(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 684 297**

(51) Int. CI.:

G10L 19/20 (2013.01)
G10L 19/22 (2013.01)
G10L 25/00 (2013.01)
G10L 25/51 (2013.01)
G10L 25/78 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **16.06.2009 PCT/EP2009/004339**

(87) Fecha y número de publicación internacional: **14.01.2010 WO10003521**

(96) Fecha de presentación y número de la solicitud europea: **16.06.2009 E 09776747 (9)**

(97) Fecha y número de publicación de la concesión europea: **25.07.2018 EP 2301011**

(54) Título: **Método y discriminador para clasificar diferentes segmentos de una señal de audio que comprende segmentos de voz y música**

(30) Prioridad:

11.07.2008 US 79875

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
02.10.2018

(73) Titular/es:

FRAUNHOFER-GESELLSCHAFT ZUR FÖRDERUNG DER ANGEWANDTEN FORSCHUNG E.V. (100.0%)
Hansastrasse 27c
80686 München, DE

(72) Inventor/es:

FUCHS, GUILLAUME;
BAYER, STEFAN;
NAGEL, FREDERIK;
HERRE, JÜRGEN;
RETTELBACH, NIKOLAUS;
WABNIK, STEFAN;
YOKOTANI, YOSHIKAZU;
HIRSCHFELD, JENS y
LECOMTE, JÉRÉMIE

(74) Agente/Representante:

ARIZTI ACHA, Monica

Método y discriminador para clasificar diferentes segmentos de una señal de audio que comprende segmentos de voz y música

DESCRIPCIÓN

Antecedentes de la invención

La invención se refiere a un enfoque para la clasificación de diferentes segmentos de una señal que comprende segmentos de al menos un primer tipo y de un segundo tipo. Las realizaciones de la invención se refieren al campo de la codificación de audio y, en particular, a la discriminación de voz/música al codificar una señal de audio.

En la técnica son conocidos los esquemas de codificación en el dominio de la frecuencia tales como MP3 o AAC. Estos codificadores del dominio de la frecuencia se basan en una conversión del dominio del tiempo/dominio de la frecuencia, una etapa de cuantificación posterior, en la cual se controla el error de cuantificación usando la información de un módulo psicoacústico y una etapa de codificación, en la cual se codifican por entropía los coeficientes espectrales cuantificados y la correspondiente información secundaria usando unas tablas de códigos.

Por otro lado, existen unos codificadores que son muy adecuados para procesar la voz tales como el AMR-WB+ como se describe en el documento 3GPP TS 26.290. Tales esquemas de codificación de voz llevan a cabo un filtrado de predicción lineal de una señal del dominio del tiempo. Tal filtrado PL se deriva de un análisis de predicción lineal de la señal de entrada del dominio del tiempo. A continuación se codifican los coeficientes de filtrado PL resultantes y se transmiten como información secundaria. El proceso se conoce como codificación de predicción lineal (LPC). En la salida del filtro, la señal de predicción residual o la señal de error de predicción, que se conoce también como la señal de excitación, se codifica usando las etapas de análisis por síntesis del codificador ACELP o, de manera alternativa, se codifica usando un codificador de transformada que utiliza una transformada de Fourier con una superposición. La decisión entre la codificación por ACELP y la codificación por excitación codificada por transformada, que se llama también codificación XCT, se lleva a cabo usando un algoritmo de bucle cerrado o de bucle abierto.

Los esquemas de codificación de audio en el dominio de la frecuencia tales como el esquema de codificación de AAC de alta eficiencia, que combina un esquema de codificación de AAC y una técnica de replicación de ancho de banda espectral, se puede combinar también con una herramienta de codificación de estéreo conjunto o de canales múltiples, la cual se conoce bajo el término “MPEG envolvente”. Los esquemas de codificación en el dominio de la frecuencia son ventajosos por el hecho de que a bajas tasas de bits muestran una alta calidad para señales de música. Sin embargo, las bajas tasas de bits son problemáticas para la calidad de señales de voz.

Por otro lado, los codificadores de voz tales como el AMR-WB+ también tienen una etapa de mejoramiento de alta frecuencia y una funcionalidad de estéreo. Los esquemas de codificación de voz muestran una alta calidad para señales de voz aún a bajas tasas de bits, pero muestran una baja calidad para señales de música a bajas tasas de bits.

En vista de los esquemas de codificación disponibles que se han mencionado anteriormente, y de los cuales algunos son más adecuados para la codificación de voz y otros son más adecuados para la codificación de música, la segmentación y clasificación automáticas de una señal de audio a codificarse son importantes herramientas en muchas aplicaciones multimedia y se pueden usar para seleccionar un proceso apropiado para cada categoría diferente que ocurre en una señal de audio. El rendimiento total de la aplicación depende mucho de la fiabilidad de la clasificación de la señal de audio. De hecho, una clasificación equivocada puede generar selecciones y sintonizaciones incorrectas de los siguientes procesos.

La Figura 6 muestra un diseño de un codificador convencional usado para codificar por separado música y voz, que depende de la discriminación de una señal de audio. El diseño del codificador comprende una rama de codificación de voz 100 que incluye un codificador de voz apropiado 102, por ejemplo un codificador de voz AMR-WB+ tal como se describe en el documento "Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec", 3GPP TS 26.290 V6.3.0, 2005-06, especificación técnica. El diseño del codificador comprende además una rama de codificación de música 104 que incluye un codificador de música 106, por ejemplo un codificador de música de AAC tal como se describe, por ejemplo, en la Codificación Genérica de Imágenes en Movimiento y de Audio Asociado: Codificación de Audio Avanzada. Norma internacional 13818-7, ISO/IEC JTC1/SC29/WG11 Grupo de Expertos en Imágenes en Movimiento 1997.

Las salidas de los codificadores 102 y 106 están conectadas con la entrada de un multiplexor 108. Las entradas de los codificadores 102 y 106 se pueden conectar selectivamente con una línea de entrada 110 que lleva una señal de audio de entrada. La señal de audio de entrada se aplica selectivamente al codificador de voz 102 o al codificador de música 106 mediante un conmutador 112 que se muestra esquemáticamente en la Figura 6 y que está controlado por un control de conmutación 114. El diseño del codificador comprende además a un discriminador de voz/música

116 que también recibe en una entrada del mismo la señal de audio de entrada y que emite una señal de control al control de conmutación 114. El control de conmutación 114 emite adicionalmente una señal indicadora de modo sobre una línea 118 que se introduce en una segunda entrada del multiplexor 108 de modo que se puede enviar una señal indicadora de modo junto con una señal codificada. La señal indicadora de modo puede tener sólo un bit, que indica que un bloque de datos asociado con el indicador de modo es o una voz codificada o una música codificada, de modo que, por ejemplo, no hace falta hacer una discriminación en un decodificador. En su lugar, basándose en el bit indicador de modo transmitido junto con los datos codificados al lado del decodificador se puede generar una señal de conmutación apropiada basándose en el indicador de modo para encaminar los datos recibidos y codificados a un decodificador apropiado de voz o de música.

La Figura 6 es un diseño tradicional de un codificador que se usa para codificar digitalmente señales de voz y de música aplicadas a la línea 110. En general, los codificadores de voz trabajan mejor con señales de voz y los codificadores de audio trabajan mejor con señales de música. Un esquema de codificación universal se puede diseñar usando un sistema de múltiples codificadores que conmutan de un codificador a otro de acuerdo con la naturaleza de la señal de entrada. Aquí el problema no trivial es diseñar un clasificador de señal de entrada bien adecuado que accione el elemento de conmutación. El clasificador es el discriminador de voz/música 116 que se muestra en la Figura 6. Normalmente, una clasificación confiable de una señal de audio introduce un retardo elevado, mientras, por otro lado, el retardo es un factor importante en las aplicaciones en tiempo real.

En general, se desea que el retardo algorítmico total introducido por el discriminador de voz/música sea suficientemente corto para que permita que se usen los codificadores de conmutación en una aplicación en tiempo real.

La Figura 7 muestra los retardos que se experimentan en un diseño de codificador como se muestra en la Figura 6. Se supone que la señal aplicada sobre la línea de entrada 110 debe codificarse en una base de tramas de 1024 muestras con una tasa de muestreo de 16 kHz de modo que la discriminación de voz/música debe entregar una decisión en cada trama, es decir cada 64 milisegundos. La transición entre dos codificadores se realiza por ejemplo en una manera que se describe en el documento WO2008/071353 A2 y el discriminador de voz/música no debe aumentar significativamente el retardo algorítmico de los decodificadores de conmutación que en total es de unas 1600 muestras sin considerar el retardo que se necesita para el discriminador de voz/música. Además se desea proporcionar la decisión de voz/música para la misma trama en la que se decide la conmutación del bloque de AAC. La situación se ilustra en la Figura 7, que muestra un bloque de AAC largo 120, que tiene una longitud de 2048 muestras, es decir el bloque largo 120 comprende dos tramas de 1024 muestras, un bloque ACC corto 122 de una trama de 1024 muestras y una súper trama AMR-WB+ 124 de una trama de 1024 muestras.

En la Figura 7 se toman la decisión de conmutación de bloque de AAC y la decisión de voz/ música en las tramas 126 y 128, respectivamente, de 1024 muestras, que cubren el mismo periodo de tiempo. Las dos decisiones se toman en esta posición particular para hacer que la codificación pueda usar al mismo tiempo ventanas de transición para pasar adecuadamente de un modo al otro. En consecuencia, se introduce un retardo mínimo de 512+64 muestras por las dos decisiones. Este retardo tiene que añadirse al retardo de 1024 muestras generadas por la superposición de 50% de la AAC MDCT, que da como resultado un retardo mínimo de 1600 muestras. En una AAC convencional, sólo se hace la conmutación de bloque y el retardo es exactamente 1600 muestras. Se requiere este retardo para conmutar a la vez de un bloque largo a los bloques cortos cuando se detectan componentes transitorios en la trama 126. Esta conmutación de longitud de transformación es deseable para evitar un artefacto pre-eco. En cualquier caso (bloques largos o cortos) la trama decodificada 130 en la Figura 7 representa la primera trama completa que puede restituirse en el lado del decodificador.

