trama de música, mejorándose así la precisión de la clasificación de una señal de audio.

En otra forma de realización, las fluctuaciones de espectro de frecuencia de la memoria se actualizan según la actividad de una trama histórica de la trama de audio actual. Específicamente, en una forma de realización, si se determina que la fluctuación de espectro de frecuencia de la trama de audio actual se almacena en la memoria de fluctuaciones de espectro de frecuencia, y una trama de audio anterior es una trama inactiva, los datos de otras fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia, excepto la fluctuación de espectro de frecuencia de la trama de audio actual, se modifican pasando a ser datos ineficaces. Cuando la trama de audio anterior es una trama inactiva y la trama de audio actual es una trama activa, la actividad de voz de la trama de audio actual es diferente a la de la trama histórica y una fluctuación de espectro de frecuencia de la trama histórica se invalida, lo que puede reducir el impacto de la trama histórica en la clasificación de audio, mejorándose así la precisión de la clasificación de una señal de audio.

En otra forma de realización, si se determina que la fluctuación de espectro de frecuencia de la trama de audio actual se almacena en la memoria de fluctuaciones de espectro de frecuencia, y tres tramas consecutivas anteriores a la trama de audio actual no son todas ellas tramas activas, la fluctuación de espectro de frecuencia de la trama de audio actual se modifica pasando a ser un primer valor. El primer valor puede ser un umbral de voz, donde si la fluctuación de espectro de frecuencia de la trama de audio es mayor que el umbral de voz, el audio se clasifica como una trama de voz. En otra forma de realización, si se determina que la fluctuación de espectro de frecuencia de la trama de audio actual se almacena en la memoria de fluctuaciones de espectro de frecuencia, y el resultado de clasificación de una trama histórica es una trama de música y la fluctuación de espectro de frecuencia de la trama de audio actual es mayor que un segundo valor, la fluctuación de espectro de frecuencia de la trama de audio actual se modifica pasando a ser el segundo valor, donde el segundo valor es mayor que el primer valor.

Si se almacena la fluctuación flux de la trama de audio actual y la trama de audio anterior es una trama inactiva (ind_vad = 0), excepto la trama de audio actual, flux, almacenada recientemente en la memoria intermedia de datos históricos de flux, todos los datos restantes de la memoria intermedia de datos históricos de flux se fijan a -1 (lo que equivale a que los datos se invaliden).

Si flux se almacena en la memoria intermedia de datos históricos de flux y tres tramas consecutivas anteriores a la trama de audio actual no son todas ellas tramas activas (ind_vad = 1), la flux de trama de audio actual que acaba de almacenarse en la memoria intermedia de datos históricos de flux se modifica pasando a ser 16; es decir, se comprueba si se satisfacen las siguientes condiciones:

$$\begin{cases} ind\_vad_{-1} = 1 \\ ind\_vad_{-2} = 1 \\ ind\_vad_{-3} = 1 \end{cases};$$

si las condiciones no se satisfacen, la flux de trama de audio actual que acaba de almacenarse en la memoria intermedia de datos históricos de flux se modifica pasando a ser 16; y
si las tres tramas consecutivas anteriores a la trama de audio actual son todas ellas tramas activas (ind_vad = 1), se comprueba si se satisfacen las siguientes condiciones:

$$\begin{cases} \mathrm{mod}\,e\_mov > 0.9 \\ flux > 20 \end{cases};$$

si las condiciones se satisfacen, la flux de trama de audio actual que acaba de almacenarse en la memoria intermedia de datos históricos de flux se modifica pasando a ser 20; en caso contrario, no se realiza ninguna operación,
donde mode_mov denota un promedio móvil a largo plazo de resultados de clasificación finales históricos en la clasificación de señales; mode_mov > 0,9 denota que la señal está en una señal de música, y flux se limita según el resultado de clasificación histórica de la señal de audio para reducir la probabilidad de que una característica de voz se produzca en flux y mejorar la estabilidad a la hora de determinar la clasificación.

Cuando las tres tramas históricas consecutivas anteriores a la trama de audio actual son todas ellas tramas inactivas, y la trama de audio actual es una trama activa, o cuando las tres tramas consecutivas anteriores a la trama de audio actual no son todas ellas tramas activas, y la trama de audio actual es una trama activa, la clasificación está en una fase de inicialización. En una forma de realización, para hacer que el resultado de la clasificación tienda a ser voz (música), la fluctuación de espectro de frecuencia de la trama de audio actual puede modificarse pasando a ser un umbral de voz (música) o un valor cercano al umbral de voz (música). En otra forma de realización, si una señal anterior a una señal actual es una señal de voz (música), la fluctuación de espectro de frecuencia de la trama de audio actual puede modificarse pasando a ser un umbral de voz (música) o un valor cercano al umbral de voz (música) para mejorar la estabilidad a la hora de determinar la clasificación. En otra forma de realización, para hacer que el resultado de la clasificación tienda a ser música, la fluctuación de espectro de frecuencia puede limitarse, es decir, la fluctuación de espectro de frecuencia de la trama de audio actual puede modificarse, de manera que la fluctuación de espectro de frecuencia no es mayor que un umbral, con el fin de reducir la probabilidad de determinar que la fluctuación de espectro de frecuencia es una característica de voz.

El indicador de sonido percutante, percus_flag, denota si existe un sonido percutante en una trama de audio. Si percus_flag está fijado a 1 denota que se ha detectado un sonido percutante, y si percus_flag está fijado a 0 denota que no se ha detectado ningún sonido percutante.

Cuando un pico de energía relativamente preciso se produce en la señal actual (es decir, las últimas tramas de señal que incluyen la trama de audio actual y varias tramas históricas de la trama de audio actual) tanto de corta duración como de larga duración, y la señal actual no tiene ninguna característica de sonido sonoro perceptible, si las diversas tramas históricas anteriores a la trama de audio actual son principalmente tramas de música, se considera que la señal actual es un fragmento de música percutante; en caso contrario, si ninguna de las subtramas de la señal actual tiene una característica de sonido sonoro perceptible y además se produce un incremento relativamente evidente en la envolvente de dominio de tiempo de la señal actual con respecto a un promedio a largo plazo de la envolvente de dominio de tiempo, también se considera que la señal actual es un fragmento de música percutante.

El indicador de sonido percutante, percus_flag, se obtiene llevando a cabo la siguiente etapa.

Primero se obtiene la energía de trama logarítmica etot de una trama de audio de entrada, donde la energía de trama logarítmica etot se denota mediante la energía de subbanda total logarítmica de la trama de audio de entrada:

$$etot = 10\log\left(\sum_{j=0}^{19}\left[\frac{1}{hb(i)-lb(i)+1}\cdot\sum_{i=lb(j)}^{hb(j)}C(i)\right]\right),$$

donde hb(j) y lb(j) denotan un límite de alta frecuencia y un límite de baja frecuencia de la j-ésima subbanda en un espectro de frecuencia de la trama de entrada, respectivamente, y C(i) denota el espectro de frecuencia de la trama de audio de entrada.

Cuando se satisfacen las siguientes condiciones, percus_flag se fija a 1; en caso contrario, percus_flag se fija a 0:

$$\begin{cases} etot_{-2} - etot_{-3} > 6 \\ etot_{-2} - etot_{-1} > 0 \\ etot_{-2} - etot > 3 \\ etot_{-1} - etot > 0 \\ etot_{-2} - lp\_voz\_{-2} > 3 \\ 0.5 \cdot sonoridad_{-1}(1) + 0.25 \cdot sonoridad(0) + 0.25 \cdot sonoridad(1) < 0.75 \\ \mathrm{mod}\, e\_mov > 0.9 \end{cases},$$

o

$$\begin{cases} etot_{-2} - etot_{-3} > 6 \\ etot_{-2} - etot_{-1} > 0 \\ etot_{-2} - etot > 3 \\ etot_{-1} - etot > 0 \\ etot_{-2} - lp\_voz\_{-2} > 3 \\ 0.5 \cdot sonoridad_{-1}(1) + 0.25 \cdot sonoridad(0) + 0.25 \cdot sonoridad(1) < 0.75 \\ sonoridad_{-1}(0) < 0.8 \\ sonoridad_{-1}(1) < 0.8 \\ sonoridad(0) < 0.8 \\ \log\_\max\_spl_{-2} - mov\_\log\_\max\_spl_{-2} > 10 \end{cases},$$

donde etot denota la energía de trama logarítmica de la trama de audio actual; lp_voz denota un promedio móvil a largo plazo de la energía de trama logarítmica, etot; sonoridad(0), sonoridad$_{-1}$(0) y sonoridad$_{-1}$(1) denotan grados de correlación normalizados de tono de bucle abierto de una primera subtrama de una trama de audio de entrada actual y de una primera y segunda subtramas de una primera trama histórica, respectivamente, y un parámetro de sonoridad, sonoridad, se obtiene mediante predicción lineal y análisis, representa un grado de correlación de dominio de tiempo entre la trama de audio actual y una señal anterior a un periodo de tono, y tiene un valor comprendido entre 0 y 1; mode_mov denota un promedio móvil a largo plazo de resultados históricos de clasificación final en la clasificación de señales; log_max_spl$_{-2}$ y mov_log_max_spl$_{-2}$ denotan, respectivamente, una amplitud máxima de puntos de muestreo logarítmicos de dominio de tiempo de una segunda trama histórica y un promedio móvil a largo plazo de la amplitud máxima de puntos de muestreo logarítmicos de dominio de tiempo. Lp_voz se actualiza en cada trama de voz activa (es decir, una trama cuyo ind_vad = 1), y un procedimiento para actualizar lp_voz es:

$$lp\_voz\_{-1} = 0.99 \cdot lp\_voz\_{-1} + 0.01 \cdot etot.$$

El significado de las dos fórmulas anteriores es el siguiente: cuando un pico de energía relativamente preciso se produce en la señal actual (es decir, las últimas tramas de señal que incluyen la trama de audio actual y varias tramas históricas de la trama de audio actual) tanto de corta duración como de larga duración, y la señal actual no tiene ninguna característica de sonido sonoro perceptible, si las diversas tramas históricas anteriores a la trama de audio actual son principalmente tramas de música, se considera que la señal actual es un fragmento de música percutante; en caso contrario, si ninguna de las subtramas de la señal actual tiene una característica de sonido sonoro perceptible y además se produce un incremento relativamente evidente en la envolvente de dominio de tiempo de la señal actual con respecto a un promedio a largo plazo de la misma, también se considera que la señal actual es un fragmento de música percutante.

El parámetro de sonoridad, sonoridad, es decir, un grado de correlación de tonos de bucle abierto normalizado, denota un grado de correlación de domino de tiempo entre la trama de audio actual y una señal antes de un periodo de tono, puede obtenerse mediante una búsqueda de tonos de bucle abierto ACELP, y tiene un valor entre 0 y 1. Esto pertenece a la técnica anterior y, por lo tanto, no se describe en detalle en la presente invención. En esta forma de realización, la sonoridad se calcula para cada una de dos subtramas de la trama de audio actual, y las sonoridades se promedian para obtener un parámetro de sonoridad de la trama de audio actual. El parámetro de sonoridad de la trama de audio actual también se almacena en una memoria intermedia de datos históricos de sonoridad y, en esta forma de realización, la longitud de la memoria intermedia de datos históricos de sonoridad es 10.

Mode_mov se actualiza en cada trama de voz activa y cuando se han producido más de 30 tramas de voz activas consecutivas antes de la trama, y un procedimiento de actualización es:

$$mode\_mov = 0.95 \cdot move\_mov_{-1} + 0.05 \cdot mode,$$

donde mode es un resultado de clasificación de una trama de audio de entrada actual y tiene un valor binario, donde "0" denota una categoría de voz y "1" denota una categoría de música.

S103: Clasificar la trama de audio actual como una trama de voz o una trama de música según las estadísticas de una parte de o todos los datos de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia. Cuando las estadísticas de datos eficaces de las fluctuaciones de espectro de frecuencia satisfacen una condición de clasificación de voz, la trama de audio actual se clasifica como una trama de voz; cuando las estadísticas de los datos eficaces de las fluctuaciones de espectro de frecuencia satisfacen una condición de clasificación de música, la trama de audio actual se clasifica como una trama de música.

En el presente documento, las estadísticas son un valor obtenido llevando a cabo una operación estadística en una fluctuación de espectro de frecuencia válida (es decir, datos eficaces) almacenada en la memoria de fluctuaciones de espectro de frecuencia. Por ejemplo, la operación estadística puede ser una operación para obtener un valor promedio o una varianza. Las estadísticas de las siguientes formas de realización tienen un significado similar.

En una forma de realización, la etapa S103 incluye:

> obtener un valor promedio de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia; y
> cuando el valor promedio obtenido de los datos eficaces de las fluctuaciones de espectro de frecuencia satisface una condición de clasificación de música, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz.

Por ejemplo, cuando el valor promedio obtenido de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un umbral de clasificación de música, la trama de audio actual se clasifica como una trama de música; en caso contrario, la trama de audio actual se clasifica como una trama de voz.

Generalmente, un valor de fluctuación de espectro de frecuencia de una trama de música es relativamente pequeño, mientras que un valor de fluctuación de espectro de frecuencia de una trama de voz es relativamente grande. Por lo tanto, la trama de audio actual puede clasificarse según las fluctuaciones de espectro de frecuencia. Evidentemente, la clasificación de señales también puede llevarse a cabo en la trama de audio actual usando otro procedimiento de clasificación. Por ejemplo, se cuenta la cantidad de datos eficaces de la fluctuaciones de espectro de frecuencia almacenados en la memoria de fluctuaciones de espectro de frecuencia; la memoria de fluctuaciones de espectro de frecuencia se divide, según la cantidad de datos eficaces, en al menos dos intervalos de diferente longitud desde un extremo cercano a un extremo remoto, y se obtiene un valor promedio de datos eficaces de fluctuaciones de espectro de frecuencia correspondientes a cada intervalo, donde un punto inicial de los intervalos es una ubicación de almacenamiento de la fluctuación de espectro de frecuencia de la trama actual, el extremo cercano es un extremo en el que se almacena la fluctuación de espectro de frecuencia de la trama actual, y el extremo remoto es un extremo en el que se almacena una fluctuación de espectro de frecuencia de una trama histórica; la trama de audio se clasifica según las estadísticas de fluctuaciones de espectro de frecuencia en un intervalo relativamente corto, y si las estadísticas de los parámetros en este intervalo son suficientes para distinguir un tipo de la trama de audio, el proceso de clasificación termina; en caso contrario, el proceso de clasificación continúa en el intervalo más corto de los intervalos restantes relativamente largos, y el resto puede deducirse por analogía. En un proceso de clasificación de cada intervalo, la trama de audio actual se clasifica según un umbral de clasificación correspondiente a cada intervalo, la trama de audio actual se clasifica como una trama de voz o una trama de música, y cuando las estadísticas de los datos eficaces de las fluctuaciones de espectro de frecuencia satisfacen la condición de clasificación de voz, la trama de audio actual se clasifica como una trama de voz; cuando las estadísticas de los datos eficaces de las fluctuaciones de espectro de frecuencia satisfacen la condición de clasificación de música, la trama de audio actual se clasifica como una trama de música.

Tras la clasificación de señales, diferentes señales pueden codificarse en diferentes modos de codificación. Por ejemplo, una señal de voz se codifica usando un codificador basado en un modelo de generación de voz (tal como CELP), y una señal de música se codifica usando un codificador basado en conversión (tal como un codificador basado en MDCT).

En la forma de realización anterior, puesto que una señal de audio se clasifica según estadísticas a largo plazo de fluctuaciones de espectro de frecuencia, hay un número relativamente bajo de parámetros, la velocidad de reconocimiento es relativamente alta y la complejidad es relativamente baja. Además, las fluctuaciones de espectro de frecuencia se ajustan teniendo en cuenta factores tales como la actividad de voz y la música percutante; por lo tanto, la presente invención tiene una mayor velocidad de reconocimiento para una señal de música, y es adecuada para la clasificación de señales de audio híbridas.

Con referencia a la FIG. 4, en otra forma de realización, después de la etapa S102, el procedimiento incluye además:

S104: Obtener un máximo de banda de alta frecuencia de espectro de frecuencia, un grado de correlación de espectro de frecuencia y una pendiente de energía residual de predicción lineal de la trama de audio actual, y almacenar en memorias el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal, donde el máximo de banda de alta frecuencia de espectro de frecuencia denota un máximo o un pico de energía, en una banda de alta frecuencia, de un espectro de frecuencia de la trama de audio actual; el grado de correlación de espectro de frecuencia denota estabilidad, entre tramas adyacentes, de una estructura armónica de señal; y la pendiente de energía residual de predicción lineal denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio de entrada a medida que aumenta el orden de predicción lineal.

Opcionalmente, antes de almacenar estos parámetros, el procedimiento incluye además: determinar, según la actividad de voz de la trama de audio actual, si hay que almacenar en las memorias el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal; y si la trama de audio actual es una trama activa, almacenar los parámetros; en caso contrario, no almacenar los parámetros.