En un codificador de conmutación que utiliza una AAC como codificador de música, la decisión de conmutación que viene de una etapa de decisión debe evitar añadir demasiado retardo adicional al retardo original de la AAC. El retardo adicional viene de la trama anticipada 132 que es necesaria para el análisis de señal en la etapa de decisión. Con una tasa de muestreo de por ejemplo 16 kHz, el retardo de la AAC es de 100 ms mientras un discriminador convencional de voz/música utiliza alrededor de 500 ms de anticipación, lo cual da como resultado una estructura de codificación conmutada con un retardo de 600 ms. Entonces, el retardo total correspondería a seis veces el retardo de la AAC original.

Los enfoques convencionales como se han descrito anteriormente son desventajosos debido a que, para una clasificación confiable de una señal de audio, se introduce un retardo elevado no deseado de modo que existe la necesidad de un nuevo enfoque para la discriminación de una señal que incluye segmentos de diferentes tipos, en el que un retardo algorítmico adicional introducido por el discriminador es suficientemente pequeño de modo que los codificadores de conmutación se puedan usar también para una aplicación en tiempo real.

J. Wang, et. al. "Real-time speech/music classification with a hierarchical oblique decision tree", ICASSP 2008, IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, 31 de marzo de 2008 a 4 de abril de 2008, describen un enfoque para la clasificación de voz/música usando rasgos distintivos a corto plazo y rasgos

distintivos a largo plazo derivados de la misma cantidad de tramas. Estos rasgos distintivos a corto plazo y rasgos distintivos a largo plazo se usan para clasificar la señal, pero sólo se aprovechan propiedades limitadas de los rasgos distintivos a corto plazo, por ejemplo no se aprovecha la reactividad de la clasificación, aunque tienen un papel importante para la mayoría de las aplicaciones de codificación de audio.

Se desvelan esquemas de discriminación de voz/música para codificación e voz y audio combinada por L. Tancerel et al. "Combined speech and audio coding by discrimination", Proc. IEEE Workshop on Speech Coding, 17-20 de septiembre de 2000 y el documento US 2003/0101050 A1.

Sumario de la invención

Es un objeto de la invención proporcionar un enfoque mejorado para discriminar en una señal segmentos de diferentes tipos, mientras se mantenga bajo cualquier retardo introducido por la discriminación.

Este objeto se consigue mediante un método de la reivindicación 1, un programa informático de la reivindicación 13 y por un discriminador de la reivindicación 14.

Las realizaciones de la invención proporcionan la señal de salida basándose en una comparación del resultado de análisis a corto plazo al resultado de análisis a largo plazo.

Las realizaciones de la invención se refieren a un enfoque para clasificar diferentes segmentos de tiempo a corto plazo no superpuestos de una señal de audio ya sea de voz o no de voz o de clases adicionales. El enfoque está basado en la extracción de rasgos distintivos y el análisis de sus estadísticas a través de dos longitudes de ventana de análisis diferentes. La primera ventana es larga y mira principalmente hacia el pasado. La primera ventana se usa para obtener un indicio de decisión confiable pero retardada para la clasificación de la señal. La segunda ventana es corta y considera principalmente el segmento procesado en el momento actual del segmento actual. La segunda ventana se usa para obtener un indicio de decisión instantáneo. Los dos indicios de decisión se combinan de manera óptima, preferentemente usando una decisión de histéresis que obtiene la información de memoria desde el indicio retardado y la información instantánea desde el indicio instantáneo.

Las realizaciones de la invención usan rasgos distintivos a largo plazo tanto en el clasificador a corto plazo como en el clasificador a largo plazo de modo que los dos clasificadores aprovechan diferentes estadísticas del mismo rasgo distintivo. El clasificador a corto plazo extraerá únicamente la información instantánea puesto que tiene acceso únicamente a un conjunto de rasgos distintivos. Por ejemplo, puede aprovechar el promedio de los rasgos distintivos. Por otra parte, el clasificador a largo plazo tiene acceso a varios conjuntos de rasgos distintivos puesto que considera varias tramas. Como consecuencia, el clasificador a largo plazo puede aprovechar más características de la señal aprovechando estadísticas a través de más tramas que el clasificador a corto plazo. Por ejemplo, el clasificador a largo plazo puede aprovechar la varianza de los rasgos distintivos o la evolución de los rasgos distintivos con el tiempo. Por tanto, el clasificador a largo plazo puede aprovechar más información que el clasificador a corto plazo, pero introduce retardo o latencia. Sin embargo, los rasgos distintivos a largo plazo, a pesar de introducir retardo o latencia, harán los resultados de clasificación a largo plazo más robustos y fiables. En algunas realizaciones los clasificadores a corto plazo y a largo plazo pueden considerar los mismos rasgos distintivos a corto plazo, que pueden calcularse una vez y usarse por ambos clasificadores. Por lo tanto, en una realización de este tipo el clasificador a largo plazo puede recibir los rasgos distintivos a corto plazo directamente desde el clasificador a corto plazo.

El nuevo enfoque de esta manera permite obtener una clasificación que es robusta mientras introduce un retardo bajo. Aparte de los enfoques convencionales, las realizaciones de la invención limitan el retardo introducido por la decisión de voz/música mientras mantienen una decisión confiable.

Breve descripción de los dibujos

Las realizaciones de la invención se describirán a continuación haciendo referencia a los dibujos adjuntos, en los cuales:

- La Figura 1 es un diagrama de bloques de un discriminador de voz/música de acuerdo con una realización de la invención;
- La Figura 2 muestra la ventana de análisis utilizada por los clasificadores a largo plazo y a corto plazo del discriminador de la Figura 1;
- La Figura 3 muestra una decisión de histéresis utilizada en el discriminador de la Figura 1;
- La Figura 4 es un diagrama de bloques de un esquema de codificación ejemplar que comprende un discriminador de acuerdo con algunas realizaciones de la invención;
- La Figura 5 es un diagrama de bloques de un esquema de decodificación que corresponde al esquema de codificación de la Figura 4

La Figura 6 muestra un diseño de codificador convencional utilizado para codificar de manera separada voz y música dependiendo de una discriminación de una señal de audio; y
La Figura 7 muestra los retardos que se experimentan en el diseño de codificador que se muestra en la Figura 6.

Descripción de las realizaciones de la invención

La Figura 1 es un diagrama de bloques de un discriminador de voz/música 116 de acuerdo con una realización de la invención. El discriminador de voz/música 116 comprende un clasificador a corto plazo 150 que recibe en una entrada del mismo una señal de entrada, por ejemplo una señal de audio que comprende segmentos de voz y de música. El clasificador a corto plazo 150 emite sobre una línea de salida 152 un resultado de clasificación a corto plazo, el indicio de decisión instantáneo. El discriminador 116 comprende además un clasificador a largo plazo 154 que también recibe la señal de entrada y emite sobre una línea de salida 156 el resultado de clasificación a largo plazo, el indicio de decisión retardada. Además se proporciona un circuito de decisión de histéresis 158 que combina las señales de salida del clasificador a corto plazo 150 y del clasificador a largo plazo 154 en una manera, que se describirá a continuación con más detalle para generar una señal de decisión de voz/música que se emite a la línea 160 y que se puede utilizar para controlar el procesamiento adicional de un segmento de una señal de entrada en una manera que se ha descrito anteriormente con respecto a la Figura 6, es decir la señal de decisión de voz/música 160 se puede utilizar para encaminar el segmento de señal de entrada, que se ha clasificado, a un codificador de voz o a un codificador de audio.

De ese modo, de acuerdo con las realizaciones de la invención, se utilizan dos diferentes clasificadores 150 y 154 en paralelo sobre la señal de entrada aplicada a los respectivos clasificadores mediante la línea de entrada 110. Los dos clasificadores se llaman clasificador a largo plazo 154 y clasificador a corto plazo 150, en el que los dos clasificadores se distinguen analizando las estadísticas de los rasgos distintivos sobre los cuales operan a través de las ventanas de análisis. Los dos clasificadores entregan las señales de salida 152 y 156, en concreto el indicio de decisión instantáneo (IDC) y el indicio de decisión retardado (DDC). El clasificador a corto plazo 150 genera un IDC basándose en rasgos distintivos a corto plazo que tienen como objetivo capturar informaciones instantáneas con respecto a la naturaleza de la señal de entrada. Están relacionados con los atributos a corto plazo de la señal que puede cambiar rápidamente y en cualquier momento. En consecuencia, se espera que los rasgos distintivos a corto plazo sean reactivos y no introduzcan un gran retardo al proceso de discriminación en su totalidad. Por ejemplo, debido a que la voz se considera que es cuasi-estacionaria en duraciones de 5 a 20 ms, los rasgos distintivos a corto plazo se pueden calcular para cada trama de 16 ms en una señal muestreada a 16 kHz. El clasificador a largo plazo 154 genera los DDC basándose en rasgos distintivos que resultan de observaciones más largas de la señal (rasgos distintivos a largo plazo) y por lo tanto permiten lograr una clasificación más confiable.

La Figura 2 muestra las ventanas de análisis utilizadas por el clasificador a largo plazo 154 y el clasificador a corto plazo 150 que se muestran en la Figura 1. Suponiendo una trama de 1024 muestras con una tasa de muestreo de 16 kHz, la longitud de la ventana del clasificador a largo plazo 162 es de 4*1024+128 muestras, es decir, la ventana del clasificador a largo plazo 162 se extiende a lo largo de cuatro tramas de la señal de audio y son necesarias unas 128 muestras adicionales por el clasificador a largo plazo 154 para llevar a cabo su análisis. Este retardo adicional, que se denomina también como “anticipación” está indicado en la Figura 2 bajo el número de referencia 164. La Figura 2 muestra también la ventana del clasificador a corto plazo 166 que es de 1024+128 muestras, es decir se extiende a lo largo de una trama de la señal de audio y el retardo adicional que se necesita para analizar un segmento actual. El segmento actual está indicado con el número 128 como el segmento para el cual hace falta hacer la decisión de voz/música.

La ventana del clasificador a largo plazo indicada en la Figura 2 es suficientemente larga para obtener la característica de modulación de energía de 4 Hz de la voz. La modulación de energía de 4 Hz es una característica importante y discriminatoria de la voz, la cual se aprovecha tradicionalmente en los discriminadores robustos de voz/música usados, por ejemplo, por Scheirer E. y Slaney M., "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", ICASSP’97, Múnich, 1997. La modulación de energía de 4 Hz es un rasgo distintivo que sólo se puede extraer observando la señal sobre un segmento de tiempo largo. El retardo adicional que se introduce por el discriminador de voz/música es igual a la anticipación 164 de 128 muestras que es necesaria por cada uno de las clasificadores 150 y 154 para llevar a cabo el respectivo análisis como un análisis de predicción lineal perceptivo que se describe por H. Hermansky, "Perceptive linear prediction (plp) analysis of speech", Journal of the Acoustical Society of America, vol. 87, n.º 4, págs. 1738-1752, 1990 y H. Hermansky, et al., "Perceptually based linear predictive analysis of speech", ICASSP 5.509-512, 1985. De ese modo, cuando se utiliza el discriminador de la realización anteriormente descrita en un diseño de codificador tal como se muestra en la Figura 6, el retardo total de los codificadores de conmutación 102 y 106 será de 1600+128 muestras, lo cual es igual a 108 milisegundos que es suficientemente bajo para aplicaciones en tiempo real.

Ahora se hace referencia a la Figura 3 que describe la combinación de las señales de salida 152 y 156 de los clasificadores 150 y 154 del discriminador 116 para obtener una señal de decisión de voz/música 160. El indicio de decisión retardada DDC y el indicio de decisión instantáneo IDC de acuerdo con una realización de la invención se

combinan utilizando una decisión de histéresis. Los procesos de histéresis se utilizan ampliamente para post procesar decisiones para estabilizarlas. La Figura 3 muestra una decisión de histéresis de dos estados como una función del DDC y del IDC para determinar si la señal de decisión de voz o música debe indicar que un segmento que se está procesando actualmente de la señal de entrada como un segmento de voz o un segmento de música. El ciclo de histéresis característico se puede ver en la Figura 3 y el IDC y el DDC están normalizados por los clasificadores 150 y 154 de tal manera que los valores están entre -1 y 1, en el que -1 significa que la probabilidad es totalmente del tipo música, y 1 significa que la probabilidad es totalmente del tipo voz.

La decisión se basa sobre el valor de una función F(IDC, DDC), unos ejemplos de la cual se describirán a continuación. En la Figura 3 F1(DDC, IDC) indica un umbral, el cual F(IDC, DDC) debe cruzar para ir de un estado de música a un estado de voz. F2(DDC, IDC) muestra un umbral, el cual F(IDC, DDC) debe cruzar para ir de un estado de voz a un estado de música. La decisión final D(n) para un segmento actual o una trama actual que tiene el índice n, se puede calcular entonces basándose en el siguiente pseudo código:

```
% Pseudo código de decisión de histéresis
If(D(n-1) == música)
       If(F(IDC, DDC) < F1(DDC, IDC))
              D(n) == música
       Else
              D(n) == voz
Else
       If(F(IDC, DDC) > F2(DDC, IDC))
              D(n) == voz
       Else
              D(n) == música
% Fin de pseudo código de decisión de histéresis
```

De acuerdo con algunas realizaciones de la invención, la función F(IDC, DDC) y los umbrales anteriormente mencionados se fijan de la siguiente manera:

$$F(IDC, DDC) = IDC$$

$$F1(IDC, DDC) = 0,4-0,4*DDC$$

$$F2(IDC, DDC) = -0,4-0,4*DDC$$

De manera alternativa, se pueden usar las definiciones siguientes:

$$F(IDC, DDC) = (2*IDC + DDC)/3$$

$$F1(IDC, DDC) = -0,75 * DDC$$

$$F2(IDC, DDC) = -0,75 * DDC$$

Cuando se utiliza la última definición, el ciclo de histéresis se anula y la decisión se toma sólo basándose en un umbral adaptivo único.

La invención no se limita a la decisión de histéresis anteriormente descrita. A continuación, se describirán las realizaciones adicionales para la combinación de los resultados de análisis para obtener la señal de salida.

Una determinación simple de umbrales se puede utilizar en lugar de la decisión de histéresis constituyendo el umbral de una manera que el mismo aprovecha las características tanto del DDC como del IDC. Se considera que el DDC es un indicio de discriminación más confiable porque viene de una observación más larga de la señal. Sin embargo, el DDC se calcula parcialmente basándose en una observación del pasado de la señal. Un clasificador convencional, que sólo compara el valor de DDC con el umbral 0 y que clasifica un segmento como tipo voz cuando DDC > 0 o como tipo música en el caso contrario, tendrá una decisión retardada. En una realización de la invención, podemos adaptar la determinación de umbrales aprovechando el IDC y hacer la decisión más reactiva. Por este propósito, se puede adaptar el umbral basándose en el siguiente pseudo código:

```
% Pseudo código de la determinación adaptiva de umbrales
If(DDC > -0,5*IDC)
   D(n) == voz
   Else
   D(n) == música
   % Fin de la determinación adaptiva de umbrales
```

En otra realización se puede utilizar el DDC para hacer más confiable al IDC. El IDC se conoce que es reactivo pero no tan confiable como el DDC. Además, la observación de la evolución del DDC entre el segmento pasado y el actual puede dar otra indicación de cómo la trama 166 en la Figura 2 influye sobre el DDC calculado para el segmento 162. La notación DDC(n) se utiliza para el valor actual del DDC y el DDC(n-1) para el valor pasado. Utilizando ambos valores, DDC(n) e DDC(n-1), se puede hacer que el IDC sea más confiable utilizando un árbol de decisión tal como se describe a continuación:

```
% Pseudo código del árbol de decisión
If(IDC > 0 && DDC(n) > 0)
   D(n) = voz
   Else if (IDC < 0 && DDC(n) < 0)
   D(n) =música
   Else if (IDC > 0 && DDC(n) - DDC(n-1)>0)
   D(n) = voz
   Else if (IDC < 0 && DDC(n) - DDC(n-1)<0)
   D(n) = música
   Else if (DDC > 0)
   D(n) = voz
   Else
   D(n) = música
   % Fin del árbol de decisión
```

En el árbol de decisión, la decisión se toma directamente si ambos indicios muestran la misma probabilidad. Si los dos indicios dan indicaciones contradictorias, miramos a la evolución del DDC. Si la diferencia DDC(n) - DDC(n–1) es positiva, podemos suponer que el segmento actual es del tipo voz. De otra manera, podemos suponer que el segmento actual es del tipo música. Si esta nueva indicación va en la misma dirección que el IDC, se toma la decisión final. Si ambos intentos fracasan en dar una decisión clara, se toma la decisión considerando sólo el indicio retardado DDC, porque la fiabilidad del IDC no se pudo validar.

A continuación se describirán los respectivos clasificadores 150 y 154 con más detalle de acuerdo con una realización de la invención.

Volviendo en primer lugar al clasificador a largo plazo 154, se observa que el mismo es para extraer un conjunto de rasgos distintivos de cada sub-trama de 256 muestras. El primer rasgo distintivo es el coeficiente cepstral de predicción lineal perceptiva (PLPCC) como se describe por H. Hermansky, "Perceptive linear prediction (plp) analysis of speech", Journal of the Acoustical Society of America, vol. 87, n.º 4, págs. 1738-1752, 1990 y H. Hermansky, et al., "Perceptually based linear predictive analysis of speech", ICASSP 5.509-512, 1985. Los PLPCC son eficientes para la clasificación de hablantes usando la estimación de la percepción auditiva humana. Este rasgo distintivo se puede usar para discriminar voz y música y, de hecho, permite distinguir tanto los formantes característicos de la voz como la modulación silábica de 4 Hz de la voz observando la variación de los rasgos distintivos en el tiempo.

Sin embargo, para ser más robustos los PLPCC se combinan con otro rasgo distintivo que es capaz de capturar la información de tono, que es otra característica importante de la voz y puede ser crítica para la codificación. De hecho, la codificación de voz se apoya en la suposición que una señal de entrada es una señal pseudo mono-periódica. Los esquemas de codificación de voz son eficientes para una señal de este tipo. Por otro lado, la característica de tono de la voz perjudica mucho la eficiencia de codificación de codificadores de música. La fluctuación suave de retardo de tono, dado por el vibrato natural de la voz, hace que la representación de frecuencia en los codificadores de música no pueda compactar eficientemente la energía que se requiere para obtener una alta eficiencia de codificación.

Se pueden determinar los siguientes rasgos distintivos característicos de tono:

Relación de energía de pulsos glotales:

Este rasgo distintivo calcula la relación de energía entre los pulsos glotales y la señal residual de LPC. Los pulsos glotales se extraen desde la señal residual de LPC usando un algoritmo de selección de picos. Normalmente, la señal residual de LPC de un segmento vocalizado muestra una gran estructura de tipo pulso que proviene de la vibración glotal. Este rasgo distintivo es alto durante segmentos vocalizados.

Predicción de ganancia a largo plazo:

Normalmente se calcula la ganancia en los codificadores de voz (véase por ejemplo "Extended Adaptive Multi-Rate - Wideband (AMRWB+) codec", 3GPP TS 26.290 V6.3.0, 06-2005, especificación técnica) durante la predicción a

largo plazo. Este rasgo distintivo mide la periodicidad de la señal y se basa en la estimación de retardo de tono.

Fluctuación de retardo de tono:

Este rasgo distintivo determina la diferencia de la estimación de retardo de tono actual cuando se compara con la última sub-trama. Para la voz vocalizada este rasgo distintivo debe ser bajo pero no cero y debe evolucionar suavemente.

Una vez que el clasificador a largo plazo ha extraído el conjunto necesario de rasgos distintivos, se utiliza un clasificador estadístico sobre estos rasgos distintivos extraídos. El clasificador se ha entrenado en primer lugar extrayendo los rasgos distintivos de un conjunto de entrenamiento de voz y un conjunto de entrenamiento de música. Los rasgos distintivos extraídos se normalizan a un valor promedio de 0 y una varianza de 1 sobre ambos conjuntos de entrenamiento. Para cada conjunto de entrenamiento, se recogen los rasgos distintivos extraídos y normalizados dentro de una ventana de clasificador a largo plazo y se modelan con un modelo de mezcla gaussiana (GMM) que usa cinco gaussianos. Al final de la secuencia de entrenamiento, se obtienen y se guardan un conjunto de parámetros de normalización y dos conjuntos de parámetros de GMM.

Para cada trama a clasificarse, en primer lugar se extraen y se normalizan los rasgos distintivos con los parámetros de normalización. Se calcula la probabilidad máxima para voz (lld_speech) y la probabilidad máxima para música (lld_music) para los rasgos distintivos extraídos y normalizados utilizando el GMM de la clase de voz y el GMM de la clase de música, respectivamente. El indicio de decisión retardada DDC se calcula entonces de la siguiente manera:

$$DDC = (lld\_speech - lld\_music) / (abs(lld\_music) + abs (lld\_speech))$$

El DDC está delimitado entre los valores -1 y 1, y es positivo cuando la probabilidad máxima para voz es más alta que probabilidad máxima para música, lld_speech > lld_music.