El máximo de banda de alta frecuencia de espectro de frecuencia denota un máximo o un pico de energía, en una banda de alta frecuencia, de un espectro de frecuencia de la trama de audio actual. En una forma de realización, el máximo de banda de alta frecuencia de espectro de frecuencia, ph, se calcula usando la siguiente fórmula:

$$ph = \sum_{i=64}^{126} p2v\_map(i),$$

donde p2v_map(i) denota un máximo de la i-ésima celda de frecuencia de un espectro de frecuencia, y el máximo p2v_map(i) se obtiene usando la siguiente fórmula:

$$p2v\_map(i) = \begin{cases} 20\log(\text{máx}(i)) - 10\log(vl(i)) - 10\log(vr(i)) & máx(i) \neq 0 \\ 0 & máx(i) = 0 \end{cases},$$

donde máx(i) = C(i) si la i-ésima celda de frecuencia es un valor pico local del espectro de frecuencia; en caso contrario máx(i) = 0; y vl(i) y vr(i) denotan valores valle locales de espectro de frecuencia v(n) que son los más adyacentes a la i-ésima celda de frecuencia en un lado de alta frecuencia y un lado de baja frecuencia de la i-ésima celda de frecuencia, respectivamente, donde

$$máx(i) = \begin{cases} C(i) & C(i) > C(i-1), C(i) > C(i+1) \\ 0 & en\ caso\ contrario \end{cases},$$

y

$$v = \forall C(i) \qquad C(i) < C(i-1), C(i) < C(i+1).$$

El máximo de banda de alta frecuencia de espectro de frecuencia, ph, de la trama de audio actual también se almacena en una memoria intermedia de datos históricos de ph y, en esta forma de realización, la longitud de la memoria intermedia de datos históricos de ph es 60.

El grado de correlación de espectro de frecuencia, cor_map_sum, denota estabilidad, entre tramas adyacentes, de una estructura armónica de señal, y se obtiene llevando a cabo las siguientes etapas:

En primer lugar, se obtiene un espectro de frecuencia C'(i) sin el límite inferior (*floor*) de una trama de audio de entrada C(i), donde

$$C'(i) = C(i) - floor(i),$$

donde floor(i) denota un límite inferior de espectro de un espectro de frecuencia de la trama de audio de entrada, donde i = 0, 1, ..., 127; y

$$floor(i) = \begin{cases} C(i) & C(i) \in v \\ vl(i) + (i - idx[vl(i)]) \cdot \dfrac{vr(i) - vl(i)}{idx[vr(i)] - idx[vl(i)]} & en\ caso\ contrario \end{cases},$$

donde idx[x] denota una ubicación de x en el espectro de frecuencia, donde idx[x] = 0, 1, ..., 127.

Después, entre cada dos valores valle adyacentes de espectro de frecuencia, se obtiene una correlación, cor(n), entre el espectro de frecuencia sin límite inferior de la trama de audio de entrada y un espectro de frecuencia sin límite inferior de una trama anterior, donde

$$cor(n) = \frac{\left(\sum_{i=lb(n)}^{hb(n)} C'(i) \cdot C'_{-1}(i) \right)^2}{\left(\sum_{i=lb(n)}^{hb(n)} C'(i) \cdot C'(i) \right) \cdot \left(\sum_{i=lb(n)}^{hb(n)} C'_{-1}(i) \cdot C'_{-1}(i) \right)},$$

donde lb(n) y hb(n) denotan respectivamente ubicaciones de puntos finales del enésimo intervalo de valores valle de espectro de frecuencia (es decir, un área situada entre dos valores valle adyacentes), es decir, ubicaciones que limitan dos valores valle de espectro de frecuencia del intervalo de valores valle.

Finalmente se calcula el grado de correlación de espectro de frecuencia, cor_map_sum, de la trama de audio de entrada usando la siguiente fórmula:

$$cor\_map\_sum = \sum_{i=0}^{127} cor(inv[lb(n) \le i, hb(n) \ge i]),$$

donde inv[f] denota una función inversa de una función f.

La pendiente de energía residual de predicción lineal, epsP_tilt, denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio de entrada a medida que aumenta el orden de predicción lineal, y puede calcularse y obtenerse usando la siguiente fórmula:

$$epsP\_tilt = \frac{\sum_{i=1}^{n} epsP(i) \cdot epsP(i+1)}{\sum_{i=1}^{n} epsP(i) \cdot epsP(i)}.$$

donde epsP(i) denota la energía residual de predicción de predicción lineal de orden i-ésimo; y n es un entero positivo, denota un orden de predicción lineal y es inferior o igual a un orden máximo de predicción lineal. Por ejemplo, en una forma de realización, n = 15.

Por lo tanto, la etapa S103 puede sustituirse por la siguiente etapa:

S105: Obtener estadísticas de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, estadísticas de datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, estadísticas de datos eficaces de grados de correlación de espectro de frecuencia almacenados, y estadísticas de datos eficaces de pendientes de energía residual de predicción lineal almacenadas, y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de los datos eficaces, donde las estadísticas de los datos eficaces se refieren a un valor de datos obtenido tras realizar una operación de cálculo en los datos eficaces almacenados en las memorias, donde la operación de cálculo puede incluir una operación para obtener un valor promedio, una operación para obtener una varianza o similares.

En una forma de realización, esta etapa incluye:

obtener por separado un valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, un valor promedio de los datos eficaces de los máximos almacenados de banda de alta frecuencia de espectro de frecuencia, un valor promedio de los datos eficaces de los grados de correlación de

espectro de frecuencia almacenados y una varianza de los datos eficaces de pendientes de energía residual de predicción lineal almacenadas; y

cuando se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz: el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

Generalmente, un valor de fluctuación de espectro de frecuencia de una trama de música es relativamente pequeño, mientras que un valor de fluctuación de espectro de frecuencia de una trama de voz es relativamente grande; un valor pico de banda de alta frecuencia de espectro de frecuencia de una trama de música es relativamente grande, y un valor pico de banda de alta frecuencia de espectro de frecuencia de una trama de voz es relativamente pequeño; un valor de grado de correlación de espectro de frecuencia de una trama de música es relativamente grande, y un valor de grado de correlación de espectro de frecuencia de una trama de voz es relativamente pequeño; un cambio en una pendiente de energía residual de predicción lineal de una trama de música es relativamente pequeño, y un cambio en una pendiente de energía residual de predicción lineal de una trama de voz es relativamente grande. Por lo tanto, la trama de audio actual puede clasificarse según las estadísticas de los parámetros anteriores. Evidentemente, la clasificación de señales también puede llevarse a cabo en la trama de audio actual usando otro procedimiento de clasificación. Por ejemplo, se cuenta la cantidad de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia; la memoria se divide, según la cantidad de datos eficaces, en al menos dos intervalos de diferente longitud desde un extremo cercano a un extremo remoto, se obtiene un valor promedio de datos eficaces de fluctuaciones de espectro de frecuencia correspondientes a cada intervalo, un valor promedio de datos eficaces de máximos de banda de alta frecuencia de espectro de frecuencia, un valor promedio de datos eficaces de grados de correlación de espectro de frecuencia y una varianza de datos eficaces de pendientes de energía residual de predicción lineal, donde un punto inicial de los intervalos es una ubicación de almacenamiento de la fluctuación de espectro de frecuencia de la trama actual, el extremo cercano es un extremo en el que se almacena la fluctuación de espectro de frecuencia de la trama actual, y el extremo remoto es un extremo en el que se almacena una fluctuación de espectro de frecuencia de una trama histórica; la trama de audio se clasifica según las estadísticas de datos eficaces de los parámetros anteriores en un intervalo relativamente corto, y si las estadísticas de los parámetros en este intervalo son suficientes para distinguir el tipo de la trama de audio, el proceso de clasificación termina; en caso contrario, el proceso de clasificación continúa en el intervalo más corto de los intervalos restantes relativamente largos, y el resto puede deducirse por analogía. En un proceso de clasificación de cada intervalo, la trama de audio actual se clasifica según un umbral de clasificación correspondiente a cada intervalo, y cuando una de las siguientes condiciones se satisface, la trama de audio actual se clasifica como una trama de música; en caso contrario, la trama de audio actual se clasifica como una trama de voz: el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

Tras la clasificación de señales, diferentes señales pueden codificarse en diferentes modos de codificación. Por ejemplo, una señal de voz se codifica usando un codificador basado en un modelo de generación de voz (tal como CELP), y una señal de música se codifica usando un codificador basado en conversión (tal como un codificador basado en MDCT).

En la anterior forma de realización, una señal de audio se clasifica según estadísticas a largo plazo de fluctuaciones de espectro de frecuencia, máximos de banda de alta de frecuencia de espectro de frecuencia, grados de correlación de espectro de frecuencia y pendientes de energía residual de predicción lineal; por lo tanto, hay número relativamente bajo de parámetros, la velocidad de reconocimiento es relativamente alta y la complejidad es relativamente baja. Además, las fluctuaciones de espectro de frecuencia se ajustan teniendo en cuenta factores tales como la actividad de voz y la música percutante, y las fluctuaciones de espectro de frecuencia se modifican según un entorno de señal en el que está ubicada la trama de audio actual; por lo tanto, la presente invención mejora la velocidad de reconocimiento de clasificación y es adecuada para la clasificación de señales de audio híbridas.

Con referencia a la FIG. 5, otra forma de realización de un procedimiento de clasificación de señales de audio incluye:

S501: Llevar a cabo un procesamiento de división en tramas en una señal de audio de entrada.

La clasificación de señales de audio se lleva a cabo generalmente en cada trama, y un parámetro se extrae de cada trama de señal de audio para realizar la clasificación, determinar si la trama de señal de audio pertenece a una trama de voz o una trama de música, y realizar una codificación en un modo de codificación correspondiente.

S502: Obtener una pendiente de energía residual de predicción lineal de una trama de audio actual, donde la pendiente de energía residual de predicción lineal denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio a medida que aumenta el orden de predicción lineal.

En una forma de realización, la pendiente de energía residual de predicción lineal, epsP_tilt, puede calcularse y obtenerse usando la siguiente fórmula:

$$epsP\_tilt = \frac{\sum_{i=1}^{n} epsP(i) \cdot epsP(i+1)}{\sum_{i=1}^{n} epsP(i) \cdot epsP(i)},$$

donde epsP(i) denota la energía residual de predicción de predicción lineal de orden i-ésimo; y n es un entero positivo, denota un orden de predicción lineal y es inferior o igual a un orden máximo de predicción lineal. Por ejemplo, en una forma de realización, n = 15.

S503: Almacenar la pendiente de energía residual de predicción lineal en una memoria.

La pendiente de energía residual de predicción lineal puede almacenarse en la memoria. En una forma de realización, la memoria puede ser una memoria intermedia FIFO, y la longitud de la memoria intermedia es de 60 unidades de almacenamiento (es decir, puede almacenarse 60 pendientes de energía residual de predicción lineal).

Opcionalmente, antes de almacenar la pendiente de energía residual de predicción lineal, el procedimiento incluye además: determinar, según la actividad de voz de la trama de audio actual, si hay que almacenar la pendiente de energía residual de predicción lineal en la memoria; y si la trama de audio actual es una trama activa, almacenar la pendiente de energía residual de predicción lineal; en caso contrario, no almacenar la pendiente de energía residual de predicción lineal.

S504: Clasificar la trama de audio según estadísticas de una parte de datos de pendientes de energía residual de predicción en la memoria.

En una forma de realización, las estadísticas de la parte de datos de las pendientes de energía residual de predicción es una varianza de la parte de los datos de las pendientes de energía residual de predicción y, por lo tanto, la etapa S504 incluye:

> comparar la varianza de la parte de los datos de las pendientes de energía residual de predicción con un umbral de clasificación de música, y cuando la varianza de la parte de los datos de las pendientes de energía residual de predicción es menor que el umbral de clasificación de música, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz.

Generalmente, un cambio en un valor de pendiente de energía residual de predicción lineal de una trama de música es relativamente pequeño, y un cambio en un valor de pendiente de energía residual de predicción lineal de una trama de voz es relativamente grande. Por lo tanto, la trama de audio actual puede clasificarse según las estadísticas de las pendientes de energía residual de predicción lineal. Evidentemente, la clasificación de señales también puede llevarse a cabo en la trama de audio actual con referencia a otro parámetro usando otro procedimiento de clasificación.

En otra forma de realización, antes de la etapa S504, el procedimiento incluye además: obtener una fluctuación de espectro de frecuencia, un máximo de banda de alta frecuencia de espectro de frecuencia y un grado de correlación de espectro de frecuencia de la trama de audio actual, y almacenar la fluctuación de espectro de frecuencia, el máximo de banda de alta frecuencia de espectro de frecuencia y el grado de correlación de espectro de frecuencia en memorias correspondientes. Por lo tanto, la etapa S504 incluye específicamente:

> obtener estadísticas de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, estadísticas de datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, estadísticas de datos eficaces de grados de correlación de espectro de frecuencia almacenados, y estadísticas de datos eficaces de las pendientes de energía residual de predicción lineal almacenadas, y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de los datos eficaces, donde las estadísticas de los datos eficaces se refieren a un valor de datos obtenido tras realizar una operación de cálculo en los datos eficaces almacenados en las memorias.

Además, obtener estadísticas de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, estadísticas de datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, estadísticas de datos eficaces de grados de correlación de espectro de frecuencia almacenados, y estadísticas de

datos eficaces de pendientes de energía residual de predicción lineal almacenadas, así como clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de los datos eficaces incluye:

obtener por separado un valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, un valor promedio de los datos eficaces de los máximos almacenados de banda de alta frecuencia de espectro de frecuencia, un valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia almacenados y una varianza de los datos eficaces de las pendientes de energía residual de predicción lineal almacenadas; y
cuando una de las siguientes condiciones se satisface, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz: el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

Generalmente, un valor de fluctuación de espectro de frecuencia de una trama de música es relativamente pequeño, mientras que un valor de fluctuación de espectro de frecuencia de una trama de voz es relativamente grande; un valor pico de banda de alta frecuencia de espectro de frecuencia de una trama de música es relativamente grande, y un valor pico de banda de alta frecuencia de espectro de frecuencia de una trama de voz es relativamente pequeño; un valor de grado de correlación de espectro de frecuencia de una trama de música es relativamente grande, y un valor de grado de correlación de espectro de frecuencia de una trama de voz es relativamente pequeño; un cambio en un valor de pendiente de energía residual de predicción lineal de una trama de música es relativamente pequeño, y un cambio en un valor de pendiente de energía residual de predicción lineal de una trama de voz es relativamente grande. Por lo tanto, la trama de audio actual puede clasificarse según las estadísticas de los parámetros anteriores.

En otra forma de realización, antes de la etapa S504, el procedimiento incluye además: obtener una cantidad de tonos de espectro de frecuencia de la trama de audio actual y una relación de la cantidad de tonos de espectro de frecuencia en una banda de baja frecuencia, y almacenar en memorias correspondientes la cantidad de tonos de espectro de frecuencia y la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia. Por lo tanto, la etapa S504 incluye específicamente:

obtener por separado estadísticas de las pendientes de energía residual de predicción lineal almacenadas y estadísticas de cantidades de tonos de espectro de frecuencia almacenadas; y
clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de las pendientes de energía residual de predicción lineal, las estadísticas de las cantidades de tonos de espectro de frecuencia y la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia, donde las estadísticas se refieren a un valor de datos obtenido tras realizar una operación de cálculo en datos almacenados en las memorias.

Además, obtener por separado estadísticas de las pendientes de energía residual de predicción lineal almacenadas y estadísticas de cantidades de tonos de espectro de frecuencia almacenadas incluye: obtener una varianza de las pendientes de energía residual de predicción lineal almacenadas; y obtener un valor promedio de las cantidades de tonos de espectro de frecuencia almacenadas. Clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de las pendientes de energía residual de predicción lineal, las estadísticas de las cantidades de tonos de espectro de frecuencia y la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia incluye:

cuando la trama de audio actual es una trama activa y se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz;
la varianza de las pendientes de energía residual de predicción lineal es menor que un quinto umbral; o
el valor promedio de las cantidades de tonos de espectro de frecuencia es mayor que un sexto umbral; o
la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia es menor que un séptimo umbral.

Obtener una cantidad de tonos de espectro de frecuencia de la trama de audio actual y una relación de la cantidad de tonos de espectro de frecuencia en una banda de baja frecuencia incluye:

contar una cantidad de celdas de frecuencia de la trama de audio actual que están en un banda de frecuencia entre 0 y 8 kHz y tienen valores pico de celda de frecuencia mayores que un valor predeterminado, con el fin de usar la cantidad como la cantidad de tonos de espectro de frecuencia; y
calcular una relación de una cantidad de celdas de frecuencia de la trama de audio actual que están en una banda de frecuencia entre 0 y 4 kHz y tienen valores pico de celda de frecuencia mayores que el valor predeterminado con respecto a la cantidad de celdas de frecuencia de la trama de audio actual que están en la banda de frecuencia entre 0 y 8 kHz y tienen valores pico de celda de frecuencia mayores que el valor

predeterminado, con el fin de usar la relación como la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia. En una forma de realización, el valor predeterminado es 50.

La cantidad de tonos de espectro de frecuencia, Ntonal, denota una cantidad de celdas de frecuencia de la trama de audio actual que están en una banda de frecuencia entre 0 y 8 kHz y tienen valores pico de celda de frecuencia mayores que un valor predeterminado. En una forma de realización, la cantidad puede obtenerse de la siguiente manera: contar una cantidad de celdas de frecuencia de la trama de audio actual que están en una banda de frecuencia entre 0 y 8 kHz y tienen valores pico p2v_map(i) mayores que 50, es decir, Ntonal, donde p2v_map(i) denota un máximo de la i-ésima celda de frecuencia del espectro de frecuencia; en lo que respecta a un modo de calcular p2v_map(i) se hace referencia a la descripción de la anterior forma de realización.