El clasificador a corto plazo utiliza como un rasgo distintivo a corto plazo los PLPCC. Diferente al clasificador a largo plazo, este rasgo distintivo sólo se analiza en la ventana 128. Se aprovechan las estadísticas de este rasgo distintivo en este tiempo corto mediante un modelo de mezcla gaussiana (GMM) que usa cinco gaussianos. Dos modelos se entrenan, uno para música y el otro para voz. Vale la pena mencionar, que los dos modelos son diferentes de los modelos que se obtienen para el clasificador a largo plazo. Para cada trama a clasificarse, en primer lugar se extraen los PLPCC y se calcula la probabilidad máxima para voz (lld_speech) y la probabilidad máxima para música (lld_music) para el uso del GMM de la categoría de voz y del GMM de la categoría de música, respectivamente. A continuación se calcula el indicio de decisión instantáneo IDC de la siguiente manera:

$$IDC = (lld\_speech – lld\_music) / (abs(lld\_music) + abs (lld\_speech))$$

El IDC está delimitado entre los valores -1 y 1.

De ese modo, el clasificador a corto plazo 150 genera el resultado de clasificación a corto plazo de la señal basándose en el rasgo distintivo del “coeficiente cepstral de predicción lineal perceptiva” (PLPCC), y el clasificador a largo plazo 154 genera el resultado de clasificación a largo plazo de la señal basándose en el mismo rasgo distintivo “coeficiente cepstral de predicción lineal perceptiva” (PLPCC) y el rasgo distintivo o los rasgos distintivos adicionales anteriormente mencionados, por ejemplo, el rasgo distintivo o los rasgos distintivos de las características de tono. Más aún, el clasificador a largo plazo puede aprovechar distintas características del rasgo distintivo compartido, es decir los PLPCC, ya que tiene acceso a una ventana de observación más larga. De ese modo, al combinar los resultados de clasificación a corto plazo y a largo plazo, se consideran suficientemente los rasgos distintivos a corto plazo para la clasificación, es decir, sus propiedades se aprovechan suficientemente.

A continuación se describirá con más detalle otra realización para los respectivos clasificadores 150 y 154.

Los rasgos distintivos a corto plazo analizados por el clasificador a corto plazo de acuerdo con esta realización corresponden principalmente a los coeficientes cepstral de predicción lineal perceptiva (PLPCC) anteriormente mencionados. Tanto los PLPCC como los MFCC (véase anteriormente) se utilizan ampliamente en el reconocimiento de voz y del hablante. Se mantienen los PLPCC porque comparten una gran parte de la funcionalidad de la predicción lineal (LP) que se utiliza en la mayoría de los codificadores de voz modernos y si ya están implementados también en un codificador de audio de conmutación. Los PLPCC pueden extraer le estructura de formato de la voz como lo hace también el LP pero teniendo en cuenta las consideraciones perceptivas, los PLPCC son más independientes del hablante y por lo tanto más importantes con respecto a la información lingüística. Se utiliza un orden de 16 en la señal de entrada muestreada de 16 kHz.

Aparte de los PLPCC, se calcula una intensidad de vocalización como un rasgo distintivo a corto plazo. No se considera intensidad de vocalización como realmente discriminatoria por sí misma, pero es beneficiosa en

asociación con los PLPCC en la dimensión del rasgo distintivo. La intensidad de vocalización permite extraer la dimensión de rasgos distintivos al menos dos agrupamientos que corresponden respectivamente a las pronunciaciones vocalizadas y no vocalizadas de la voz. Se basa sobre un cálculo de calidad de sistema utilizando diferentes parámetros, en particular, un contador de cruce en cero (zc), la inclinación espectral (tilt), la estabilidad de tono (ps) y la correlación normalizada del tono (nc). Los cuatro parámetros están normalizados entre 0 y 1 en una manera que 0 corresponde a una señal típicamente no vocalizada y 1 corresponde a una señal típicamente vocalizada. En esta realización, la intensidad de vocalización está inspirada a partir de los criterios de clasificación de voz usados en el codificador de voz VMR-WB descrito por Milan Jelinek y Redwan Salami, "Wideband speech coding advances in vmr-wb standard", IEEE Trans. on Audio, Speech and Language Processing, vol. 15, n.º 4, págs. 1167-1179, mayo de 2007. Se basa sobre un rastreador de tono evolucionado basándose en una autocorrelación. Para la trama con el índice k, la intensidad de vocalización u(k) tienen la siguiente fórmula:

$$v(k)=\frac{1}{5}(2*nc(k)+2*ps(k)+tilt(k)+zc(k))$$

La capacidad discriminatoria de los rasgos distintivos a corto plazo se evalúa por los modelos de mezcla gaussiana (GMMS) como un clasificador. Se aplican dos GMM, uno para la categoría de voz y el otro para la categoría de música. La cantidad de mezclas se hace para evaluar el efecto sobre el rendimiento. La Tabla 1 muestra las tasas de precisión para los diferentes números de mezclas. Se calcula una decisión para cada segmento de cuatro tramas sucesivas. El retardo global es entonces de 64 ms, que es apropiado para una codificación de audio de conmutación. Se puede observar, que el rendimiento aumenta con la cantidad de mezclas. La brecha entre 1 GMM y 5 GMM es particularmente importante y se puede explicar por el hecho de que la representación de formante de la voz es demasiado compleja para definirse suficientemente por sólo un gaussiano.

Tabla 1: precisión de clasificación de rasgos distintivos a corto plazo en %

	1 GMM	**5 GMM**	**10 GMM**	**20 GMM**
Voz	95,33	96,52	97,02	97,60
Música	92,17	91,97	9161	91,77
Promedio	93,75	94,25	94,31	94,68

Tratando ahora el clasificador a largo plazo 154, se observa que muchos trabajos, por ejemplo, M. J. Carey, et. al. "A comparison of features for speech and music discrimination", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP, vol. 12, págs. 149 a 152, marzo de 1999, consideran que las varianzas de los rasgos distintivos estadísticos son más discriminatorias que los rasgos distintivos mismos. Como una regla general aproximada, se puede considerar que la música es más estacionaria y presenta generalmente menos varianza. Al contrario, la voz se puede distinguir fácilmente por su modulación de energía de 4 Hz notable ya que la señal cambia periódicamente entre los segmentos vocalizados y no vocalizados. Más aún, la sucesión de diferentes fonemas hace que los rasgos distintivos de voz sean menos constantes. En esta realización, se consideran dos rasgos distintivos a largo plazo, uno basándose en el cálculo de una varianza y el otro basándose en un conocimiento previo del contorno de tono de la voz. Se adaptan los rasgos distintivos a largo plazo a la SMD (discriminación de voz/música) de bajo retardo.

La varianza móvil de los PLPCC consiste en el cálculo de la varianza para cada conjunto de PLPCC a lo largo de una ventana de análisis superpuesta que cubre varias tramas para resaltar la última trama. Para limitar la latencia introducida, la ventana de análisis es asimétrica y considera sólo la trama actual y la historia pasada. En una primera etapa, se calcula la media móvil $ma_m(k)$ de los PLPCC sobre las últimas N tramas como se describe a continuación:

$$ma_m(k)=\sum_{i=0}^{N-1} PLPC_m(k-i)\cdot w(i)$$

donde PLPPm(k) es el coeficiente cepstral de orden m sobre un total de M coeficientes provenientes de la trama de orden k. La varianza móvil $mv_m(k)$ se define entonces como:

$$mv_m(k)=\sum_{i=0}^{N-1}(PLPC_m(k-i)-ma_m(k))^2\cdot w(i)$$

donde w es una ventana de longitud N, que en esta realización es una pendiente de rampa definida como:

$$w(i)=(N-i)/N\cdot(N+1)/2$$

Finalmente se promedia la varianza móvil sobre la dimensión cepstral:

$$mv(k) = \frac{1}{M}\sum_{m=0}^{M} mv_m(k)$$

El tono de la voz tiene unas propiedades notables y una parte de las mismas se puede observar sólo en ventanas de análisis largas. De hecho, el tono de la voz fluctúa suavemente durante los segmentos vocalizados pero es rara vez constante. Al contrario, la música presenta muy frecuentemente un tono constante durante la duración completa de una nota y un cambio abrupto durante los componentes transitorios. Los rasgos distintivos a largo plazo abarcan está característica observando el contorno del tono sobre un segmento de tiempo largo. Un parámetro de contorno de tono pc(k) se define como:

$$pc(k) = \begin{cases} 0 & \text{si} \quad |p(k)-p(k-1)|<1 \\ 0,5 & \text{si} \quad 1\leq|p(k)-p(k-1)|<2 \\ 1 & \text{si} \quad 2\leq|p(k)-p(k-1)|<20 \\ 0,5 & \text{si} \quad 20\leq|p(k)-p(k-1)|<25 \\ 0 & \text{de lo contrario} \end{cases}$$

donde p(k) es el retardo del tono calculado en el índice de trama k sobre la señal residual LP muestreada a 16 KHz. A partir del parámetro de contorno de tono se calcula una calidad de voz, sm(k), en una manera que se espera que la voz muestre un retardo de tono de una suave fluctuación durante los segmentos vocalizados y una fuerte inclinación espectral hacia las frecuencias altas durante los segmentos no vocalizados:

$$sm(k) = \begin{cases} nc(k)\cdot pc(k) & \text{si} \quad v(k)\geq 0,5 \\ (1-nc(k))\cdot(1-tilt(k)) & \text{de lo contrario} \end{cases}$$

donde nc(k), tilt(k) y v(k) se definen como anteriormente (véase el clasificador a corto plazo). A continuación se pondera la calidad de voz con la ventana w anteriormente definida y se integra sobre las últimas N tramas:

$$ams(k) = \sum_{i=0}^{N} m(k-1)w(i)$$

El contorno del tono también es una indicación importante que una señal sea apropiada para una codificación de voz o de música. De hecho, los codificadores de voz trabajan principalmente en el dominio del tiempo y suponen que la se es armónica y cuasi-estacionaria en segmentos de tiempos cortos de aproximadamente 5 ms. De esta manera ellos pueden modelar eficientemente la fluctuación natural del tono de la voz. Al contrario, la misma fluctuación daña la eficiencia de los codificadores de audio general que aprovechan las transformaciones lineales sobre ventanas de análisis largas. A continuación se distribuye la energía principal de la señal sobre varios coeficientes transformados.

Tal como para los rasgos distintivos a corto plazo, también se evalúan los rasgos distintivos a largo plazo utilizando un clasificador estadístico, y de ese modo se obtiene un resultado de clasificación a largo plazo (DDC). Los dos rasgos distintivos se calculan utilizando N = 25 tramas, por ejemplo, considerando 400 ms de historial pasado de la señal. Un análisis de discriminación lineal (ADL) se aplica en primer lugar antes de utilizar 3 GMM en el espacio unidimensional reducido. La Tabla 2 muestra el rendimiento medido sobre los conjuntos de entrenamiento y de prueba para la clasificación de segmentos de cuatro tramas sucesivas.

Tabla 2: precisión de clasificación de rasgos distintivos a largo plazo en %

	Conjunto de entrenamiento	**Conjunto de prueba**
Voz	97,99	97,84
Música	95,93	95,44
Promedio	96,96	96,64

El sistema de clasificadores combinados de acuerdo con las realizaciones de la invención combina apropiadamente

los rasgos distintivos a corto plazo y a largo plazo en una manera que aportan su propia contribución específica a la decisión final. Para este propósito se puede utilizar la etapa de decisión final de histéresis, como se ha descrito anteriormente, donde el efecto de memoria se acciona por el DDC o el indicio de discriminación a largo plazo (LTDC) mientras la entrada instantánea proviene del IDC o del indicio de discriminación a corto plazo (STDC). Los dos indicios son las salidas de los clasificadores a largo plazo y a corto plazo como se muestra en la Figura 1. Se toma la decisión basándose en el IDC pero se estabiliza con el DDC que controla dinámicamente los umbrales que desencadenan un cambio de estado.

El clasificador a largo plazo 154 utiliza ambos, los rasgos distintivos a largo plazo y a corto plazo previamente definidos por un LDA seguido por 3 GMM. El DDC es igual a la relación logarítmica entre la probabilidad del clasificador a largo plazo para la categoría de voz y para la categoría de música calculada sobre las últimas 4 x K tramas. El número de tramas, que se tienen en cuenta, puede variar con el parámetro K para agregar más o menos efecto de memoria a la decisión final. Al contrario, el clasificador a corto plazo utiliza sólo los rasgos distintivos a corto plazo con 5 GMM que muestran un buen compromiso entre el rendimiento y la complejidad. El IDC es igual a la relación logarítmica entre la probabilidad del clasificador a largo plazo para la categoría de voz y para la categoría de música calculada sólo sobre las últimas 4 tramas.

Para evaluar el enfoque inventivo, en particular, para la codificación de audio de conmutación, se evaluaron tres diferentes tipos de rendimiento. Una primera medición de rendimiento es el rendimiento convencional de voz frente a música (SvM). Se evalúa sobre un conjunto grande de elementos de música y de voz. Una segunda medición de rendimiento se hace sobre un largo elemento único que tienen segmentos de voz y de música que alternan cada 3 segundos. La precisión de discriminación se denomina entonces el rendimiento de voz antes/después de música (SabM) y refleja principalmente la reactividad del sistema. Finalmente, la estabilidad de la decisión se evalúa llevando a cabo la clasificación sobre un conjunto grande de elementos de voz sobre música. La mezcla entre voz y música se hace en diferentes niveles de un elemento a otro. Entonces se obtiene el rendimiento de voz sobre música (VsM) calculando la relación de la cantidad de conmutaciones de categoría que tuvo lugar durante la cantidad total de tramas.

Se utiliza el clasificador a largo plazo y el clasificador a corto plazo como referencias para evaluar los enfoques de clasificadores individuales convencionales. El clasificador a corto plazo muestra una buena reactividad, mientras tiene una estabilidad y una capacidad de discriminación total más bajas. Por otro lado, el clasificador a largo plazo, especialmente cuando aumenta el número de tramas a 4 x K, puede alcanzar una mejor estabilidad y un mejor comportamiento de discriminación comprometiendo la reactividad para la decisión. En comparación con el enfoque convencional recién mencionado anteriormente, los rendimientos del sistema clasificador combinado de acuerdo con la invención tienen varias ventajas. Una ventaja es que mantiene un buen rendimiento de voz pura frente a la discriminación de música mientras conserva la reactividad del sistema. Otra ventaja es el buen equilibrio entre reactividad y estabilidad.

A continuación, se hace referencia a las Figuras 4 y 5 que muestran esquemas de codificación y decodificación ejemplares, los cuales incluyen una etapa de discriminación o decisión que funciona de acuerdo con las realizaciones de la invención.

De acuerdo con el esquema de codificación ejemplar que se muestra en la Figura 4, se introduce una señal mono, una señal estéreo o una señal de múltiples canales en una etapa de pre-procesamiento común 200.

La etapa de pre-procesamiento común 200 puede tener una funcionalidad de estéreo conjunta, una funcionalidad de envolvente y/o una funcionalidad de extensión de ancho de banda. En la salida de la etapa 200 hay un canal mono, un canal estéreo o múltiples canales que es la entrada para uno o más conmutadores 202. El conmutador 202 puede proporcionarse para cada salida de la etapa 200, cuando la etapa 200 tiene dos o más salidas, es decir, cuando la etapa 200 emite una señal estéreo o una señal de múltiples canales. A modo de ejemplo, el primer canal de una señal estéreo puede ser un canal de voz y el segundo canal de la señal estéreo puede ser un canal de música. En este caso, la decisión en una etapa de decisión 204 puede ser diferente entre los dos canales para el mismo instante de tiempo.

Se controla el conmutador 202 mediante la etapa de decisión 204. La etapa de decisión comprende un discriminador de acuerdo con las realizaciones de la invención y recibe, como una entrada, una señal de entrada hacia la etapa 200 o una salida de señal desde la etapa 200. De manera alternativa, la etapa de decisión 204 puede recibir también una información secundaria, que está incluida en la señal mono, la señal estéreo o la señal de múltiples canales o está al menos asociada con tal señal, donde existe la información que se generó, por ejemplo, cuando se produjo originalmente la señal mono, la señal estéreo o la señal de múltiples canales.

En una realización, la etapa de decisión no controla la etapa de pre-procesamiento 200 y la flecha entre las etapas 204 y 200 no existe. En otra realización, el procesamiento en la etapa 200 está controlado en cierto grado por la etapa de decisión 204 para fijar uno o más parámetros en la etapa 200 basándose en la decisión. Sin embargo, esto

no influirá al algoritmo general en la etapa 200 de modo que la funcionalidad principal en la etapa 200 está activa con independencia de la decisión en la etapa 204.

La etapa de decisión 204 acciona el conmutador 202 para alimentar la salida de la etapa de pre-procesamiento común en una porción de codificación de frecuencia 206 ilustrada en una rama superior de la Figura 4 o en una porción de codificación del dominio del LPC 208 ilustrada en una rama inferior de la Figura 4.

En una realización, el conmutador 202 conmuta entre las dos ramas de codificación 206, 208. En otra realización, puede haber otras ramas de codificación tales como una tercera rama de codificación o incluso una cuarta rama de codificación o incluso más ramas de codificación. En una realización con tres ramas de codificación, la tercera rama de codificación puede ser similar a la segunda rama de codificación, pero incluye un codificador de excitación diferente del codificador de excitación 210 en la segunda rama de codificación 208. En tal realización, la segunda rama de codificación comprende la etapa LPC 212 y un codificador de excitación 210 basado en un libro de códigos tal como en el ACELP, y la tercera rama de codificación comprende una etapa LPC y un codificador de excitación que funciona con una representación espectral de la señal de salida de la etapa LPC.

La rama de codificación del dominio de la frecuencia comprende un bloque de conversión 214 que funciona para convertir la señal de salida de la etapa de pre-procesamiento común en un dominio espectral. El bloque de conversión espectral puede incluir un algoritmo MDCT, un QMF, un algoritmo FFT, un análisis de ondícula o un banco de filtros tal como un banco de filtros críticamente muestreado que tiene una cierta cantidad de canales de banco de filtros, donde las señales de subbanda en este banco de filtros pueden ser señales de valores reales o señales de valores complejos. La salida del bloque de conversión espectral 214 está codificada utilizando un codificador de audio espectral 216 que puede incluir bloques de procesamiento como se conoce a partir del esquema de codificación de AAC.

La rama inferior de codificación 208 comprende un analizador de modelo de fuente tal como el LPC 212 que emite dos tipos de señales. Una señal es una señal de información de LPC que se utiliza para controlar la característica de filtro de un filtro de síntesis de LPC. Esta información de LPC se transmite hacia un decodificador. La otra señal de salida de la etapa LPC 212 es una señal de excitación o una señal de dominio del LPC que se introduce en un codificador de excitación 210. El codificador de excitación 210 puede provenir de cualquier codificador de modelo de filtro de fuente tal como un codificador de CELP, un codificador de ACELP o cualquier otro codificador que procese una señal de dominio del LPC.

Otra implementación de un codificador de excitación puede ser una codificación de transformada de la señal de excitación. En tal realización, la señal de excitación no se codifica utilizando un mecanismo de libro de códigos ACELP, sino que la señal de excitación se convierte en una representación espectral y los valores de la representación espectral tales como las señales de subbanda en el caso de un banco de filtros o los coeficientes de frecuencia en el caso de una transformada tal como una FFT se codifican para obtener una compresión de datos. Una implementación de este tipo de codificador de excitación es el modo de codificación de TCX que se conoce a partir del AMR-WB+.

La decisión en la etapa de decisión 204 puede ser adaptable a la señal de modo que la etapa de decisión 204 realiza una discriminación de música/voz y controla el conmutador 202 de tal manera que las señales de música se introducen en la rama superior 206 y las señales de voz se introducen en la rama inferior 208. En una realización, la etapa de decisión 204 alimenta su información de decisión a un flujo de bits de salida, de modo que un decodificador puede utilizar está información de decisión para llevar a cabo las funciones correctas de decodificación.

Tal decodificador se ilustra en la Figura 5. Después de la transmisión, la señal emitida por el codificador de audio espectral 216 se introduce en un decodificador de audio espectral 218. La salida del decodificador de audio espectral 218 se introduce en un convertidor de dominio del tiempo 220. La salida del codificador de excitación 210 de la Figura 4, se introduce en un decodificador de excitación 222 que emite una señal de dominio del LPC. La señal de dominio del LPC se introduce en una etapa de síntesis de LPC 224, que recibe, como una entrada adicional, la información de LPC generada por la correspondiente etapa de análisis de LPC 212. La salida del convertidor de dominio del tiempo 220 y/o la salida de la etapa de síntesis de LPC 224 se introducen a un conmutador 226. El conmutador 226 está controlado mediante una señal de control de conmutación que, por ejemplo, se generó por la etapa de decisión 204, o que se proporcionó externamente tal como por un creador de la señal mono original, señal estéreo o señal de múltiples canales.