La relación ratio_Ntonal_lf de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia denota una relación de una cantidad de tonos de banda de baja frecuencia con respecto a la cantidad de tonos de espectro de frecuencia. En una forma de realización, la relación puede obtenerse de la siguiente manera: contar una cantidad Ntonal_lf de la trama de audio actual que está en una banda de frecuencia entre 0 y 4 kHz y tiene p2v_map(i) mayor que 50. Ratio_Ntonal_lf es una relación de Ntonal_lf con respecto a Ntonal, es decir, Ntonal_lf/Ntonal. P2v_map(i) denota un máximo de la i-ésima celda de frecuencia del espectro de frecuencia; en lo que respecta a un modo de calcular p2v_map(i) se hace referencia a la descripción de la anterior forma de realización. En otra forma de realización se obtienen por separado un promedio de múltiples valores Ntonal almacenados y un promedio de múltiples valores Ntonal_lf almacenados, y una relación del promedio de los valores Ntonal_lf con respecto al promedio de los valores Ntonal se calcula para usarse como la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia.

En esta forma de realización, una señal de audio se clasifica según estadísticas a largo plazo de pendientes de energía residual de predicción lineal. Además, se tiene en cuenta tanto la robustez de la clasificación como la velocidad de reconocimiento de la clasificación; por lo tanto, el número de parámetros de clasificación es relativamente bajo, pero el resultado es relativamente preciso, la complejidad es baja y las sobrecargas de memoria son bajas.

Con referencia a la FIG. 6, otra forma de realización de un procedimiento de clasificación de señales de audio incluye:

S601: Llevar a cabo un procesamiento de división en tramas en una señal de audio de entrada.
S602: Obtener una fluctuación de espectro de frecuencia, un máximo de banda de alta frecuencia de espectro de frecuencia, un grado de correlación de espectro de frecuencia y una pendiente de energía residual de predicción lineal de una trama de audio actual.

La fluctuación de espectro de frecuencia, flux, denota una fluctuación de energía de corta duración o de larga duración de un espectro de frecuencia de una señal, y es un valor promedio de valores absolutos de diferencias de energía logarítmica entre frecuencias correspondientes de una trama de audio actual y una trama histórica en un espectro de banda baja y media, donde la trama histórica se refiere a cualquier trama anterior a la trama de audio actual. El máximo de banda de alta frecuencia de espectro de frecuencia, ph, denota un máximo o un pico de energía, en una banda de alta frecuencia, de un espectro de frecuencia de la trama de audio actual. El grado de correlación de espectro de frecuencia, cor_map_sum, denota estabilidad, entre tramas adyacentes, de una estructura armónica de señal. La pendiente de energía residual de predicción lineal epsP_tilt denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio de entrada a medida que aumenta el orden de predicción lineal. En lo que respecta a un procedimiento específico para calcular estos parámetros se hace referencia a la anterior forma de realización.

Además, puede obtenerse un parámetro de sonoridad; y el parámetro de sonoridad, *sonoridad*, denota un grado de correlación de dominio de tiempo entre la trama de audio actual y una señal antes de un periodo de tono. El parámetro de sonoridad, *sonoridad*, se obtiene mediante predicción lineal y análisis, representa un grado de correlación de dominio de tiempo entre la trama de audio actual y una señal antes de un periodo de tono y tiene un valor entre 0 y 1. Esto pertenece a la técnica anterior y, por lo tanto, no se describe en detalle en la presente invención. En esta forma de realización, la sonoridad se calcula para cada una de dos subtramas de la trama de audio actual, y las sonoridades se promedian para obtener un parámetro de sonoridad de la trama de audio actual. El parámetro de sonoridad de la trama de audio actual también se almacena en una memoria intermedia de datos históricos de sonoridad y, en esta forma de realización, la longitud de la memoria intermedia de datos históricos de sonoridad es 10.

S603: Almacenar la fluctuación de espectro de frecuencia, el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal en memorias correspondientes.

Opcionalmente, antes de almacenar estos parámetros, el procedimiento incluye además:

En una forma de realización se determina, según la actividad de voz de la trama de audio actual, si hay que almacenar la fluctuación de espectro de frecuencia en la memoria de fluctuaciones de espectro de frecuencia. Si la trama de audio actual es una trama activa, la fluctuación de espectro de frecuencia de la trama de audio actual se almacena en la memoria de fluctuaciones de espectro de frecuencia.

En otra forma de realización, se determina, según la actividad de voz de la trama de audio y si la trama de audio es un ataque de energía, si hay que almacenar la fluctuación de espectro de frecuencia en la memoria. Si la trama de audio actual es una trama activa y la trama de audio actual no pertenece a un ataque de energía, la fluctuación de espectro de frecuencia de la trama de audio actual se almacena en la memoria de fluctuaciones de espectro de frecuencia. En otra forma de realización, si la trama de audio actual es una trama activa y ninguna de múltiples tramas consecutivas, que incluyen la trama de audio actual y una trama histórica de la trama de audio actual, pertenece a un ataque de energía, la fluctuación de espectro de frecuencia de la trama de audio se almacena en la memoria de fluctuaciones de espectro de frecuencia; en caso contrario, la fluctuación de espectro de frecuencia no se almacena. Por ejemplo, si la trama de audio actual es una trama activa y ni una trama anterior de la trama de audio actual ni una segunda trama histórica de la trama de audio actual pertenece a un ataque de energía, la fluctuación de espectro de frecuencia de la trama de audio se almacena en la memoria de fluctuaciones de espectro de frecuencia; en caso contrario, la fluctuación de espectro de frecuencia no se almacena.

En lo que respecta a definiciones y maneras de obtener el indicador de actividad de voz, ind_vad, y el indicador de ataque de voz, ind_ataque, se hace referencia a la descripción de las anterior forma de realización.

Opcionalmente, antes de almacenar estos parámetros, el procedimiento incluye además:

determinar, según la actividad de voz de la trama de audio actual, si hay que almacenar en las memorias el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal; y si la trama de audio actual es una trama activa, almacenar los parámetros; en caso contrario, no almacenar los parámetros.

S604: Obtener estadísticas de datos eficaces de fluctuaciones de espectro de frecuencia almacenadas, estadísticas de datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, estadísticas de datos eficaces de grados de correlación de espectro de frecuencia almacenados, y estadísticas de datos eficaces de pendientes de energía residual de predicción lineal almacenadas, y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de los datos eficaces, donde las estadísticas de los datos eficaces se refieren a un valor de datos obtenido tras realizar una operación de cálculo en los datos eficaces almacenados en las memorias, donde la operación de cálculo puede incluir una operación para obtener un valor promedio, una operación para obtener una varianza o similares.

Opcionalmente, antes de la etapa S604, el procedimiento puede incluir además:

actualizar, dependiendo de si la trama de audio actual es música percutante, las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia. En una forma de realización, si la trama de audio actual es música percutante, los valores válidos de fluctuación de espectro de frecuencia de la memoria de fluctuaciones de espectro de frecuencia se modifican pasando a ser un valor inferior o igual a un umbral de música, donde si una fluctuación de espectro de frecuencia de una trama de audio es menor que el umbral de música, el audio se clasifica como una trama de música. En una forma de realización, si la trama de audio actual es música percutante, los valores válidos de fluctuación de espectro de frecuencia de la memoria de fluctuaciones de espectro de frecuencia se fijan a 5.

Opcionalmente, antes de la etapa S604, el procedimiento puede incluir además:

actualizar las fluctuaciones de espectro de frecuencia de la memoria según la actividad de una trama histórica de la trama de audio actual. En una forma de realización, si se determina que la fluctuación de espectro de frecuencia de la trama de audio actual se almacena en la memoria de fluctuaciones de espectro de frecuencia, y una trama de audio anterior es una trama inactiva, los datos de otras fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia, excepto la fluctuación de espectro de frecuencia de la trama de audio actual, se modifican pasando a ser datos eficaces. En otra forma de realización, si se determina que la fluctuación de espectro de frecuencia de la trama de audio actual se almacena en la memoria de fluctuaciones de espectro de frecuencia, y tres tramas consecutivas anteriores a la trama de audio actual no son todas ellas tramas activas, la fluctuación de espectro de frecuencia de la trama de audio actual se modifica pasando a ser un primer valor. El primer valor puede ser un umbral de voz, donde si la fluctuación de espectro de frecuencia de la trama de audio es mayor que el umbral de voz, el audio se clasifica como una trama de voz. En otra forma de realización, si se determina que la fluctuación de espectro de frecuencia de la trama de audio actual se almacena en la memoria de fluctuaciones de espectro de frecuencia, y el resultado de clasificación de una trama histórica es una trama de música y la fluctuación de espectro de frecuencia de la trama de audio actual es mayor que un segundo valor, la fluctuación de

espectro de frecuencia de la trama de audio actual se modifica pasando a ser el segundo valor, donde el segundo valor es mayor que el primer valor.

Por ejemplo, si una trama anterior de la trama de audio actual es una trama inactiva (ind_vad = 0), excepto la flux de trama de audio actual almacenada recientemente en la memoria intermedia de datos históricos de flux, todos los datos restantes de la memoria intermedia de datos históricos de flux se fijan a -1 (lo que equivale a que los datos se invaliden). Si tres tramas consecutivas anteriores a la trama de audio actual no son todas tramas activas (ind_vad = 1), la flux de trama de audio actual que acaba de almacenarse en la memoria intermedia de datos históricos de flux se modifica pasando a ser 16. Si las tres tramas consecutivas anteriores a la trama de audio actual son todas ellas tramas activas (ind_vad = 1), un resultado uniforme a largo plazo de un resultado de clasificación de señal histórica es una señal de música y la flux de trama de audio actual es mayor que 20, la fluctuación de espectro de frecuencia de la trama de audio actual almacenada pasa a ser 20. En lo que respecta al cálculo de la trama activa y del resultado uniforme a largo plazo del resultado de clasificación de señal histórica, se hace referencia a la forma de realización anterior.

En una forma de realización, la etapa S604 incluye:

obtener por separado un valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, un valor promedio de los datos eficaces de los máximos almacenados de banda de alta frecuencia de espectro de frecuencia, un valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia almacenados y una varianza de los datos eficaces de las pendientes de energía residual de predicción lineal almacenadas; y
cuando se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz; el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

Generalmente, un valor de fluctuación de espectro de frecuencia de una trama de música es relativamente pequeño, mientras que un valor de fluctuación de espectro de frecuencia de una trama de voz es relativamente grande; un valor pico de banda de alta frecuencia de espectro de frecuencia de una trama de música es relativamente grande, y un máximo de banda de alta frecuencia de espectro de frecuencia de una trama de voz es relativamente pequeño; un valor de grado de correlación de espectro de frecuencia de una trama de música es relativamente grande, y un valor de grado de correlación de espectro de frecuencia de una trama de voz es relativamente pequeño; un valor de pendiente de energía residual de predicción lineal de una trama de música es relativamente pequeño, y un valor de pendiente de energía residual de predicción lineal de una trama de voz es relativamente grande. Por lo tanto, la trama de audio actual puede clasificarse según las estadísticas de los parámetros anteriores. Evidentemente, la clasificación de señales también puede llevarse a cabo en la trama de audio actual usando otro procedimiento de clasificación. Por ejemplo, se cuenta una cantidad de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia; la memoria se divide, según la cantidad de datos eficaces, en al menos dos intervalos de diferente longitud desde un extremo cercano a un extremo remoto, se obtiene un valor promedio de datos eficaces de fluctuaciones de espectro de frecuencia correspondientes a cada intervalo, un valor promedio de datos eficaces de máximos de banda de alta frecuencia de espectro de frecuencia, un valor promedio de datos eficaces de grados de correlación de espectro de frecuencia y una varianza de datos eficaces de pendientes de energía residual de predicción lineal, donde un punto inicial de los intervalos se una ubicación de almacenamiento de la fluctuación de espectro de frecuencia de la trama actual, el extremo cercano es un extremo en el que se almacena la fluctuación de espectro de frecuencia de la trama actual, y el extremo remoto es un extremo en el que se almacena una fluctuación de espectro de frecuencia de una trama histórica; la trama de audio se clasifica según las estadísticas de los datos eficaces de los parámetros anteriores en un intervalo relativamente corto, y si las estadísticas de los parámetros en este intervalo son suficientes para distinguir un tipo de la trama de audio, el proceso de clasificación termina; en caso contrario, el proceso de clasificación continúa en el intervalo más corto de los intervalos restantes relativamente largos, y el resto puede deducirse por analogía. En un proceso de clasificación de cada intervalo, la trama de audio actual se clasifica según un umbral de clasificación correspondiente a cada intervalo, y cuando se satisface una de las siguientes condiciones, la trama de audio actual se clasifica como una trama de música; en caso contrario, la trama de audio actual se clasifica como una trama de voz: el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

Tras la clasificación de señales, diferentes señales pueden codificarse en diferentes modos de codificación. Por ejemplo, una señal de voz se codifica usando un codificador basado en un modelo de generación de voz (tal como

CELP), y una señal de música se codifica usando un codificador basado en conversión (tal como un codificador basado en MDCT).

En esta forma de realización, la clasificación se realiza según estadísticas a largo plazo de fluctuaciones de espectro de frecuencia, máximos de banda de alta frecuencia de espectro de frecuencia, grados de correlación de espectro de frecuencia y pendientes de energía residual de predicción lineal. Además, se tiene en cuenta tanto la robustez de la clasificación como la velocidad de reconocimiento de la clasificación; por lo tanto, el número de parámetros de clasificación es relativamente bajo, pero el resultado es relativamente preciso, la velocidad de reconocimiento es relativamente alta y la complejidad es relativamente baja.

En una forma de realización, después de que la fluctuación de espectro de frecuencia, flux, el máximo de banda de alta frecuencia de espectro de frecuencia, ph, el grado de correlación de espectro de frecuencia, cor_map_sum, y la pendiente de energía residual de predicción lineal, epsP_tilt, se almacenen en las memorias correspondientes, puede llevarse a cabo la clasificación según una cantidad de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas usando diferentes procesos de determinación. Si el indicador de actividad de voz está fijado a 1, es decir, la trama de audio actual es una trama de voz activa, se comprueba la cantidad N de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas.

Si cambia un valor de la cantidad N de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria, también cambia un proceso de determinación.

(1) Con referencia a la FIG. 7, si N = 60 se obtiene un valor promedio de todos los datos de la memoria intermedia de datos históricos de flux y se denota como flux60, se obtiene un valor promedio de 30 datos en un extremo cercano y se denota como flux30 y se obtiene un valor promedio de 10 datos en un extremo cercano y se denota como flux10. Se obtiene un valor promedio de todos los datos de la memoria intermedia de datos históricos de ph y se denota como ph60, se obtiene un valor promedio de 30 datos en un extremo cercano y se denota como ph30 y se obtiene un valor promedio de 10 datos en el extremo cercano y se denota como ph10. Se obtiene un valor promedio de todos los datos de la memoria intermedia de datos históricos de cor_map_sum y se denota como cor_map_sum60, se obtiene un valor promedio de 30 datos en un extremo cercano y se denota como cor_map_sum30 y se obtiene un valor promedio de 10 datos en el extremo cercano y se denota como cor_map_sum10. Además, se obtiene una varianza de todos los datos de la memoria intermedia de datos históricos de epsP_tilt y se denota como epsP_tilt60, se obtiene una varianza de 30 datos en un extremo cercano y se denota como epsP_tilt30 y se obtiene una varianza de 10 datos en el extremo cercano y se denota como epsP_tilt10. Se obtiene una cantidad cnt_sonoridad de datos cuyo valor es mayor que 0,9 en la memoria intermedia de datos históricos de sonoridad. El extremo cercano es un extremo en el que están almacenados los parámetros anteriores correspondientes a la trama de audio actual.
En primer lugar, se comprueba si flux10, ph10, epsP_tilt10, cor_map_sum10, y cnt_sonoridad satisfacen las siguientes condiciones: flux10 < 10 o epsPtilt10 < 0,0001 o ph10 > 1050 o cor_map_sum10 > 95, y cnt_sonoridad < 6. Si se satisfacen las condiciones, la trama de audio actual se clasifica como un tipo de música (es decir, Modo = 1). En caso contrario, se comprueba si flux10 es mayor que 15 y si cnt_sonoridad es mayor que 2, o si flux10 es mayor que 16. Si se satisfacen las condiciones, la trama de audio actual se clasifica como un tipo de voz (es decir, Modo = 0). En caso contrario, se comprueba si flux30, flux10, ph30, epsP_tilt30, cor_map_sum30 y cnt_sonoridad satisfacen las siguientes condiciones: flux30 < 13 y flux10 < 15, o epsPtilt30 < 0,001 o ph30 > 800 o cor_map_sum30 > 75. Si se satisfacen las condiciones, la trama de audio actual se clasifica como un tipo de música. En caso contrario, se comprueba si flux60, flux30, ph60, epsP_tilt60 y cor_map_sum60 satisfacen las siguientes condiciones: flux60 < 14,5 o cor_map_sum30 > 75 o ph60 > 770 o epsP_tilt10 < 0,002 y flux30 < 14. Si se satisfacen las condiciones, la trama de audio actual se clasifica como un tipo de música; en caso contrario, la trama de audio actual se clasifica como un tipo de voz.
(2) Con referencia a la FIG. 8, si N < 60 y N ≥ 30, un valor promedio de N datos en un extremo cercano de la memoria intermedia de datos históricos de flux, un valor promedio de N datos en un extremo cercano de la memoria intermedia de datos históricos de ph, y un valor promedio de N datos en un extremo cercano de la memoria intermedia de datos históricos de cor_map_sum se obtienen por separado y se denotan como fluxN, phN y cor_map_sumN. Además, se obtiene una varianza de N datos en un extremo cercano de la memoria intermedia de datos históricos de epsP_tilt y se denota como epsP_tiltN. Se comprueba si fluxN, phN, epsP_tiltN, y cor_map_sumN satisfacen la siguiente condición: fluxN < 13 + (N - 30)/20 o cor_map_sumN > 75 + (N - 30)/6 o phN > 800 o epsP_tiltN < 0,001. Si se satisface la condición, la trama de audio actual se clasifica como un tipo de música; en caso contrario, la trama de audio actual se clasifica como un tipo de voz.
(3) Con referencia a la FIG. 9, si N < 30 y N ≥ 10, un valor promedio de N datos en un extremo cercano de la memoria intermedia de datos históricos de flux, un valor promedio de N datos en un extremo cercano de la memoria intermedia de datos históricos de ph, y un valor promedio de N datos en un extremo cercano de la memoria intermedia de datos históricos de cor_map_sum se obtienen por separado y se denotan como fluxN, phN y cor_map_sumN. Además, se obtiene una varianza de N datos en un extremo cercano de la memoria intermedia de datos históricos de epsP_tilt y se denota como epsP_tiltN.
En primer lugar, se comprueba si un promedio móvil a largo plazo, mode_mov, de un resultado de clasificación de datos históricos es mayor que 0,8. Si es así, se comprueba si fluxN, phN, epsP_tiltN y cor_map_sumN satisfacen la siguiente condición: fluxN < 16 + (N - 10)/20 o phN > 1000 -12,5 x (N-10) o

epsP_tiltN < 0,0005 + 0,000045 x (N - 10) o cor_map_sumN > 90 - (N - 10). En caso contrario, se obtiene una cantidad cnt_sonoridad de datos cuyo valor es mayor que 0,9 en la memoria intermedia de datos históricos de sonoridad, y se comprueba si se satisfacen las siguientes condiciones: fluxN < 12 + (N -10)/20 o phN > 1050- 12,5 x (N - 10) o epsP_tiltN < 0,0001 + 0,000045 x (N - 10) o cor_map_sumN > 95 - (N - 10) y cnt_sonoridad < 6. Si se satisface cualquier grupo de los dos grupos de condiciones anteriores, la trama de audio actual se clasifica como un tipo de música; en caso contrario, la trama de audio actual se clasifica como un tipo de voz.