La salida del conmutador 226 es una señal mono completa que se introduce subsiguientemente en una etapa de post-procesamiento común 228, que puede llevar a cabo un procesamiento estéreo conjunto o un procesamiento de extensión de ancho de banda, etc. De manera alternativa, la salida del conmutador también puede ser una señal estéreo o una señal de múltiples canales. Es una señal estéreo cuando el pre-procesamiento incluye una reducción de canales a dos canales. Inclusive, puede ser una señal de múltiples canales, cuando se lleva a cabo una reducción de canales a tres canales o ninguna reducción de canales en absoluto, sino una replicación de banda

espectral.

Dependiendo de la funcionalidad específica de la etapa de post-procesamiento común, se emite una señal mono, una señal estéreo o una señal de múltiples canales, que tiene, cuando la etapa de post-procesamiento común 228 lleva a cabo una operación de extensión de ancho de banda, un ancho de banda mayor que la señal que se introdujo en el bloque 228.

En una realización, el conmutador 226 conmuta entre las dos ramas de decodificación 218, 220 y 222, 224. En otra realización, puede haber ramas de decodificación adicionales tales como una tercera rama de decodificación o incluso una cuarta rama de decodificación o incluso más ramas de decodificación. En una realización con tres ramas de decodificación, la tercera rama de decodificación puede ser similar a la segunda rama de decodificación, pero incluye un decodificador de excitación que es diferente al decodificador de excitación 222 en la segunda rama de decodificación 222, 224. En una realización de este tipo, la segunda rama comprende la etapa de LPC 224 y un decodificador de excitación basado en un libro de códigos tal como en el ACELP, y la tercera rama comprende una etapa de LPC y un decodificador de excitación que funciona sobre una representación espectral de la señal de salida de la etapa de LPC 224.

En otra realización, la etapa pre-procesamiento común comprende un bloque de estéreo envolvente/conjunto que genera, como una salida, parámetros de estéreo conjunto y una señal mono de salida, que se genera mezclando de manera descendente la señal de entrada que es una señal que tiene dos o más canales. En general, la señal en la salida del bloque también puede ser una señal que tiene más canales, pero debido a la operación de mezcla descendente, el número de canales en la salida del bloque será menor que el número de canales introducidos en el bloque. En esta realización, la rama de codificación de frecuencia comprende una etapa de conversión espectral y una etapa conectada subsiguientemente de cuantificación/ codificación. La etapa de cuantificación/codificación puede incluir cualquiera de las funcionalidades que se conocen de los codificadores modernos del dominio de la frecuencia tal como el codificador de AAC. Adicionalmente, se puede controlar la operación de cuantificación en la etapa de cuantificación/codificación mediante un módulo psicoacústico que genera información psicoacústica tal como un umbral de enmascaramiento psicoacústico sobre la frecuencia donde esta información se introduce en esta etapa. Preferentemente, la conversión espectral se hace usando una operación de MDCT que, incluso más preferentemente, es la función de MDCT deformada en el tiempo, donde la intensidad, o, en general, la intensidad de deformación, puede controlarse entre cero y una alta intensidad de deformación. En una intensidad de deformación cero, la operación de MDCT es una operación de MDCT sencilla que es conocida en la materia. El codificador del dominio de LPC puede incluir un núcleo de ACELP que calcula una ganancia de tono, un retraso de tono y/o una información de libro de códigos tal como un índice de libro de códigos y una ganancia de códigos.

A pesar de que algunas figuras muestran diagramas de bloques de un aparato, se observa que estas figuras, al mismo tiempo muestran un método, en el que las funcionalidades del bloque corresponden a las etapas de método.

Las realizaciones de la invención se describieron anteriormente basándose en una señal de entrada de audio que comprende diferentes segmentos o tramas, asociándose los diferentes segmentos o tramas con información de voz o información de música. La invención no está limitada a tales realizaciones, en su lugar el enfoque para clasificar diferentes segmentos de una señal que comprende segmentos de al menos un primer tipo y un segundo tipo se puede aplicar también a señales de audio que comprenden tres o más tipos de segmentos diferentes, cada uno de los cuales se desea codificar con diferentes esquemas de codificación. Ejemplos para tales tipos de segmentos son:

- Segmentos estacionarios y no estacionarios pueden ser útiles para el uso de diferentes bancos de filtros, ventanas o adaptaciones de codificación. Por ejemplo, un transitorio se debe codificar con un banco de filtros de una resolución de tiempo precisa, mientras una sinusoidal pura se debe codificar con un banco de filtros de una resolución de frecuencia precisa.
- Vocalizados/no vocalizados: los segmentos vocalizados están bien tratados con un codificador de voz como CELP, pero para los segmentos no vocalizados se desperdician demasiados bits. La codificación paramétrica será más eficiente.
- Silencio/activo: puede codificarse silencio con menos bits que segmentos activos.
- armónico/no armónico: será beneficios utilizar para la codificación de segmentos armónicos una predicción lineal en el dominio de la frecuencia.

Además la invención no está limitado al campo de las técnicas de audio, más bien el enfoque anteriormente descrito para clasificar una señal se puede aplicar a otros tipos de señales, como señales de vídeo o señales de datos, en el que estas respectivas señales incluyen segmentos de diferentes tipos que requieren un procesamiento diferente, como por ejemplo:
La presente invención puede adaptarse para todas las aplicaciones en tiempo real que requieren una segmentación de una señal de tiempo. Por ejemplo, el reconocimiento de una cara desde una cámara de video de vigilancia puede basarse en un clasificador que determina para cada píxel de una trama (aquí una trama corresponde a una imagen tomada en un momento de tiempo n) si pertenece a la cara de una persona o no. La clasificación (es decir, la

segmentación de la cara) se debe hacer para cada trama individual del flujo de video. Sin embargo, utilizando la presente invención, la segmentación de la trama actual puede tener en cuenta las sucesivas tramas pasadas para obtener una mejor precisión de la segmentación aprovechando la ventaja que las sucesivas imágenes están fuertemente correlacionadas. Entonces se pueden aplicar dos clasificadores. Uno considera sólo la trama actual y otro que considera un conjunto de tramas incluyendo la trama actual y las tramas pasadas. El último clasificador puede integrar un conjunto de tramas y determinar la región de probabilidad para la posición de la cara. La decisión del clasificador que se hace sólo sobre la trama actual, se comparará a continuación a las regiones de probabilidad. A continuación la decisión puede validarse o modificarse.

Las realizaciones de la invención utilizan un conmutador para conmutar entre ramas de modo que sólo una rama reciba una señal a procesarse y que la otra rama no reciba la señal. Sin embargo, en una realización alternativa el conmutador puede también estar dispuesto después de las etapas de procesamiento o ramas, por ejemplo, el codificador de audio y el codificador de voz, de modo que ambas ramas procesen la misma señal en paralelo. Se selecciona la señal emitida por una de estas ramas para emitirse, por ejemplo, para escribirse en un flujo de bits de salida.

Mientras algunas realizaciones de la invención se describieron basándose en señales digitales, en las cuales se determinaron los segmentos mediante una cantidad predeterminada de muestras obtenidas con una tasa de muestreo específica, la invención no está limitada a esas señales, más bien, se puede aplicar también a señales analógicas en las cuales se determinaría el segmento mediante un rango de frecuencia específico o un periodo de tiempo específico de la señal analógica. Además, algunas realizaciones de la invención se describieron en combinación con codificadores que incluyen un discriminador. Se observa que, básicamente, el enfoque de acuerdo con las realizaciones de la invención para clasificar señales se puede aplicar también a decodificadores que reciben una señal codificada, para la que se pueden clasificar diferentes esquemas de codificación, permitiendo de ese modo que se suministre la señal codificada a un decodificador apropiado.

Dependiendo de ciertos requisitos de implementación de los métodos inventivos, los métodos inventivos se pueden implementar mediante hardware o software. Se puede llevar a cabo la implementación utilizando un medio de almacenamiento digital, en particular, un disco, un DVD o un CD, que tiene almacenadas en el mismo señales de control electrónicamente legibles, las cuales cooperan con sistemas informáticos programables de modo que se llevan a cabo los métodos inventivos. Por lo tanto, la presente invención es, por lo tanto, un producto de programa informático con un código de programa almacenado en un portador legible por máquina, operándose el código de programa para llevar a cabo los métodos inventivos, cuando se ejecuta el producto de programa informático en un ordenador. En otras palabras, los métodos inventivos son, por lo tanto, un programa informático que tiene un código de programa para llevar a cabo al menos uno de los métodos inventivos cuando se ejecuta el programa informático en un ordenador.

Las realizaciones anteriormente descritas son meramente ilustrativas de los principios de la presente invención. Se entiende que las modificaciones y variaciones posibles de las disposiciones y de los detalles descritos en el presente documento serán evidentes para los expertos en la materia. Por lo tanto, es la intención que la invención esté limitada sólo por el alcance de las siguientes reivindicaciones de patente y no por los detalles específicos presentados a modo de descripción y explicación de las realizaciones del presente documento.

En las realizaciones anteriores, la señal se describe como que comprende una pluralidad de tramas, en el que se evalúa una trama actual para una decisión de conmutación. Se observa que el segmento actual de la señal que se está evaluando para una decisión de conmutación puede ser una trama, sin embargo, la invención no está limitada a tales realizaciones. Más bien, un segmento de la señal también puede comprender una pluralidad, es decir dos o más tramas.

Además, en las realizaciones anteriormente descritas, tanto el clasificador a corto plazo como el clasificador a largo plazo utilizan el mismo rasgo distintivo o los mismos rasgos distintivos. Este enfoque se puede utilizar por distintas razones, como la necesidad de calcular los rasgos distintivos a corto plazo sólo una vez y aprovechar el mismo por los dos clasificadores de distintas maneras que reducirá la complejidad del sistema, como por ejemplo, el rasgo distintivo a corto plazo puede calcularse por uno de los clasificadores a corto plazo y a largo plazo y se proporciona al otro clasificador. También, la comparación entre los resultados de los clasificadores a corto plazo y a largo plazo puede ser más importante, ya que se puede deducir más fácilmente la contribución de la trama actual en el resultado de clasificación a largo plazo comparándolo con el resultado de clasificación a corto plazo, debido a que los dos clasificadores comparten rasgos distintivos comunes

Sin embargo, la invención no se restringe a este enfoque y el clasificador a largo plazo no se restringe al uso del mismo rasgo distintivo o rasgos distintivos que el clasificador a corto plazo, es decir tanto el clasificador a corto plazo como el clasificador a largo plazo pueden calcular su respectivo rasgo distintivo o rasgos distintivos a corto plazo que son diferentes entre sí.