(4) Con referencia a la FIG. 10, si N < 10 y N ≥ 5, se obtiene un valor promedio de N datos en un extremo cercano de la memoria intermedia de datos históricos de ph y un valor promedio de N datos en un extremo cercano de la memoria intermedia de datos históricos de cor_map_sum, y se denotan como phN y cor_map_sumN, y se obtiene una varianza de N datos en un extremo cercano de la memoria intermedia de datos históricos de epsP_tilt y se denota como epsP_tiltN. Además, se obtiene una cantidad cnt6_sonoridad de datos cuyo valor es mayor que 0,9 entre seis datos en un extremo cercano de la memoria intermedia de datos históricos de sonoridad.

Se comprueba si se satisfacen las siguientes condiciones: epsP_tiltN < 0,00008 o phN > 1100 o cor_map_sumN > 100, y cnt_sonoridad < 4. Si se satisfacen las condiciones, la trama de audio actual se clasifica como un tipo de música; en caso contrario, la trama de audio actual se clasifica como un tipo de voz.

(5) Si N ≤ 5, un resultado de clasificación de una trama de audio anterior se usa como un tipo de clasificación de la trama de audio actual.

La anterior forma de realización es un proceso de clasificación específico en el que la clasificación se lleva a cabo según estadísticas a largo plazo de fluctuaciones de espectro de frecuencia, máximos de banda de alta frecuencia de espectro de frecuencia, grados de correlación de espectro de frecuencia y pendientes de energía residual de predicción lineal, y los expertos en la técnica pueden entender que la clasificación puede llevarse a cabo usando otro proceso. El proceso de clasificación en esta forma de realización puede aplicarse en etapas correspondientes de la anterior forma de realización para servir, por ejemplo, como un procedimiento de clasificación específico de la etapa 103 de la FIG. 2, la etapa 105 de la FIG. 4 o la etapa 604 de la FIG 6.

Con referencia a la FIG. 11, otra forma de realización de un procedimiento de clasificación de señales de audio incluye:

S1101: Llevar a cabo un procesamiento de división en tramas en una señal de audio de entrada.

S1102: Obtener una pendiente de energía residual de predicción lineal y una cantidad de tonos de espectro de frecuencia de una trama de audio actual y una relación de la cantidad de tonos de espectro de frecuencia en una banda de baja frecuencia.

La pendiente de energía residual de predicción lineal, epsP_tilt, denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio de entrada a medida que aumenta el orden de predicción lineal; la cantidad de tonos de espectro de frecuencia, Ntonal, denota una cantidad de celdas de frecuencia de la trama de audio actual que están en un banda de frecuencia entre 0 y 8 kHz y tienen valores pico de celda de frecuencia mayores que un valor predeterminado; la relación ratio_Ntonal_lf de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia denota una relación de una cantidad de tonos de banda de baja frecuencia con respecto a la cantidad de tonos de espectro de frecuencia. En lo que respecta a un cálculo específico, se hace referencia a la descripción de la anterior forma de realización.

S1103: Almacenar en memorias correspondientes la pendiente de energía residual de predicción lineal epsP_tilt, la cantidad de tonos de espectro de frecuencia y la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia.

La pendiente de energía residual de predicción lineal, epsP_tilt, y la cantidad de tonos de espectro de frecuencia de la trama de audio actual se almacenan en respectivas memorias intermedias de datos históricos y, en esta forma de realización, las longitudes de las dos memorias intermedias son también de 60.

Opcionalmente, antes de almacenar estos parámetros, el procedimiento incluye además: determinar, según la actividad de voz de la trama de audio actual, si hay que almacenar en las memorias la pendiente de energía residual de predicción lineal, la cantidad de tonos de espectro de frecuencia y la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia; y almacenar la pendiente de energía residual de predicción lineal en una memoria cuando se determina que es necesario almacenar la pendiente de energía residual de predicción lineal. Si la trama de audio actual es una trama activa, los parámetros se almacenan; en caso contrario, los parámetros no se almacenan.

S1104: Obtener por separado estadísticas de pendientes de energía residual de predicción lineal almacenadas y estadísticas de cantidades de tonos de espectro de frecuencia almacenadas, donde las estadísticas se refieren a un valor de datos obtenido tras realizar una operación de cálculo en datos almacenados en las memorias, donde la operación de cálculo puede incluir una operación para obtener un valor promedio, una operación para obtener una varianza, o similares.

En una forma de realización, obtener por separado estadísticas de las pendientes de energía residual de predicción lineal almacenadas y estadísticas de cantidades de tonos de espectro de frecuencia almacenadas incluye: obtener una varianza de las pendientes de energía residual de predicción lineal almacenadas; y obtener un valor promedio de las cantidades de tonos de espectro de frecuencia almacenadas.

S1105: Clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de las pendientes de energía residual de predicción lineal, las estadísticas de las cantidades de tonos de espectro de frecuencia y la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia.

En una forma de realización, esta etapa incluye:

> cuando la trama de audio actual es una trama activa y se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz;
> la varianza de las pendientes de energía residual de predicción lineal es menor que un quinto umbral; o
> el valor promedio de las cantidades de tonos de espectro de frecuencia es mayor que un sexto umbral; o
> la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia es menor que un séptimo umbral.

Generalmente, un valor de pendiente de energía residual de predicción lineal de una trama de música es relativamente pequeño, y un valor de pendiente de energía residual de predicción lineal de una trama de voz es relativamente grande; una cantidad de tonos de espectro de frecuencia de una trama de música es relativamente grande, y una cantidad de tonos de espectro de frecuencia de una trama de voz es relativamente pequeña; una relación de una cantidad de tonos de espectro de frecuencia de una trama de música en una banda de baja frecuencia es relativamente baja, y una relación de una cantidad de tonos de espectro de frecuencia de una trama de voz en la banda de baja frecuencia es relativamente alta (la energía de la trama de voz está concentrada principalmente en la banda de baja frecuencia). Por lo tanto, la trama de audio actual puede clasificarse según las estadísticas de los parámetros anteriores. Evidentemente, la clasificación de señales también puede llevarse a cabo en la trama de audio actual usando otro procedimiento de clasificación.

Tras la clasificación de señales, diferentes señales pueden codificarse en diferentes modos de codificación. Por ejemplo, una señal de voz se codifica usando un codificador basado en un modelo de generación de voz (tal como CELP), y una señal de música se codifica usando un codificador basado en conversión (tal como un codificador basado en MDCT).

En la anterior forma de realización, una señal de audio se clasifica según estadísticas a largo plazo de pendientes de energía residual de predicción lineal y cantidades de tonos de espectro de frecuencia y una relación de una cantidad de tonos de espectro de frecuencia en una banda de baja frecuencia; por lo tanto, hay un número relativamente bajo de parámetros, la tasa de reconocimiento es relativamente alta y la complejidad es relativamente baja.

En una forma de realización, después de almacenar en memorias intermedias correspondientes la pendiente de energía residual de predicción lineal, epsP_tilt, la cantidad de tonos de espectro de frecuencia, Ntonal, y la relación ratio_Ntonal_lf de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia, se obtiene una varianza de todos los datos de la memoria intermedia de datos históricos de epsP_tilt y se denota como epsP_tilt60. Se obtiene un valor promedio de todos los datos de la memoria intermedia de datos históricos de Ntonal y se denota como Ntonal60. Se obtiene un valor promedio de todos los datos de la memoria intermedia de datos históricos de Ntonal_lf, y se calcula una relación del valor promedio con respecto a Ntonal60 y se denota como ratio_Ntonal_lf60. Con referencia a la FIG. 12, una trama de audio actual se clasifica según la siguiente regla:

> Si un indicador de actividad de voz es 1 (es decir, ind_vad = 1), es decir, la trama de audio actual es una trama de voz activa, se comprueba si se satisface la siguiente condición: epsP_tilt60 < 0,002 o Ntonal60 > 18 o ratio_Ntonal_lf60 < 0,42; si se satisface la condición, la trama de audio actual se clasifica como un tipo de música (es decir, Modo = 1); en caso contrario, la trama de audio actual se clasifica como un tipo de voz (es decir, Modo = 0).

La anterior forma de realización es un proceso de clasificación específico en el que la clasificación se lleva a cabo según estadísticas de pendientes de energía residual de predicción lineal, estadísticas de cantidades de tonos de espectro de frecuencia y una relación de una cantidad de tonos de espectro de frecuencia en una banda de baja frecuencia, y los expertos en la técnica pueden entender que la clasificación puede llevarse a cabo usando otro proceso. El proceso de clasificación en esta forma de realización puede aplicarse en etapas correspondientes de la anterior forma de realización para servir, por ejemplo, como un procedimiento de clasificación específico de la etapa 504 de la FIG. 5 o la etapa 1105 de la FIG. 11.

La presente invención proporciona un procedimiento de selección de modo de codificación de audio que tiene una baja complejidad y bajas sobrecargas de memoria. Además, se tiene en cuenta tanto la robustez de la clasificación como la velocidad de reconocimiento de la clasificación.

En relación con la anterior forma de realización de procedimiento, la presente invención proporciona además un aparato de clasificación de señales de audio, y el aparato puede estar ubicado en un dispositivo terminal o un dispositivo de red. El aparato de clasificación de señales de audio puede llevar a cabo las etapas de la anterior forma de realización de procedimiento.

Con referencia a la FIG. 13, la presente invención proporciona una forma de realización de un aparato de clasificación de señales de audio, donde el aparato está configurado para clasificar una señal de audio de entrada, e incluye:

una unidad de determinación de almacenamiento 1301, configurada para determinar, según la actividad de voz de la trama de audio actual, si hay que obtener y almacenar una fluctuación de espectro de frecuencia de la trama de audio actual, donde la fluctuación de espectro de frecuencia denota una fluctuación de energía de un espectro de frecuencia de una señal de audio;
una memoria 1302, configurada para almacenar la fluctuación de espectro de frecuencia cuando la unidad de determinación de almacenamiento proporciona un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia;
una unidad de actualización 1304, configurada para actualizar, dependiendo de si una trama de voz es música percutante, las fluctuaciones de espectro de frecuencia almacenadas en la memoria; y
una unidad de clasificación 1303, configurada para clasificar la trama de audio actual como una trama de voz o una trama de música según estadísticas de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria; y cuando las estadísticas de datos eficaces de las fluctuaciones de espectro de frecuencia satisfacen una condición de clasificación de voz, clasificar la trama de audio actual como una trama de voz; o cuando las estadísticas de los datos eficaces de las fluctuaciones de espectro de frecuencia satisfacen una condición de clasificación de música, clasificar la trama de audio actual como una trama de música.

En una forma de realización, la unidad de determinación de almacenamiento 1301 está configurada específicamente para: cuando se determina que la trama de audio actual es una trama activa, proporcionar un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia de la trama de audio actual.

En otra forma de realización, la unidad de determinación de almacenamiento está configurada específicamente para: cuando se determina que la trama de audio actual es una trama activa y la trama de audio actual no pertenece a un ataque de energía, proporcionar un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia de la trama de audio actual.

En otra forma de realización, la unidad de determinación de almacenamiento está configurada específicamente para: cuando se determina que la trama de audio actual es una trama activa y ninguna de múltiples tramas consecutivas que incluyen la trama de audio actual y una trama histórica de la trama de audio actual pertenece a un ataque de energía, proporcionar un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia de la trama de audio actual.

Según la invención, la unidad de actualización está configurada específicamente para: si la trama de audio actual pertenece a música percutante, modificar los valores de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia.

En otra forma de realización, la unidad de actualización está configurada específicamente para: si la trama de audio actual es una trama activa, y una trama de audio anterior es una trama inactiva, modificar los datos de otras fluctuaciones de espectro de frecuencia almacenadas en la memoria excepto la fluctuación de espectro de frecuencia de la trama de audio actual pasando a ser datos ineficaces; o si la trama de audio actual es una trama activa, y tres tramas consecutivas anteriores a la trama de audio actual no son todas ellas tramas activas, modificar la fluctuación de espectro de frecuencia de la trama de audio actual pasando a ser un primer valor; o si la trama de audio actual es una trama activa y un resultado de clasificación de datos históricos es una señal de música y la fluctuación de espectro de frecuencia de la trama de audio actual es mayor que un segundo valor, modificar la fluctuación de espectro de frecuencia de la trama de audio actual pasando a ser el segundo valor, donde el segundo valor es mayor que el primer valor.

Con referencia a la FIG. 14, en una forma de realización, la unidad de clasificación 1303 incluye:

una unidad de cálculo 1401, configurada para obtener un valor promedio de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria; y
una unidad de determinación 1402, configurada para comparar el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia con una condición de clasificación de música; y cuando el valor

promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia satisface la condición de clasificación de música, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz.

Por ejemplo, cuando el valor promedio obtenido de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un umbral de clasificación de música, la trama de audio actual se clasifica como una trama de música; en caso contrario, la trama de audio actual se clasifica como una trama de voz.

En la forma de realización anterior, puesto que una señal de audio se clasifica según estadísticas a largo plazo de fluctuaciones de espectro de frecuencia, hay un número relativamente bajo de parámetros, la velocidad de reconocimiento es relativamente alta y la complejidad es relativamente baja. Además, las fluctuaciones de espectro de frecuencia se ajustan teniendo en cuenta factores tales como la actividad de voz y la música percutante; por lo tanto, la presente invención tiene una mayor velocidad de reconocimiento para una señal de música, y es adecuada para la clasificación de señales de audio híbridas.

En otra forma de realización, el aparato de clasificación de señales de audio incluye además:

una unidad de obtención de parámetros, configurada para obtener un máximo de banda de alta frecuencia de espectro de frecuencia, un grado de correlación de espectro de frecuencia y una pendiente de energía residual de predicción lineal de la trama de audio actual, donde el máximo de banda de alta frecuencia de espectro de frecuencia denota un máximo o un pico de energía, en una banda de alta frecuencia, de un espectro de frecuencia de la trama de audio actual; el grado de correlación de espectro de frecuencia denota estabilidad, entre tramas adyacentes, de una estructura armónica de señal de la trama de audio actual; y la pendiente de energía residual de predicción lineal denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio a medida que aumenta el orden de predicción lineal; donde
la unidad de determinación de almacenamiento está configurada además para determinar, según la actividad de voz de la trama de audio actual, si hay que almacenar el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal;
la unidad de almacenamiento está configurada además para: cuando la unidad de determinación de almacenamiento proporciona un resultado que indica que es necesario almacenar el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal, almacenar el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal; y
la unidad de clasificación está configurada específicamente para obtener estadísticas de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, estadísticas de datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, estadísticas de datos eficaces de grados de correlación de espectro de frecuencia almacenados, y estadísticas de datos eficaces de pendientes de energía residual de predicción lineal almacenadas, y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de los datos eficaces; y cuando las estadísticas de los datos eficaces de las fluctuaciones de espectro de frecuencia satisfacen una condición de clasificación de voz, clasificar la trama de audio actual como una trama de voz; o cuando las estadísticas de los datos eficaces de las fluctuaciones de espectro de frecuencia satisfacen una condición de clasificación de música, clasificar la trama de audio actual como una trama de música.

En una forma de realización, la unidad de clasificación incluye específicamente:

una unidad de cálculo, configurada para obtener por separado un valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, un valor promedio de los datos eficaces de los máximos almacenados de banda de alta frecuencia de espectro de frecuencia, un valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia almacenados y una varianza de los datos eficaces de las pendientes de energía residual de predicción lineal almacenadas; y
una unidad de determinación, configurada para: cuando se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz: el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

En la anterior forma de realización, una señal de audio se clasifica según estadísticas a largo plazo de fluctuaciones de espectro de frecuencia, máximos de banda de alta de frecuencia de espectro de frecuencia, grados de correlación de espectro de frecuencia y pendientes de energía residual de predicción lineal; por lo tanto, hay un número relativamente bajo de parámetros, la tasa de reconocimiento es relativamente alta y la complejidad es

relativamente baja. Además, las fluctuaciones de espectro de frecuencia se ajustan teniendo en cuenta factores tales como la actividad de voz y la música percutante, y las fluctuaciones de espectro de frecuencia se modifican según un entorno de señal en el que está ubicada la trama de audio actual; por lo tanto, la presente invención mejora la tasa de reconocimiento de clasificación y es adecuada para la clasificación de señales de audio híbridas.

Con referencia a la FIG. 15, se representa una forma de realización de un aparato de clasificación de señales de audio, donde el aparato está configurado para clasificar una señal de audio de entrada, e incluye:

una unidad de división en tramas 1501, configurada para llevar a cabo un procesamiento de división en tramas en una señal de audio de entrada;
una unidad de obtención de parámetros 1502, configurada para obtener una pendiente de energía residual de predicción lineal de una trama de audio actual, donde la pendiente de energía residual de predicción lineal denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio a medida que aumenta el orden de predicción lineal;
una unidad de almacenamiento 1503, configurada para almacenar la pendiente de energía residual de predicción lineal; y
una unidad de clasificación 1504, configurada para clasificar la trama de audio según estadísticas de una parte de datos de pendientes de energía residual de predicción en una memoria.

Con referencia a la FIG. 16, el aparato de clasificación de señales de audio incluye además:

una unidad de determinación de almacenamiento 1505, configurada para determinar, según la actividad de voz de la trama de audio actual, si hay que almacenar la pendiente de energía residual de predicción lineal en la memoria, donde
la unidad de almacenamiento 1503 está configurada específicamente para: cuando la unidad de determinación de almacenamiento determina que es necesario almacenar la pendiente de energía residual de predicción lineal, almacenar la pendiente de energía residual de predicción lineal en la memoria.

En una forma de realización, las estadísticas de la parte de los datos de las pendientes de energía residual de predicción es una varianza de la parte de los datos de las pendientes de energía residual de predicción; y
la unidad de clasificación está configurada específicamente para comparar la varianza de la parte de los datos de las pendientes de energía residual de predicción con un umbral de clasificación de música, y cuando la varianza de la parte de los datos de las pendientes de energía residual de predicción es menor que el umbral de clasificación de música, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz.

En otra forma de realización, la unidad de obtención de parámetros está configurada además para: obtener una fluctuación de espectro de frecuencia, un máximo de banda de alta frecuencia de espectro de frecuencia y un grado de correlación de espectro de frecuencia de la trama de audio actual, y almacenar en memorias correspondientes la fluctuación de espectro de frecuencia, el máximo de banda de alta frecuencia de espectro de frecuencia y el grado de correlación de espectro de frecuencia; y
la unidad de clasificación está configurada específicamente para obtener estadísticas de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, estadísticas de datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, estadísticas de datos eficaces de grados de correlación de espectro de frecuencia almacenados, y estadísticas de datos eficaces de pendientes de energía residual de predicción lineal almacenadas, y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de los datos eficaces, donde las estadísticas de los datos eficaces se refieren a un valor de datos obtenido tras realizar una operación de cálculo en los datos eficaces almacenados en las memorias.

Con referencia a la FIG. 17, específicamente, en una forma de realización, la unidad de clasificación 1504 incluye:

una unidad de cálculo 1701, configurada para obtener por separado un valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, un valor promedio de los datos eficaces de los máximos almacenados de banda de alta frecuencia de espectro de frecuencia, un valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia almacenados y una varianza de los datos eficaces de pendientes de energía residual de predicción lineal almacenadas; y
una unidad de determinación 1702, configurada para: cuando se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz: el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

En otra forma de realización, la unidad de obtención de parámetros está configurada además para obtener una cantidad de tonos de espectro de frecuencia de la trama de audio actual y una relación de la cantidad de tonos de espectro de frecuencia en una banda de baja frecuencia, y almacenar en memorias la cantidad de tonos de espectro de frecuencia y la relación de la cantidad de tono de espectro de frecuencia en la banda de baja frecuencia; y la unidad de clasificación está configurada específicamente para obtener por separado estadísticas de las pendientes de energía residual de predicción lineal almacenadas y estadísticas de cantidades de tonos de espectro de frecuencia almacenadas; y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de las pendientes de energía residual de predicción lineal, las estadísticas de las cantidades de tono de espectro de frecuencia y la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia, donde las estadísticas de los datos eficaces se refieren a un valor de datos obtenido tras realizar una operación de cálculo en datos almacenados en las memorias.

Específicamente, la unidad de clasificación incluye:

una unidad de cálculo, configurada para obtener una varianza de datos eficaces de las pendientes de energía residual de predicción lineal almacenadas y un valor promedio de las cantidades de tonos de espectro de frecuencia almacenadas; y
una unidad de determinación, configurada para: cuando la trama de audio actual es una trama activa y se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz: la varianza de las pendientes de energía residual de predicción lineal es menor que un quinto umbral; o el valor promedio de las cantidades de tonos de espectro de frecuencia es mayor que un sexto umbral; o la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia es menor que un séptimo umbral.

Específicamente, la unidad de obtención de parámetros obtiene la pendiente de energía residual de predicción lineal de la trama de audio actual según la siguiente fórmula:

$$epsP\_tilt = \frac{\sum_{i=1}^{n} epsP(i) \cdot epsP(i+1)}{\sum_{i}^{n} epsP(i) \cdot epsP(i)}$$

donde epsP(i) denota la energía residual de predicción lineal de orden i-ésimo de la trama de audio actual; y n es un entero positivo, denota un orden de predicción lineal y es inferior o igual a un orden máximo de predicción lineal.

Específicamente, la unidad de obtención de parámetros está configurada para contar una cantidad de celdas de frecuencia de la trama de audio actual que están en una banda de frecuencia entre 0 y 8 kHz y tienen valores pico de celda de frecuencia mayores que un valor predeterminado, para usar la cantidad como la cantidad de tonos de espectro de frecuencia; y la unidad de obtención de parámetros está configurada para calcular una relación de una cantidad de celdas de frecuencia de la trama de audio actual que están en una banda de frecuencia entre 0 y 4 kHz y tienen valores pico de celda de frecuencia mayores que el valor predeterminado con respecto a la cantidad de las celdas de frecuencia de la trama de audio actual que están en la banda de frecuencia entre 0 y 8 kHz y tienen valores pico de celda de frecuencia mayores que el valor predeterminado, para usar la relación como la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia.

En esta forma de realización, una señal de audio se clasifica según estadísticas a largo plazo de pendientes de energía residual de predicción lineal. Además, se tiene en cuenta tanto la robustez de la clasificación como la velocidad de reconocimiento de la clasificación; por lo tanto, el número de parámetros de clasificación es relativamente bajo, pero el resultado es relativamente preciso, la complejidad es baja y las sobrecargas de memoria son bajas.

La presente descripción proporciona otro ejemplo de un aparato de clasificación de señales de audio, donde el aparato está configurado para clasificar una señal de audio de entrada, e incluye:

una unidad de división en tramas, configurada para llevar a cabo un procesamiento de división en tramas en una señal de audio de entrada;
una unidad de obtención de parámetros, configurada para obtener una fluctuación de espectro de frecuencia, un máximo de banda de alta frecuencia de espectro de frecuencia, un grado de correlación de espectro de frecuencia y una pendiente de energía residual de predicción lineal de una trama de audio actual, donde la fluctuación de espectro de frecuencia denota una fluctuación de energía de un espectro de frecuencia de la señal de audio; el máximo de banda de alta frecuencia de espectro de frecuencia denota un máximo o un pico de energía, en una banda de alta frecuencia, de un espectro de frecuencia de la trama de audio actual; el grado de correlación de espectro de frecuencia denota estabilidad, entre tramas adyacentes, de una estructura armónica de señal de la trama de audio actual, y la pendiente de energía residual de predicción

lineal denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio a medida que aumenta un orden de predicción lineal;
una unidad de almacenamiento, configurada para almacenar la fluctuación de espectro de frecuencia, el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal; y
una unidad de clasificación, configurada para obtener estadísticas de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, estadísticas de datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, estadísticas de datos eficaces de grados de correlación de espectro de frecuencia almacenados, y estadísticas de datos eficaces de pendientes de energía residual de predicción lineal almacenadas, y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de los datos eficaces, donde las estadísticas de los datos eficaces se refieren a un valor de datos obtenido tras realizar una operación de cálculo en los datos eficaces almacenados en las memorias, donde la operación de cálculo puede incluir una operación para obtener un valor promedio, una operación para obtener una varianza o similares.

En una forma de realización, el aparato de clasificación de señales de audio puede incluir además:

una unidad de determinación de almacenamiento, configurada para determinar, según la actividad de voz de la trama de audio actual, si hay que almacenar la fluctuación de espectro de frecuencia, el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal de la trama de audio actual; y
la unidad de almacenamiento está configurada específicamente para: cuando la unidad de determinación de almacenamiento proporciona un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia, el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal, almacenar la fluctuación de espectro de frecuencia, el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal.

Específicamente, en una forma de realización, la unidad de determinación de almacenamiento determina, según la actividad de voz de la trama de audio actual, si hay que almacenar la fluctuación de espectro de frecuencia en la memoria de fluctuaciones de espectro de frecuencia. Si la trama de audio actual es una trama activa, la unidad de determinación de almacenamiento proporciona un resultado que indica que es necesario almacenar el parámetro; en caso contrario, la unidad de determinación de almacenamiento proporciona un resultado que indica que no es necesario almacenar el parámetro. En otra forma de realización, la unidad de determinación de almacenamiento determina, según la actividad de voz de la trama de audio y si la trama de audio es un ataque de energía, si hay que almacenar la fluctuación de espectro de frecuencia en la memoria. Si la trama de audio actual es una trama activa y la trama de audio actual no pertenece a un ataque de energía, la fluctuación de espectro de frecuencia de la trama de audio actual se almacena en la memoria de fluctuaciones de espectro de frecuencia. En otra forma de realización, si la trama de audio actual es una trama activa y ninguna de múltiples tramas consecutivas, que incluyen la trama de audio actual y una trama histórica de la trama de audio actual, pertenece a un ataque de energía, la fluctuación de espectro de frecuencia de la trama de audio se almacena en la memoria de fluctuaciones de espectro de frecuencia; en caso contrario, la fluctuación de espectro de frecuencia no se almacena. Por ejemplo, si la trama de audio actual es una trama activa y ni una trama anterior de la trama de audio actual ni una segunda trama histórica de la trama de audio actual pertenece a un ataque de energía, la fluctuación de espectro de frecuencia de la trama de audio se almacena en la memoria de fluctuaciones de espectro de frecuencia; en caso contrario, la fluctuación de espectro de frecuencia no se almacena.

En una forma de realización, la unidad de clasificación incluye:

una unidad de cálculo, configurada para obtener por separado un valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, un valor promedio de los datos eficaces de los máximos almacenados de banda de alta frecuencia de espectro de frecuencia, un valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia almacenados y una varianza de los datos eficaces de las pendientes de energía residual de predicción lineal almacenadas; y
una unidad de determinación, configurada para: cuando se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz: el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

En lo que respecta a una manera específica de calcular la fluctuación de espectro de frecuencia, el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente

de energía residual de predicción lineal de la trama de audio actual, se hace referencia a la anterior forma de realización de procedimiento.

Además, el aparato de clasificación de señales de audio puede incluir además:

una unidad de actualización, configurada para actualizar, dependiendo de si una trama de voz es música percutante o actividad de una trama de audio histórica, las fluctuaciones de espectro de frecuencia almacenadas en la memoria. En una forma de realización según la invención, la unidad de actualización está configurada específicamente para: si la trama de audio actual pertenece a música percutante, modificar los valores de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia. En otra forma de realización, la unidad de actualización está configurada específicamente para: si la trama de audio actual es una trama activa, y una trama de audio anterior es una trama inactiva, modificar los datos de otras fluctuaciones de espectro de frecuencia almacenadas en la memoria, excepto la fluctuación de espectro de frecuencia de la trama de audio actual, pasando a ser datos ineficaces; o si la trama de audio actual es una trama activa, y tres tramas consecutivas anteriores a la trama de audio actual no son todas ellas tramas activas, modificar la fluctuación de espectro de frecuencia de la trama de audio actual pasando a ser un primer valor; o si la trama de audio actual es una trama activa y un resultado de clasificación de datos históricos es una señal de música y la fluctuación de espectro de frecuencia de la trama de audio actual es mayor que un segundo valor, modificar la fluctuación de espectro de frecuencia de la trama de audio actual pasando a ser el segundo valor, donde el segundo valor es mayor que el primer valor.

En esta forma de realización, la clasificación se realiza según estadísticas a largo plazo de fluctuaciones de espectro de frecuencia, máximos de banda de alta frecuencia de espectro de frecuencia, grados de correlación de espectro de frecuencia y pendientes de energía residual de predicción lineal. Además, se tiene en cuenta tanto la robustez de la clasificación como la velocidad de reconocimiento de la clasificación; por lo tanto, el número de parámetros de clasificación es relativamente bajo, pero el resultado es relativamente preciso, la velocidad de reconocimiento es relativamente alta y la complejidad es relativamente baja.

La presente descripción proporciona otra forma de realización de un aparato de clasificación de señales de audio, donde el aparato está configurado para clasificar una señal de audio de entrada, e incluye:

una unidad de división en tramas, configurada para llevar a cabo un procesamiento de división en tramas en una señal de audio de entrada;
una unidad de obtención de parámetros, configurada para obtener una pendiente de energía residual de predicción lineal y una cantidad de tonos de espectro de frecuencia de una trama de audio actual y una relación de la cantidad de tonos de espectro de frecuencia en una banda de baja frecuencia, donde la pendiente de energía residual de predicción lineal, epsP_tilt, denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio de entrada a medida que aumenta el orden de predicción lineal; la cantidad de tonos de espectro de frecuencia, Ntonal, denota una cantidad de celdas de frecuencia de la trama de audio actual que están en un banda de frecuencia entre 0 y 8 kHz y tienen valores pico de celda de frecuencia mayores que un valor predeterminado; y la relación ratio_Ntonal_lf de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia denota una relación de una cantidad de tonos de banda de baja frecuencia con respecto a la cantidad de tonos de espectro de frecuencia; en lo que respecta a un cálculo específico, se hace referencia a la descripción de la anterior forma de realización;
una unidad de almacenamiento, configurada para almacenar la pendiente de energía residual de predicción lineal, la cantidad de tonos de espectro de frecuencia y la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia; y
una unidad de clasificación, configurada para obtener por separado estadísticas de las pendientes de energía residual de predicción lineal almacenadas y estadísticas de cantidades de tonos de espectro de frecuencia almacenadas; y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de las pendientes de energía residual de predicción lineal, las estadísticas de las cantidades de tonos de espectro de frecuencia y la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia, donde las estadísticas de los datos eficaces se refieren a un valor de datos obtenido tras realizar una operación de cálculo en datos almacenados en las memorias.

Específicamente, la unidad de clasificación incluye:

una unidad de cálculo, configurada para obtener una varianza de datos eficaces de las pendientes de energía residual de predicción lineal almacenadas y un valor promedio de las cantidades de tonos de espectro de frecuencia almacenadas; y
una unidad de determinación, configurada para: cuando la trama de audio actual es una trama activa y se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz: la varianza de las pendientes de energía residual de predicción lineal es menor que un quinto umbral; o el valor promedio de las cantidades de

tono de espectro de frecuencia es mayor que un sexto umbral; o la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia es menor que un séptimo umbral.

Específicamente, la unidad de obtención de parámetros obtiene la pendiente de energía residual de predicción lineal de la trama de audio actual según la siguiente fórmula:

$$epsP\_tilt = \frac{\sum_{i=1}^{n} epsP(i) \cdot epsP(i+1)}{\sum_{i=1}^{n} epsP(i) \cdot epsP(i)}$$

donde epsP(i) denota la energía residual de predicción lineal de orden i-ésimo de la trama de audio actual; y n es un entero positivo, denota un orden de predicción lineal y es inferior o igual a un orden máximo de predicción lineal.

Específicamente, la unidad de obtención de parámetros está configurada para contar una cantidad de celdas de frecuencia de la trama de audio actual que están en una banda de frecuencia entre 0 y 8 kHz y tienen valores pico de celda de frecuencia mayores que un valor predeterminado, para usar la cantidad como la cantidad de tonos de espectro de frecuencia; y la unidad de obtención de parámetros está configurada para calcular una relación de una cantidad de celdas de frecuencia de la trama de audio actual que están en una banda de frecuencia entre 0 y 4 kHz y tienen valores pico de celda de frecuencia mayores que el valor predeterminado con respecto a la cantidad de las celdas de frecuencia de la trama de audio actual que están en la banda de frecuencia entre 0 y 8 kHz y tienen valores pico de celda de frecuencia mayores que el valor predeterminado, para usar la relación como la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia.

En la anterior forma de realización, una señal de audio se clasifica según estadísticas a largo plazo de pendientes de energía residual de predicción lineal y cantidades de tonos de espectro de frecuencia y una relación de una cantidad de tonos de espectro de frecuencia en una banda de baja frecuencia; por lo tanto, hay un número relativamente bajo de parámetros, la tasa de reconocimiento es relativamente alta y la complejidad es relativamente baja.