Mientras las realizaciones anteriormente descritas mencionan el uso de los PLPCC como rasgo distintivo a corto plazo, se observa que se pueden considerar otros rasgos distintivos, por ejemplo la variabilidad de los PLPCC.

REIVINDICACIONES

1. Un método para clasificar diferentes segmentos de una señal de audio, comprendiendo la señal de audio segmentos de voz y de música, comprendiendo el método:

clasificar a corto plazo, por un clasificador a corto plazo (150), la señal de audio usando al menos un rasgo distintivo a corto plazo extraído de la señal de audio y entregar un resultado de clasificación a corto plazo (152) que indica si un segmento actual de la señal de audio es un segmento de voz o un segmento de música;
clasificar a largo plazo, por un clasificador a largo plazo (154), la señal de audio usando al menos un rasgo distintivo a corto plazo y al menos un rasgo distintivo a largo plazo extraídos de la señal de audio y entregar un resultado de clasificación a largo plazo (156) que indica si el segmento actual de la señal de audio es un segmento de voz o un segmento de música; y
aplicar el resultado de clasificación a corto plazo y el resultado de clasificación a largo plazo a un circuito de decisión (158) acoplado a una salida del clasificador a corto plazo (150) y a una salida del clasificador a largo plazo (154), combinando el circuito de decisión (158) el resultado de clasificación a corto plazo (152) y el resultado de clasificación a largo plazo (156) para proporcionar una señal de salida (160) que indica si el segmento actual de la señal de audio es un segmento de voz o un segmento de música.

2. El método de la reivindicación 1, en el que la etapa de combinación comprende proporcionar la señal de salida basándose en una comparación del resultado de clasificación a corto plazo (152) con el resultado de clasificación a largo plazo (156).

3. El método de la reivindicación 1 o 2, en el que
el al menos un rasgo distintivo a corto plazo se obtiene analizando el segmento actual de la señal de audio que se va a clasificar; y
el al menos un rasgo distintivo a largo plazo se obtiene analizando el segmento actual de la señal de audio y uno o más segmentos anteriores de la señal de audio.

4. El método de una de las reivindicaciones 1 a 3, en el que
el al menos un rasgo distintivo a corto plazo se obtiene analizando una ventana de análisis (168) de una primera longitud y un primer método de análisis; y
el al menos un rasgo distintivo a largo plazo se obtiene analizando una ventana de análisis (162) de una segunda longitud y un segundo método de análisis, siendo la primera longitud más corta que la segunda longitud, y siendo el primer y segundo métodos de análisis diferentes.

5. El método de la reivindicación 4, en el que la primera longitud se extiende a lo largo del segmento actual de la señal de audio, la segunda longitud se extiende a lo largo del segmento actual de la señal de audio y uno o más segmentos anteriores de la señal de audio, y la primera y segunda longitudes comprenden un periodo adicional (164) que cubre un periodo de análisis.

6. El método de una de las reivindicaciones 1 a 5, en el que combinar el resultado de clasificación a corto plazo (152) con el resultado de clasificación a largo plazo (156) comprende una decisión de histéresis basándose en un resultado combinado, en el que el resultado combinado comprende el resultado de clasificación a corto plazo (152) y el resultado de clasificación a largo plazo (156), cada uno ponderado por un factor de ponderación predeterminado.

7. El método de una de las reivindicaciones 1 a 6, en el que la señal de audio es una señal digital y un segmento de la señal de audio comprende un número predefinido de muestras obtenido a una tasa de muestreo específica.

8. El método de una de las reivindicaciones 1 a 7, en el que
el al menos un rasgo distintivo a corto plazo comprende parámetros de coeficiente cepstral de predicción lineal perceptiva PLPCC; y
el al menos un rasgo distintivo a largo plazo comprende información característica de tono.

9. El método de una de las reivindicaciones 1 a 8, en el que el al menos un rasgo distintivo a corto plazo usado para la clasificación a corto plazo y el al menos un rasgo distintivo a corto plazo usado para la clasificación a largo plazo son los mismos o diferentes.

10. Un método para procesar una señal de audio que comprende segmentos de al menos un primer tipo y un segundo tipo, comprendiendo el método:

clasificar (116) un segmento de la señal de audio de acuerdo con el método de una de las reivindicaciones 1 a 9;
procesar (102; 206; 106; 208) el segmento de acuerdo con un primer proceso o un segundo proceso, dependiendo de la señal de salida (160) proporcionada por la etapa de clasificación (116); y
emitir el segmento procesado.

11. El método de la reivindicación 10, en el que
el segmento se procesa por un codificador de voz (102) cuando la señal de salida (160) indica que el segmento es un segmento de voz; y
el segmento se procesa por un codificador de música (106) cuando la señal de salida (160) indica que el segmento es un segmento de música.

12. El método de la reivindicación 11, que comprende además:

combinar (108) el segmento codificado e información de la señal de salida (160) que indica el tipo del segmento.

13. Un programa informático para realizar, cuando se ejecuta en un ordenador, el método de una de las reivindicaciones 1 a 12.

14. Un discriminador que comprende:

un clasificador a corto plazo (150) configurado para recibir una señal de audio y proporcionar un resultado de clasificación a corto plazo (152) que indica si un segmento actual de la señal de audio es un segmento de voz o un segmento de música usando al menos un rasgo distintivo a corto plazo extraído de la señal de audio, comprendiendo la señal de audio segmentos de voz y segmentos de música;
un clasificador a largo plazo (154) configurado para recibir la señal de audio y proporcionar un resultado de clasificación a largo plazo (156) que indica si el segmento actual de la señal de audio es un segmento de voz o un segmento de música usando al menos un rasgo distintivo a corto plazo y al menos un rasgo distintivo a largo plazo extraídos de la señal de audio; y
un circuito de decisión (158) acoplado a una salida del clasificador a corto plazo (150) y a una salida del clasificador a largo plazo (154), para recibir el resultado de clasificación a corto plazo (152) y el resultado de clasificación a largo plazo (156), el circuito de decisión (158) configurado para combinar el resultado de clasificación a corto plazo (152) y el resultado de clasificación a largo plazo (156) para proporcionar una señal de salida (160) que indica si el segmento actual de la señal de audio es un segmento de voz o un segmento de música.

15. El discriminador de la reivindicación 14, en el que el circuito de decisión (158) está configurado para proporcionar la señal de salida basándose en una comparación del resultado de clasificación a corto plazo (152) con el resultado de clasificación a largo plazo (156).

16. Un aparato de procesamiento de señal, que comprende:

una entrada (110) configurada para recibir una señal de audio a procesarse, en el que la señal de audio comprende segmentos de voz y segmentos de música;
una primera etapa de procesamiento (102; 206) configurada para procesar segmentos de voz;
una segunda etapa de procesamiento (104; 208) configurada para procesar segmentos de música;
un discriminador (116; 204) de la reivindicación 14 o 15 acoplado a la entrada (110); y
y un dispositivo de conmutación (112; 202) acoplado entre la entrada (110) y la primera y segunda etapas de procesamiento (102, 104; 206, 208) y configurado para aplicar la señal de audio desde la entrada (110) a una de la primera y segunda etapas de procesamiento (102, 104; 206, 208) dependiendo de la señal de salida (160) del discriminador (116).

17. Un codificador de audio, que comprende un aparato de procesamiento de señal de la reivindicación 16.

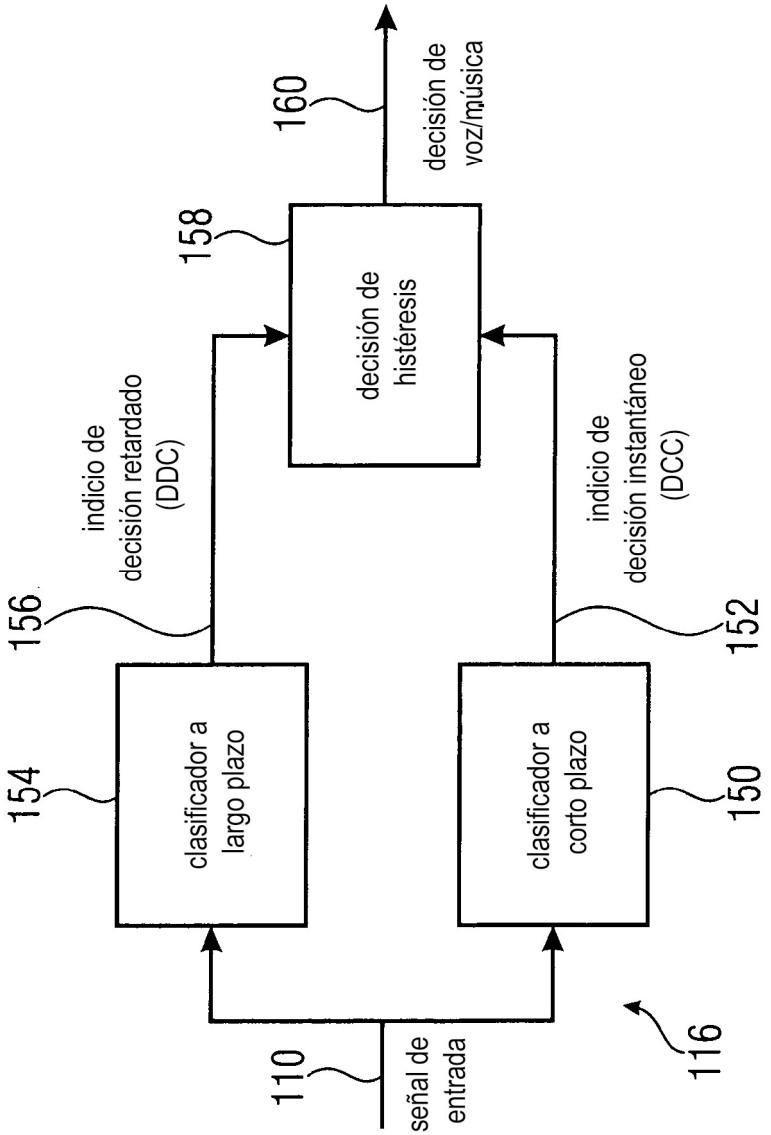
160
decisión de voz/música
158
decisión de histéresis
indicio de decisión retardado (DDC)
156
indicio de decisión instantáneo (DCC)
152
154
clasificador a largo plazo
clasificador a corto plazo
150
110
señal de entrada
116