El anterior aparato de clasificación de señales de audio puede conectarse a diferentes codificadores y codificar diferentes señales usando los diferentes codificadores. Por ejemplo, el aparato de clasificación de señales de audio está conectado a dos codificadores, codifica una señal de voz usando un codificador basado en un modelo de generación de voz (tal como CELP), y codifica una señal de música usando un codificador basado en conversión (tal como un codificador basado en MDCT). En lo que respecta a una definición y un procedimiento de obtención de cada parámetro específico de la anterior forma de realización de aparato, se hace referencia a la descripción relacionada de la forma de realización de procedimiento.

En relación con la anterior forma de realización de procedimiento, la presente invención proporciona además un aparato de clasificación de señales de audio, y el aparato puede estar ubicado en un dispositivo terminal o un dispositivo de red. El aparato de clasificación de señales de audio puede implementarse mediante un circuito de hardware o implementarse mediante software en combinación con hardware. Por ejemplo, con referencia a la FIG. 18, un procesador invoca un aparato de clasificación de señales de audio para implementar la clasificación en una señal de audio. El aparato de clasificación de señales de audio puede llevar a cabo los diversos procedimientos y procesos de la anterior forma de realización de procedimiento. En lo que respecta a módulos y funciones específicos del aparato de clasificación de señales de audio, se hace referencia a la descripción relacionada de la anterior forma de realización de aparato.

Un ejemplo de un dispositivo 1900 en la FIG. 19 es un codificador. El dispositivo 1900 incluye un procesador 1910 y una memoria 1920.

La memoria 1920 puede incluir una memoria aleatoria, una memoria flash, una memoria de solo lectura, una memoria de solo lectura programable, una memoria no volátil, un registro o similar. El procesador 1910 puede ser una unidad de procesamiento central (CPU).

La memoria 1920 está configurada para almacenar una instrucción ejecutable. El procesador 1910 puede ejecutar la instrucción ejecutable almacenada en la memoria 1920 y está configurado para:

En lo que respecta a otras funciones y operaciones del dispositivo 1900 se hace referencia a los procesos de las formas de realización de procedimiento de las FIG. 3 a 12, los cuales no se describen de nuevo en el presente documento por simplicidad.

Los expertos en la técnica pueden entender que todos o algunos de los procesos de los procedimientos de las formas de realización pueden implementarse mediante un programa informático que da instrucciones a hardware relacionado. El programa puede almacenarse en un medio de almacenamiento legible por ordenador. Cuando el programa se ejecuta se llevan a cabo los procesos de los procedimientos de las formas de realización El medio de

almacenamiento anterior puede incluir: un disco magnético, un disco óptico, una memoria de solo lectura (ROM) o una memoria de acceso aleatorio (RAM).

En las diversas formas de realización proporcionadas en la presente solicitud, debe entenderse que el sistema, el aparato y el procedimiento dados a conocer pueden implementarse de otra manera. Por ejemplo, la forma de realización de aparato descrita se proporciona simplemente a modo de ejemplo. Por ejemplo, la división en unidades es simplemente una división en funciones lógicas y puede ser otra división en una implementación real. Por ejemplo, una pluralidad de unidades o componentes pueden combinarse o integrarse en otro sistema, o algunas características pueden ignorarse o no llevarse a cabo. Además, los acoplamientos mutuos o acoplamientos directos o conexiones de comunicación ilustrados o descritos pueden implementarse usando algunas interfaces. Los acoplamientos indirectos o conexiones de comunicación entre los aparatos o unidades pueden implementarse de manera electrónica, mecánica o de otro modo.

Las unidades descritas como partes separadas pueden estar, o no, físicamente separadas, y las partes mostradas como unidades pueden ser, o no, unidades físicas, pueden estar ubicadas en una posición o pueden estar distribuidas en una pluralidad de unidades de red. Algunas o todas las unidades pueden seleccionarse según las necesidades reales para conseguir los objetivos de las soluciones de las formas de realización.

Además, las unidades funcionales de las formas de realización de la presente invención pueden estar integradas en una unidad de procesamiento, o cada una de las unidades pueden ser físicamente independientes, o dos o más unidades están integradas en una unidad.

Lo que antecede muestra simplemente formas de realización a modo de ejemplo, y no todas ellas pertenecen a la presente invención. Los expertos en la técnica pueden realizar diversas modificaciones y variaciones en la presente invención siempre que estén dentro del alcance definido por las reivindicaciones adjuntas.

REIVINDICACIONES

1. Un procedimiento de clasificación de señales de audio, que comprende:

determinar (101), según actividad de voz de una trama de audio actual, si hay que obtener una fluctuación de espectro de frecuencia de la trama de audio actual y almacenar la fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia, donde la fluctuación de espectro de frecuencia denota una fluctuación de energía de un espectro de frecuencia de una señal de audio;
actualizar (102), dependiendo de si la trama de audio es música percutante, las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia; y
clasificar (103) la trama de audio actual como una trama de voz o una trama de música según las estadísticas de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia.

2. El procedimiento según la reivindicación 1, en el que determinar, según la actividad de voz de una trama de audio actual, si hay que obtener una fluctuación de espectro de frecuencia de la trama de audio actual y almacenar la fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia comprende:

si la trama de audio actual es una trama activa, almacenar la fluctuación de espectro de frecuencia de la trama de audio actual en la memoria de fluctuaciones de espectro de frecuencia.

3. El procedimiento según la reivindicación 1, en el que determinar, según la actividad de voz de una trama de audio actual, si hay que obtener una fluctuación de espectro de frecuencia de la trama de audio actual y almacenar la fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia comprende:

si la trama de audio actual es una trama activa y la trama de audio actual no pertenece a un ataque de energía, almacenar la fluctuación de espectro de frecuencia de la trama de audio actual en la memoria de fluctuaciones de espectro de frecuencia.

4. El procedimiento según la reivindicación 1, en el que determinar, según la actividad de voz de una trama de audio actual, si hay que obtener una fluctuación de espectro de frecuencia de la trama de audio actual y almacenar la fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia comprende:

si la trama de audio actual es una trama activa y ninguna de múltiples tramas consecutivas que comprenden la trama de audio actual y una trama histórica de la trama de audio actual pertenece a un ataque de energía, almacenar la fluctuación de espectro de frecuencia de la trama de audio actual en la memoria de fluctuaciones de espectro de frecuencia.

5. El procedimiento según una cualquiera de las reivindicaciones 1 a 4, en el que actualizar, dependiendo de si la trama de actual es música percutante, las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia comprende:

si la trama de audio actual pertenece a música percutante, modificar los valores de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia.

6. El procedimiento según una cualquiera de las reivindicaciones 1 a 5, en el que clasificar la trama de audio actual como una trama de voz o una trama de música según las estadísticas de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia comprende:

obtener un valor promedio de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia; y
cuando el valor promedio obtenido de los datos eficaces de las fluctuaciones de espectro de frecuencia satisface una condición de clasificación de música, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz.

7. El procedimiento según una cualquiera de las reivindicaciones 1 a 5, que comprende además:

obtener un máximo de banda de alta frecuencia de espectro de frecuencia, un grado de correlación de espectro de frecuencia y una pendiente de energía residual de predicción lineal de la trama de audio actual, donde el máximo de banda de alta frecuencia de espectro de frecuencia denota un máximo o un pico de energía, en una banda de alta frecuencia, de un espectro de frecuencia de la trama de audio actual; el grado de correlación de espectro de frecuencia denota estabilidad, entre tramas adyacentes, de una estructura armónica de señal de la trama de audio actual, y la pendiente de energía residual de predicción lineal denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio a medida que aumenta un orden de predicción lineal; y

determinar, según la actividad de voz de la trama de audio actual, si hay que almacenar en una memoria el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal,
donde clasificar la trama de audio según las estadísticas de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia comprende:
obtener por separado un valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, un valor promedio de datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, un valor promedio de datos eficaces de grados de correlación de espectro de frecuencia almacenados y una varianza de datos eficaces de pendientes de energía residual de predicción lineal almacenadas; y
cuando se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz: el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

8. Un aparato de clasificación de señales de audio, donde el aparato está configurado para clasificar una señal de audio de entrada, y comprende:

una unidad de determinación de almacenamiento (1301), configurada para determinar, según la actividad de voz de una trama de audio actual, si hay que obtener y almacenar una fluctuación de espectro de frecuencia de la trama de audio actual, donde la fluctuación de espectro de frecuencia denota una fluctuación de energía de un espectro de frecuencia de una señal de audio;
una memoria (1302), configurada para almacenar la fluctuación de espectro de frecuencia cuando la unidad de determinación de almacenamiento proporciona un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia;
una unidad de actualización (1304), configurada para actualizar, dependiendo de si la trama de audio es música percutante, las fluctuaciones de espectro de frecuencia almacenadas en la memoria; y
una unidad de clasificación (1303), configurada para clasificar la trama de audio actual como una trama de voz o una trama de música según las estadísticas de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria.

9. El aparato según la reivindicación 8, en el que la unidad de determinación de almacenamiento está configurada específicamente para:

cuando se determina que la trama de audio actual es una trama activa, proporcionar un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia de la trama de audio actual.

10. El aparato según la reivindicación 8, en el que la unidad de determinación de almacenamiento está configurada específicamente para:

cuando se determina que la trama de audio actual es una trama activa y la trama de audio actual no pertenece a un ataque de energía, proporcionar un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia de la trama de audio actual.

11. El aparato según la reivindicación 8, en el que la unidad de determinación de almacenamiento está configurada específicamente para:

cuando se determina que la trama de audio actual es una trama activa y ninguna de múltiples tramas consecutivas, que comprenden la trama de audio actual y una trama histórica de la trama de audio actual, pertenece a un ataque de energía, proporcionar un resultado que indica que es necesario almacenar la fluctuación de espectro de frecuencia de la trama de audio actual.

12. El aparato según una cualquiera de las reivindicaciones 8 a 11, en el que la unidad de actualización está configurada específicamente para: si la trama de audio actual pertenece a música percutante, modificar los valores de las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia.

13. El aparato según una cualquiera de las reivindicaciones 8 a 12, en el que la unidad de clasificación comprende:

una unidad de cálculo, configurada para obtener un valor promedio de una parte de o todos los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas en la memoria; y

una unidad de determinación, configurada para comparar el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia con una condición de clasificación de música; y cuando el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia satisface la condición de clasificación de música, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz.

14. El aparato según una cualquiera de las reivindicaciones 8 a 12, que comprende además:

una unidad de obtención de parámetros, configurada para obtener un máximo de banda de alta frecuencia de espectro de frecuencia, un grado de correlación de espectro de frecuencia, un parámetro de sonoridad y una pendiente de energía residual de predicción lineal de la trama de audio actual, donde el máximo de banda de alta frecuencia de espectro de frecuencia denota un máximo o un pico de energía, en una banda de alta frecuencia, de un espectro de frecuencia de la trama de audio actual; el grado de correlación de espectro de frecuencia denota estabilidad, entre tramas adyacentes, de una estructura armónica de señal de la trama de audio actual; el parámetro de sonoridad denota un grado de correlación de dominio de tiempo entre la trama de audio actual y una señal antes de un periodo de tono; y la pendiente de energía residual de predicción lineal denota hasta qué punto cambia la energía residual de predicción lineal de la señal de audio a medida que aumenta un orden de predicción lineal; donde
la unidad de determinación de almacenamiento está configurada además para determinar, según la actividad de voz de la trama de audio actual, si hay que almacenar en memorias el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal;
la memoria está configurada además para: cuando la unidad de determinación de almacenamiento proporciona un resultado que indica que es necesario almacenar el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal, almacenar el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal; y
la unidad de clasificación está configurada específicamente para obtener estadísticas de datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, estadísticas de datos eficaces de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, estadísticas de datos eficaces de grados de correlación de espectro de frecuencia almacenados, y estadísticas de datos eficaces de pendientes de energía residual de predicción lineal almacenadas, y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de los datos eficaces.

15. El aparato según la reivindicación 14, en el que la unidad de clasificación comprende:

una unidad de cálculo, configurada para obtener por separado un valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia almacenadas, un valor promedio de los datos eficaces de los máximos almacenados de banda de alta frecuencia de espectro de frecuencia, un valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia almacenados y una varianza de los datos eficaces de las pendientes de energía residual de predicción lineal almacenadas; y
una unidad de determinación, configurada para: cuando se satisface una de las siguientes condiciones, clasificar la trama de audio actual como una trama de música; en caso contrario, clasificar la trama de audio actual como una trama de voz: el valor promedio de los datos eficaces de las fluctuaciones de espectro de frecuencia es menor que un primer umbral; o el valor promedio de los datos eficaces de los máximos de banda de alta frecuencia de espectro de frecuencia es mayor que un segundo umbral; o el valor promedio de los datos eficaces de los grados de correlación de espectro de frecuencia es mayor que un tercer umbral; o la varianza de los datos eficaces de las pendientes de energía residual de predicción lineal es menor que un cuarto umbral.

	Enésima trama anterior	...	Segunda trama anterior	Trama anterior	Trama actual	

FIG. 1

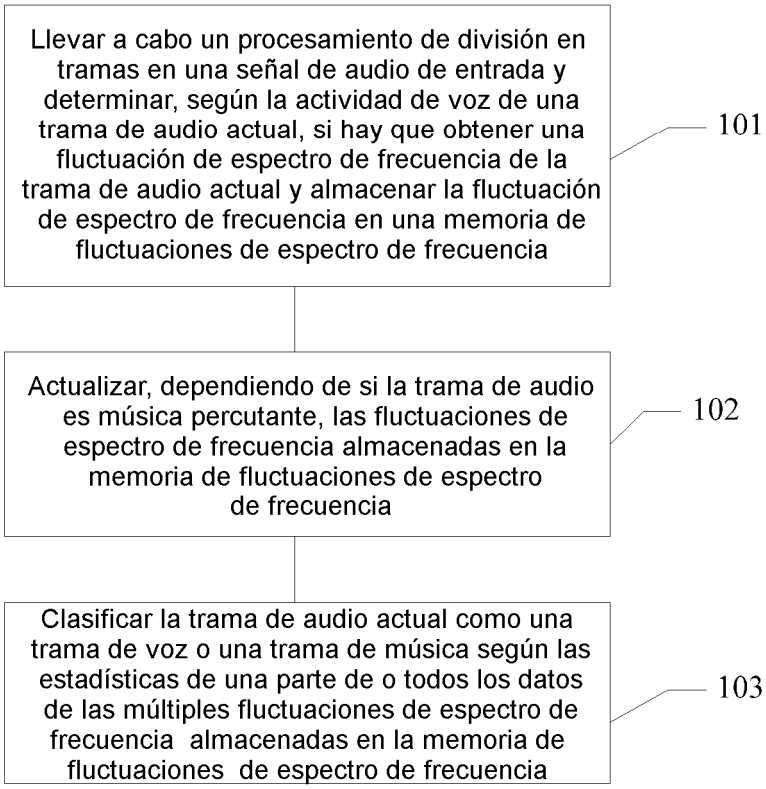

FIG. 2

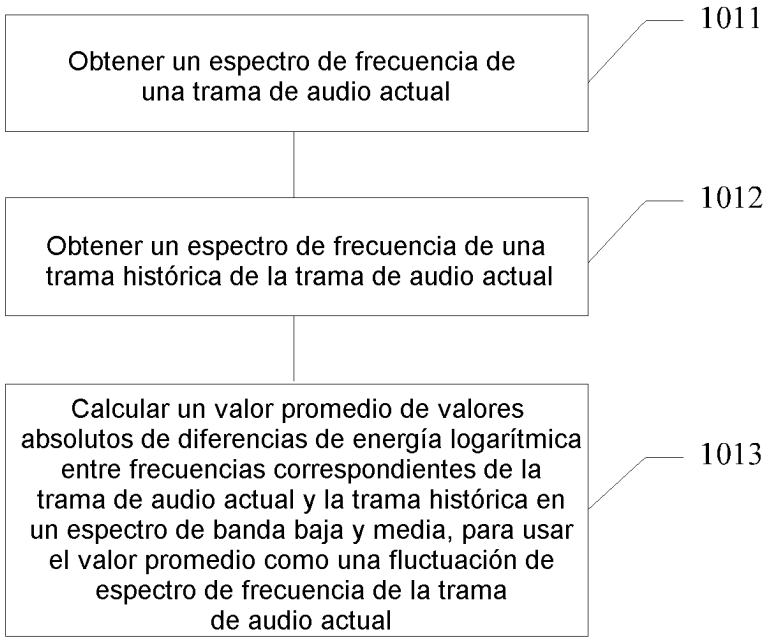
Obtener un espectro de frecuencia de una trama de audio actual
1011
Obtener un espectro de frecuencia de una trama histórica de la trama de audio actual
1012
Calcular un valor promedio de valores absolutos de diferencias de energía logarítmica entre frecuencias correspondientes de la trama de audio actual y la trama histórica en un espectro de banda baja y media, para usar el valor promedio como una fluctuación de espectro de frecuencia de la trama de audio actual
1013

FIG. 3

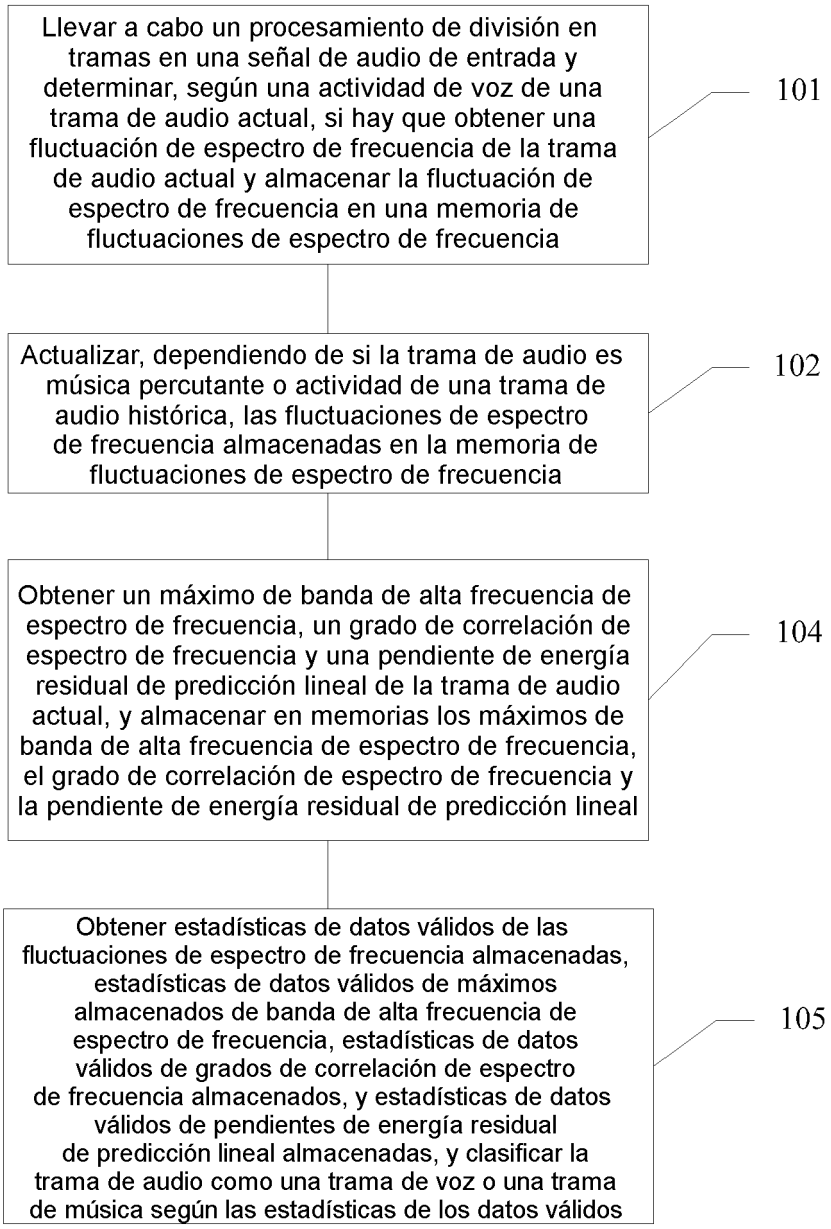
Llevar a cabo un procesamiento de división en tramas en una señal de audio de entrada y determinar, según una actividad de voz de una trama de audio actual, si hay que obtener una fluctuación de espectro de frecuencia de la trama de audio actual y almacenar la fluctuación de espectro de frecuencia en una memoria de fluctuaciones de espectro de frecuencia
101
Actualizar, dependiendo de si la trama de audio es música percutante o actividad de una trama de audio histórica, las fluctuaciones de espectro de frecuencia almacenadas en la memoria de fluctuaciones de espectro de frecuencia
102
Obtener un máximo de banda de alta frecuencia de espectro de frecuencia, un grado de correlación de espectro de frecuencia y una pendiente de energía residual de predicción lineal de la trama de audio actual, y almacenar en memorias los máximos de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal
104
Obtener estadísticas de datos válidos de las fluctuaciones de espectro de frecuencia almacenadas, estadísticas de datos válidos de máximos almacenados de banda de alta frecuencia de espectro de frecuencia, estadísticas de datos válidos de grados de correlación de espectro de frecuencia almacenados, y estadísticas de datos válidos de pendientes de energía residual de predicción lineal almacenadas, y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de los datos válidos
105

FIG. 4

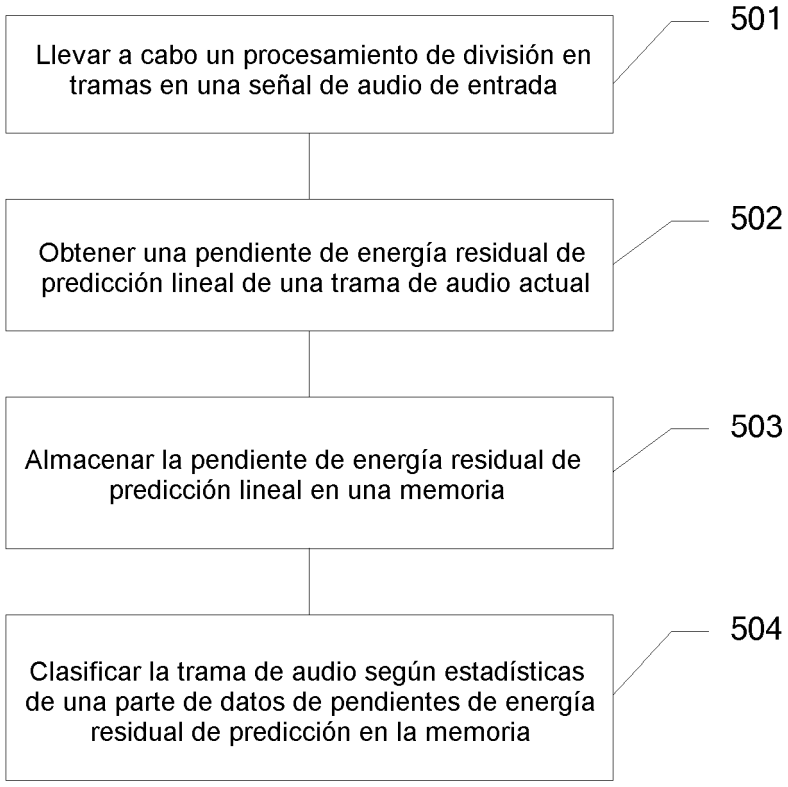
Llevar a cabo un procesamiento de división en tramas en una señal de audio de entrada
501
Obtener una pendiente de energía residual de predicción lineal de una trama de audio actual
502
Almacenar la pendiente de energía residual de predicción lineal en una memoria
503
Clasificar la trama de audio según estadísticas de una parte de datos de pendientes de energía residual de predicción en la memoria
504

FIG. 5

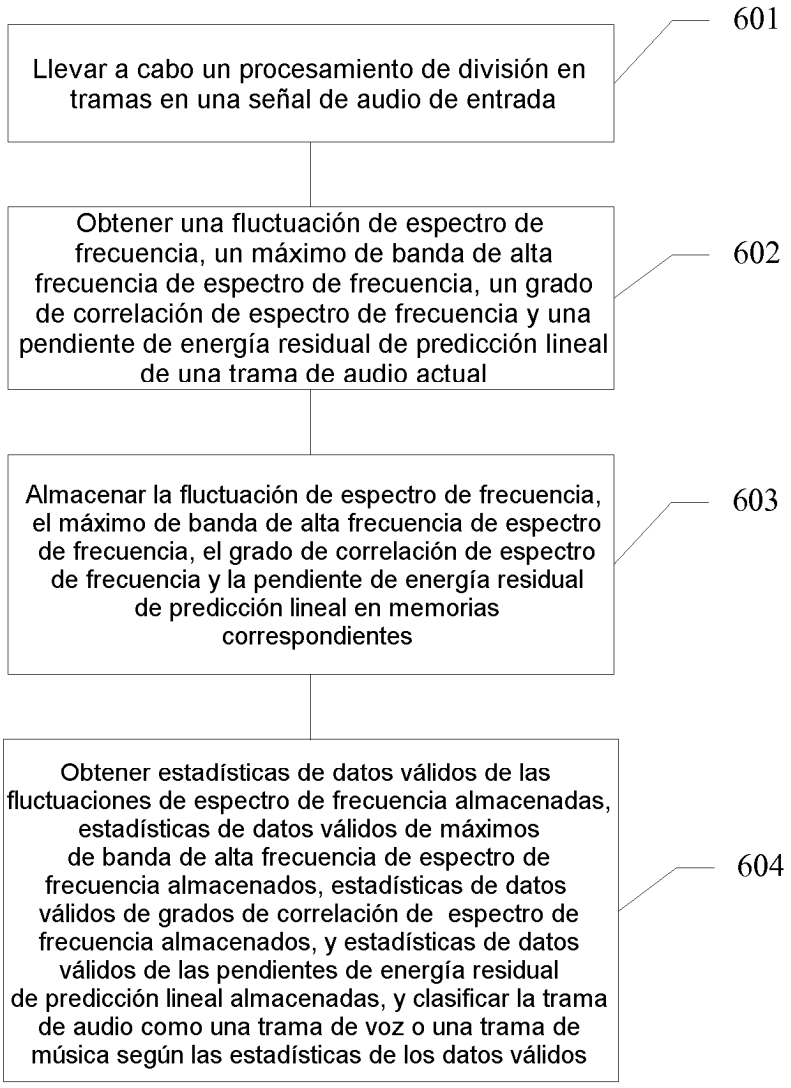
601
Llevar a cabo un procesamiento de división en tramas en una señal de audio de entrada
602
Obtener una fluctuación de espectro de frecuencia, un máximo de banda de alta frecuencia de espectro de frecuencia, un grado de correlación de espectro de frecuencia y una pendiente de energía residual de predicción lineal de una trama de audio actual
603
Almacenar la fluctuación de espectro de frecuencia, el máximo de banda de alta frecuencia de espectro de frecuencia, el grado de correlación de espectro de frecuencia y la pendiente de energía residual de predicción lineal en memorias correspondientes
604
Obtener estadísticas de datos válidos de las fluctuaciones de espectro de frecuencia almacenadas, estadísticas de datos válidos de máximos de banda de alta frecuencia de espectro de frecuencia almacenados, estadísticas de datos válidos de grados de correlación de espectro de frecuencia almacenados, y estadísticas de datos válidos de las pendientes de energía residual de predicción lineal almacenadas, y clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de los datos válidos

FIG. 6

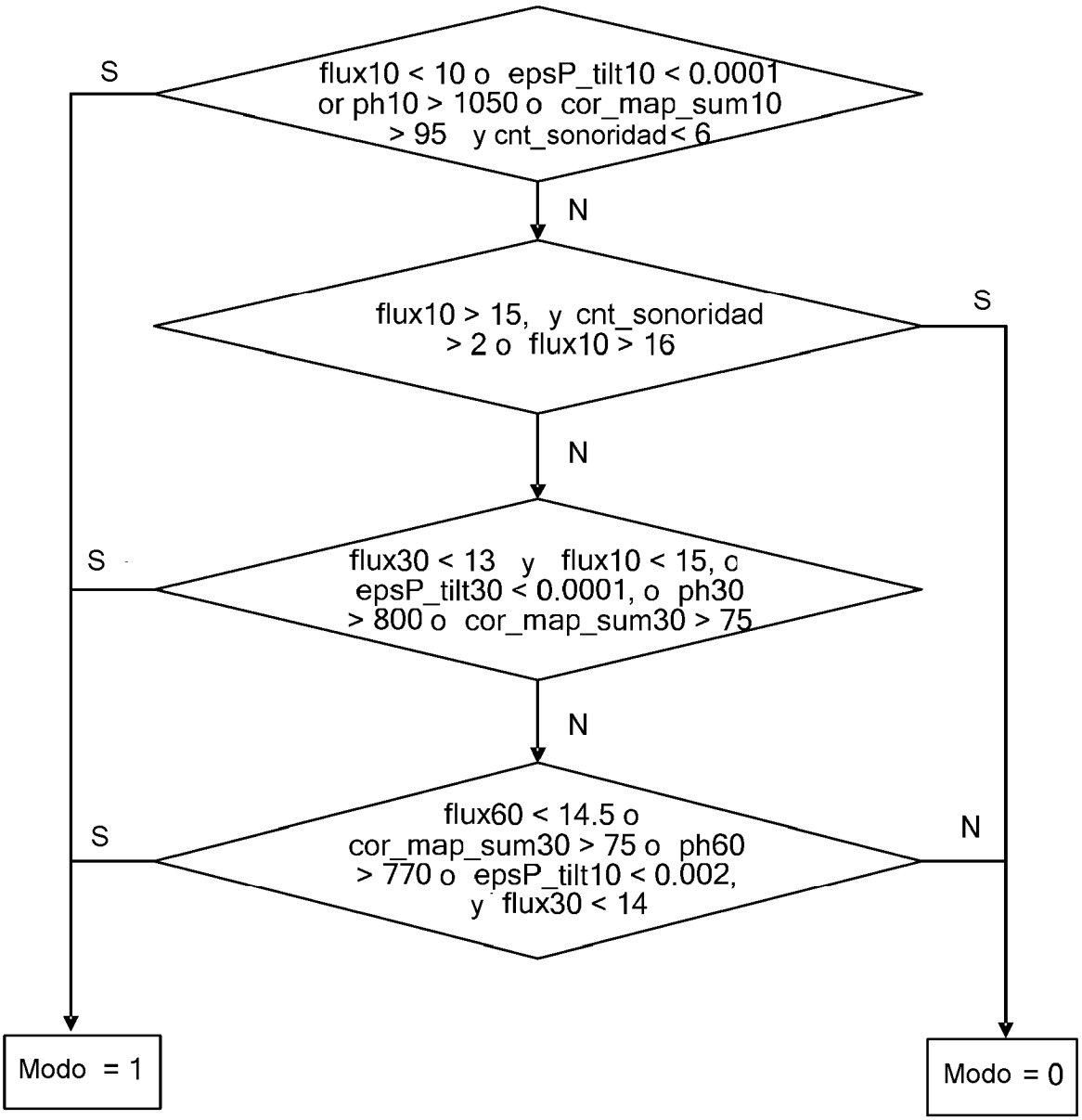
flux10 < 10 o epsP_tilt10 < 0.0001 or ph10 > 1050 o cor_map_sum10 > 95 y cnt_sonoridad < 6
S
N
flux10 > 15, y cnt_sonoridad > 2 o flux10 > 16
S
N
flux30 < 13 y flux10 < 15, o epsP_tilt30 < 0.0001, o ph30 > 800 o cor_map_sum30 > 75
S
N
flux60 < 14.5 o cor_map_sum30 > 75 o ph60 > 770 o epsP_tilt10 < 0.002, y flux30 < 14
S
N
Modo = 1
Modo = 0

FIG. 7

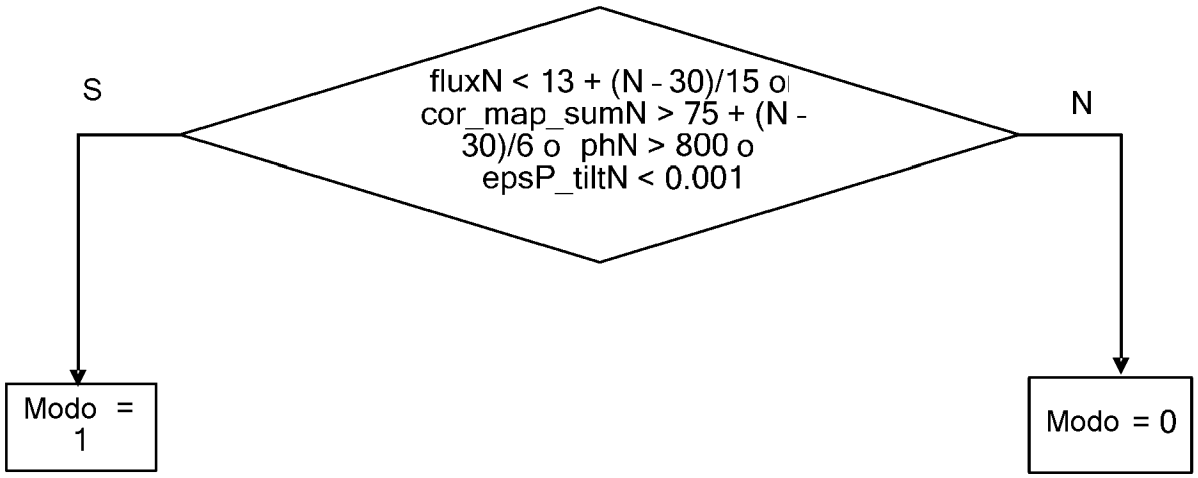
S
fluxN < 13 + (N - 30)/15 o
cor_map_sumN > 75 + (N -
30)/6 o phN > 800 o
epsP_tiltN < 0.001
N
Modo = 1
Modo = 0

FIG. 8

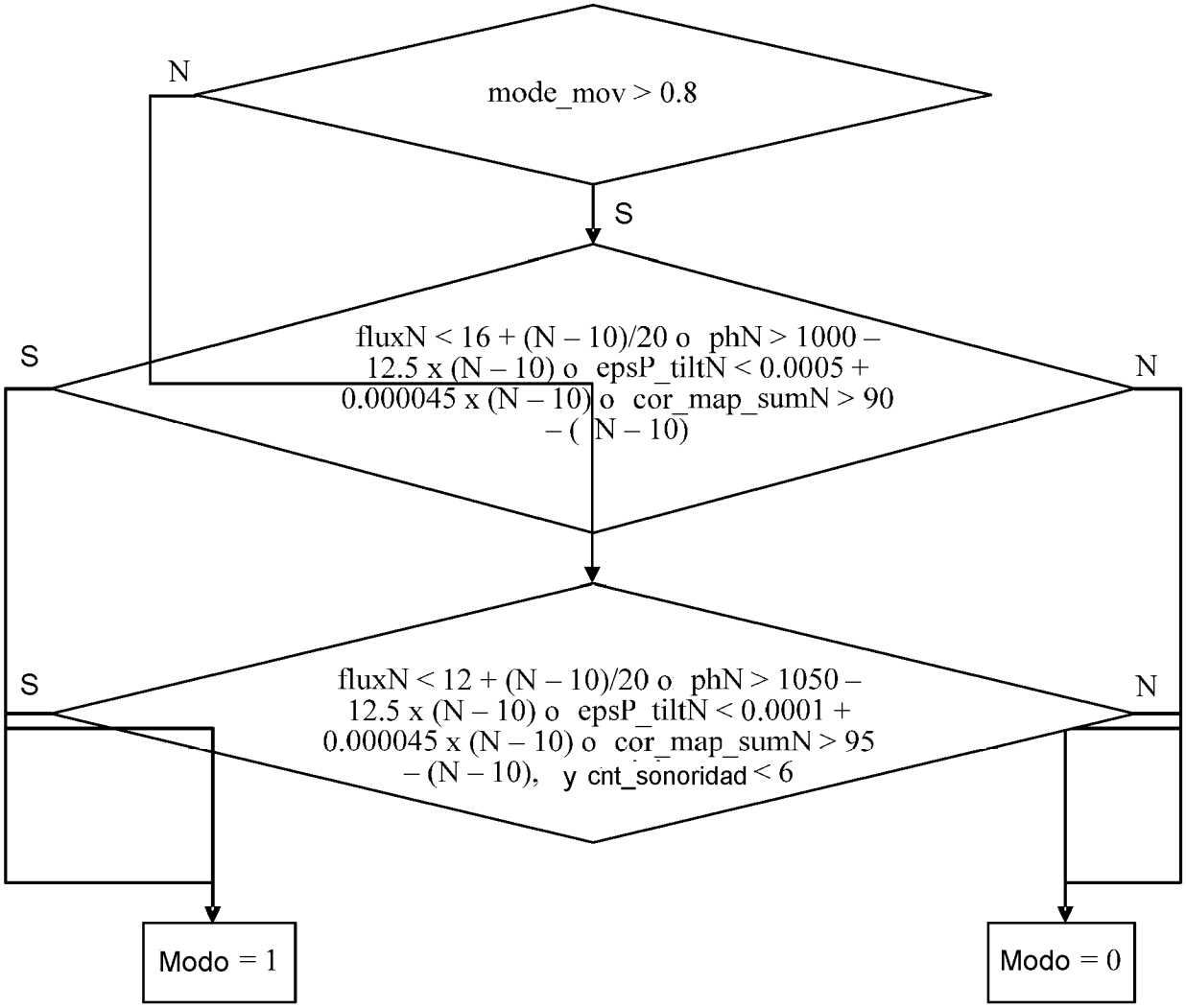
N
mode_mov > 0.8
S
S
fluxN < 16 + (N – 10)/20 o phN > 1000 –
12.5 x (N – 10) o epsP_tiltN < 0.0005 +
0.000045 x (N – 10) o cor_map_sumN > 90
– (N – 10)
N
S
fluxN < 12 + (N – 10)/20 o phN > 1050 –
12.5 x (N – 10) o epsP_tiltN < 0.0001 +
0.000045 x (N – 10) o cor_map_sumN > 95
– (N – 10), y cnt_sonoridad < 6
N
Modo = 1
Modo = 0

FIG. 9

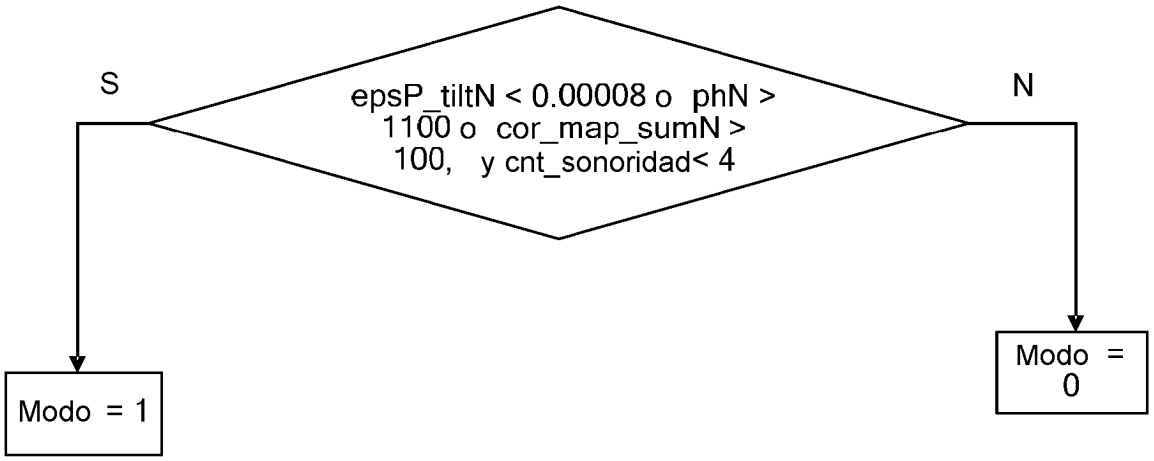
S
epsP_tiltN < 0.00008 o phN > 1100 o cor_map_sumN > 100, y cnt_sonoridad< 4
N
Modo = 1
Modo = 0

FIG. 10

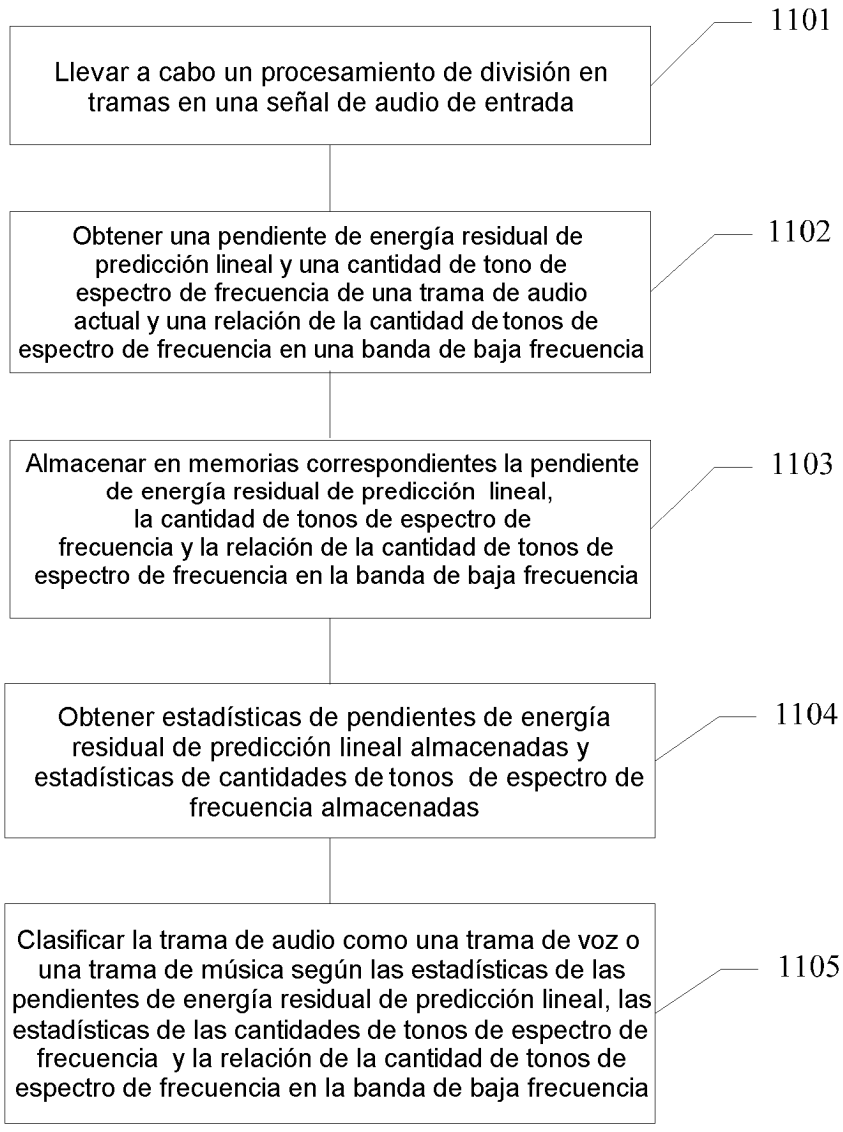
1101
Llevar a cabo un procesamiento de división en tramas en una señal de audio de entrada
1102
Obtener una pendiente de energía residual de predicción lineal y una cantidad de tono de espectro de frecuencia de una trama de audio actual y una relación de la cantidad de tonos de espectro de frecuencia en una banda de baja frecuencia
1103
Almacenar en memorias correspondientes la pendiente de energía residual de predicción lineal, la cantidad de tonos de espectro de frecuencia y la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia
1104
Obtener estadísticas de pendientes de energía residual de predicción lineal almacenadas y estadísticas de cantidades de tonos de espectro de frecuencia almacenadas
1105
Clasificar la trama de audio como una trama de voz o una trama de música según las estadísticas de las pendientes de energía residual de predicción lineal, las estadísticas de las cantidades de tonos de espectro de frecuencia y la relación de la cantidad de tonos de espectro de frecuencia en la banda de baja frecuencia

FIG. 11

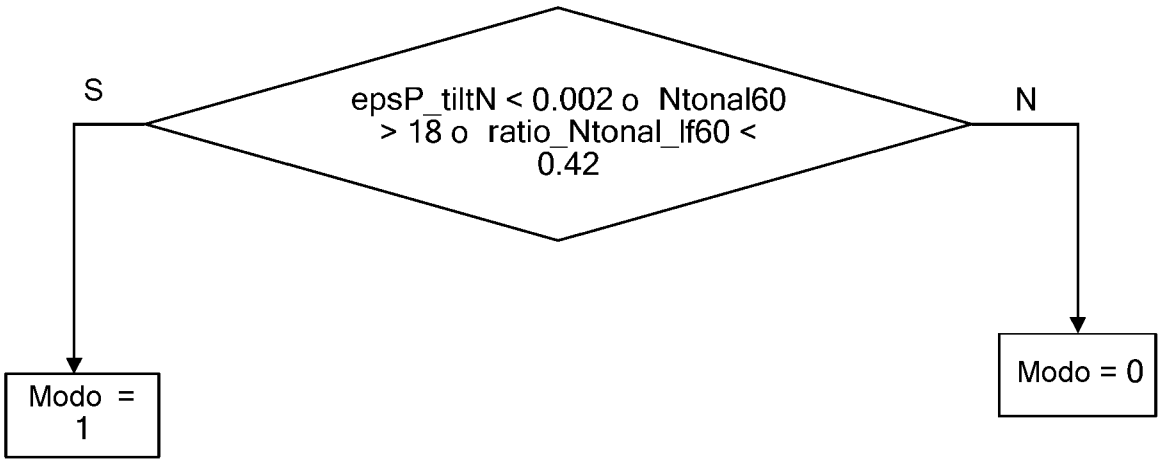

FIG. 12

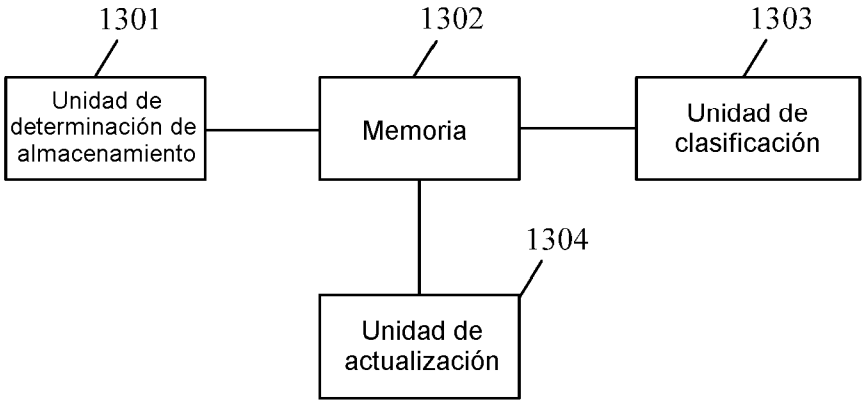

FIG. 13

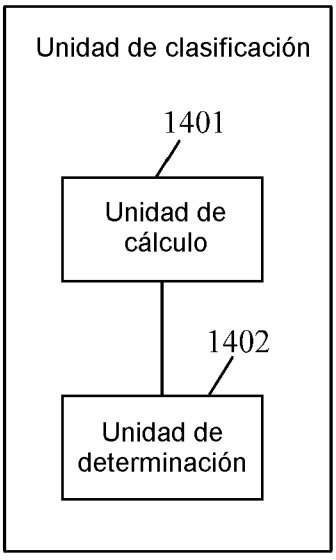

FIG. 14

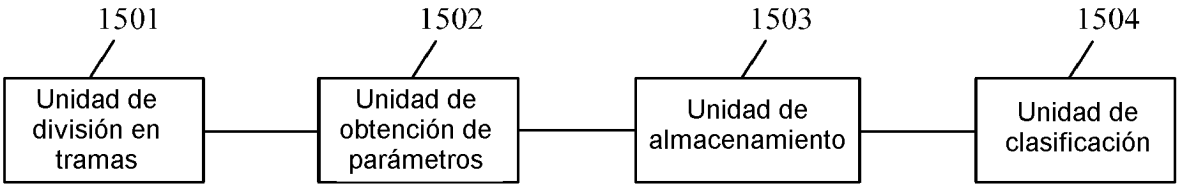

FIG. 15

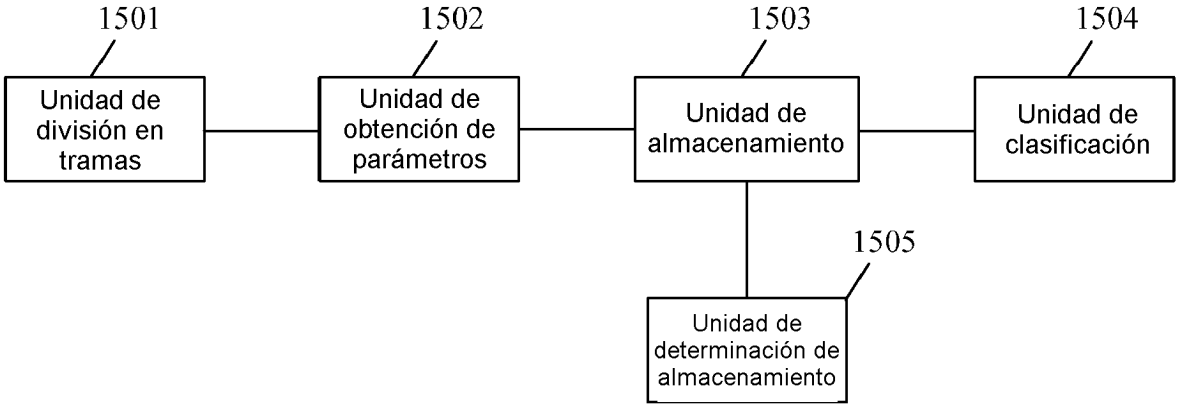

FIG. 16

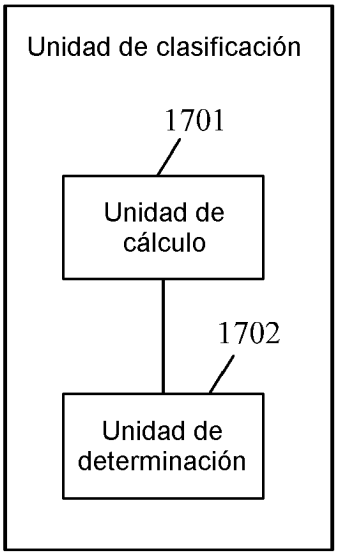

FIG. 17

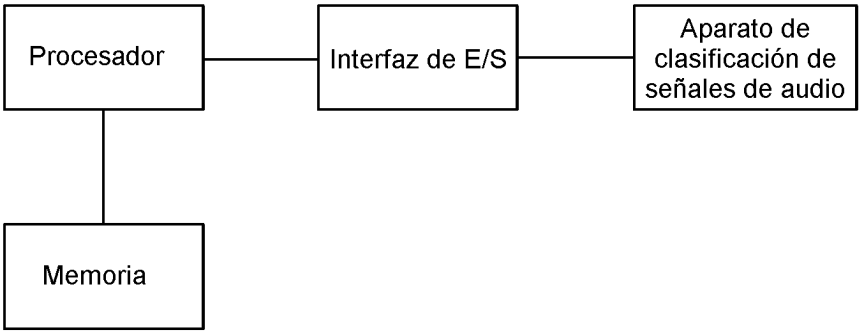
Procesador
Interfaz de E/S
Aparato de clasificación de señales de audio
Memoria

FIG. 18

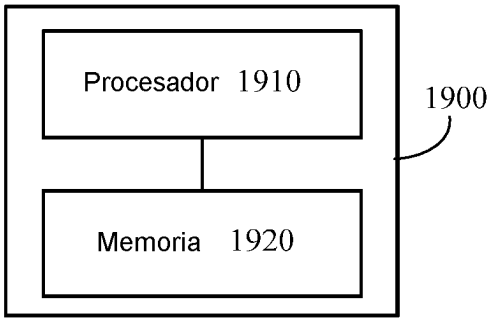
Procesador 1910
1900
Memoria 1920

FIG. 19