FIGURA 1

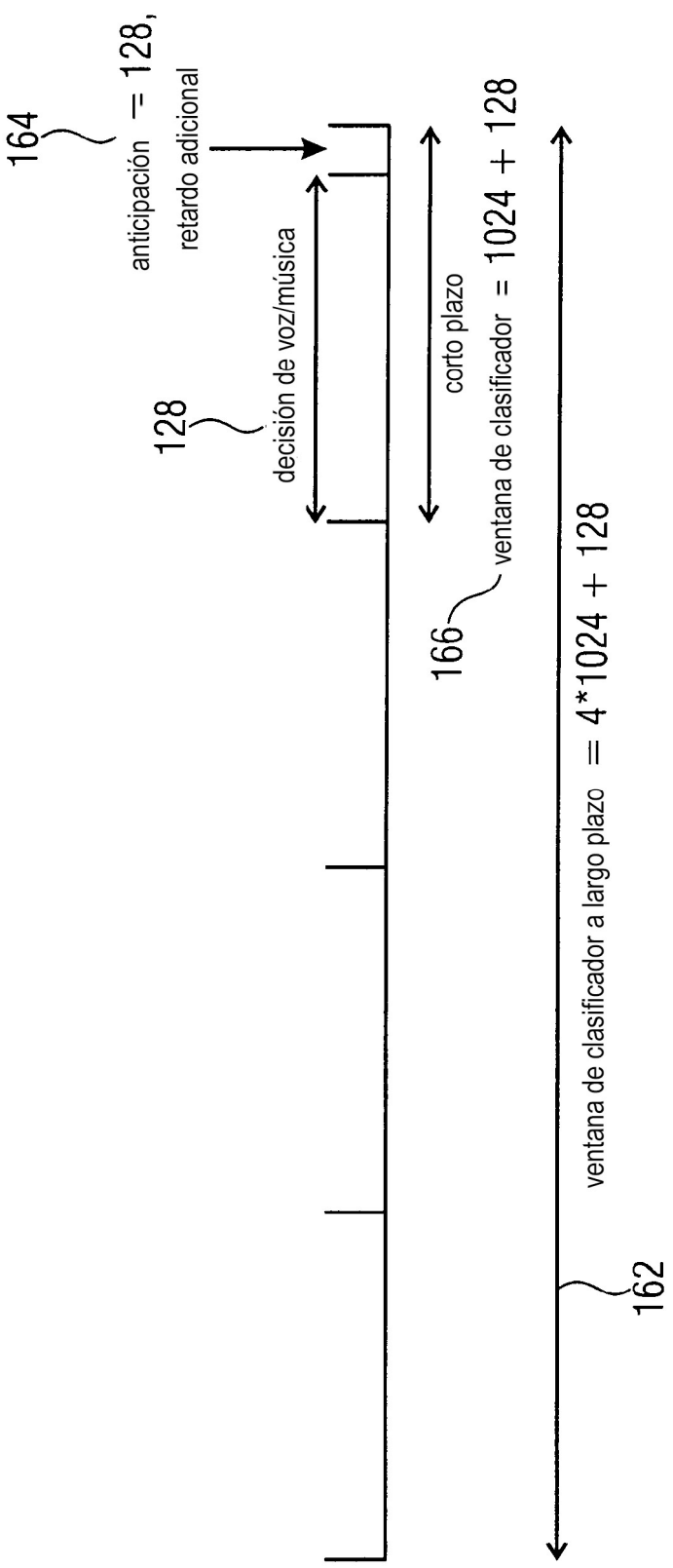
164
anticipación = 128,
retardo adicional
128
decisión de voz/música
corto plazo
ventana de clasificador = 1024 + 128
166
ventana de clasificador a largo plazo = 4*1024 + 128
162

FIGURA 2

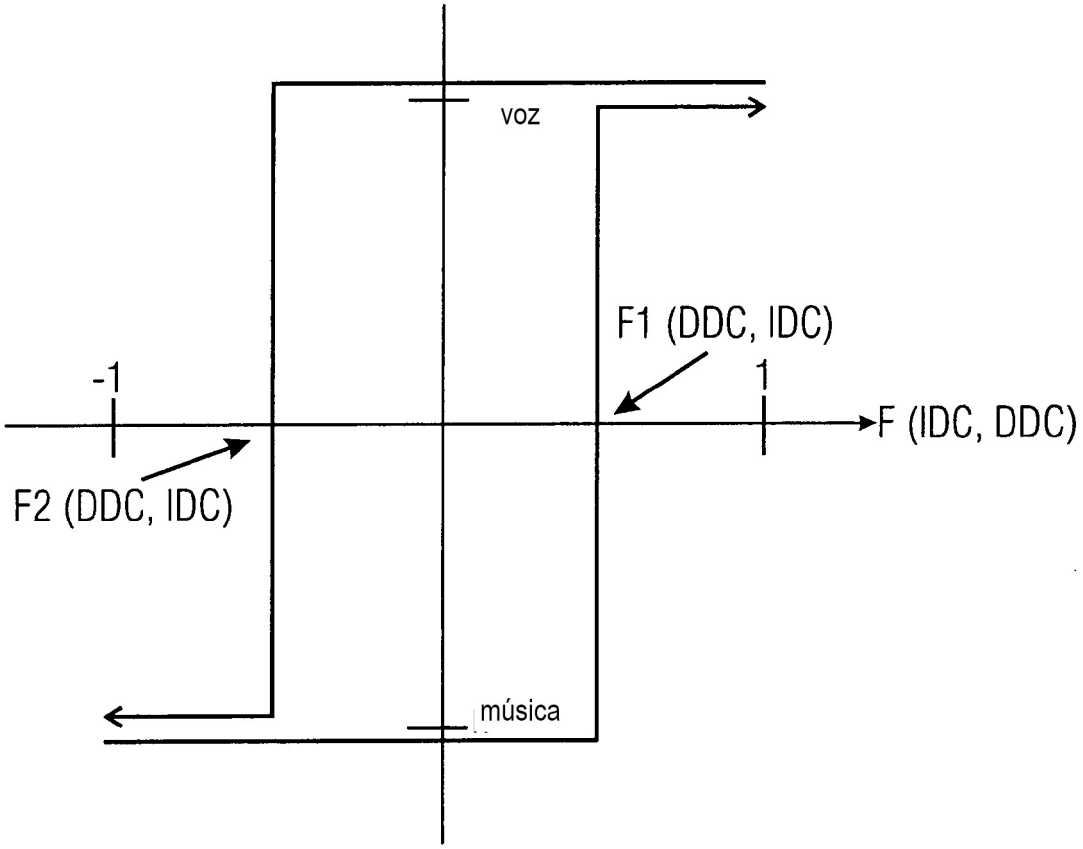

FIGURA 3

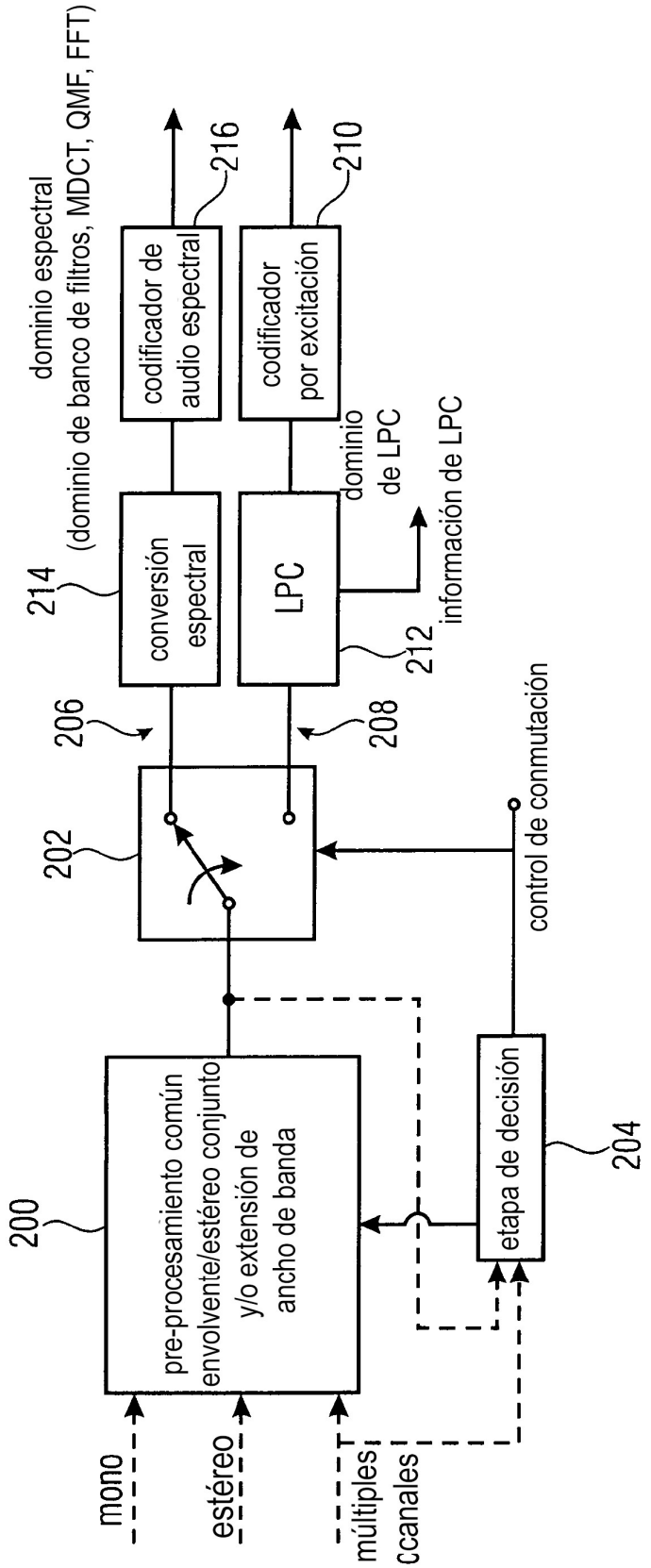

FIGURA 4
(CODIFICADOR)

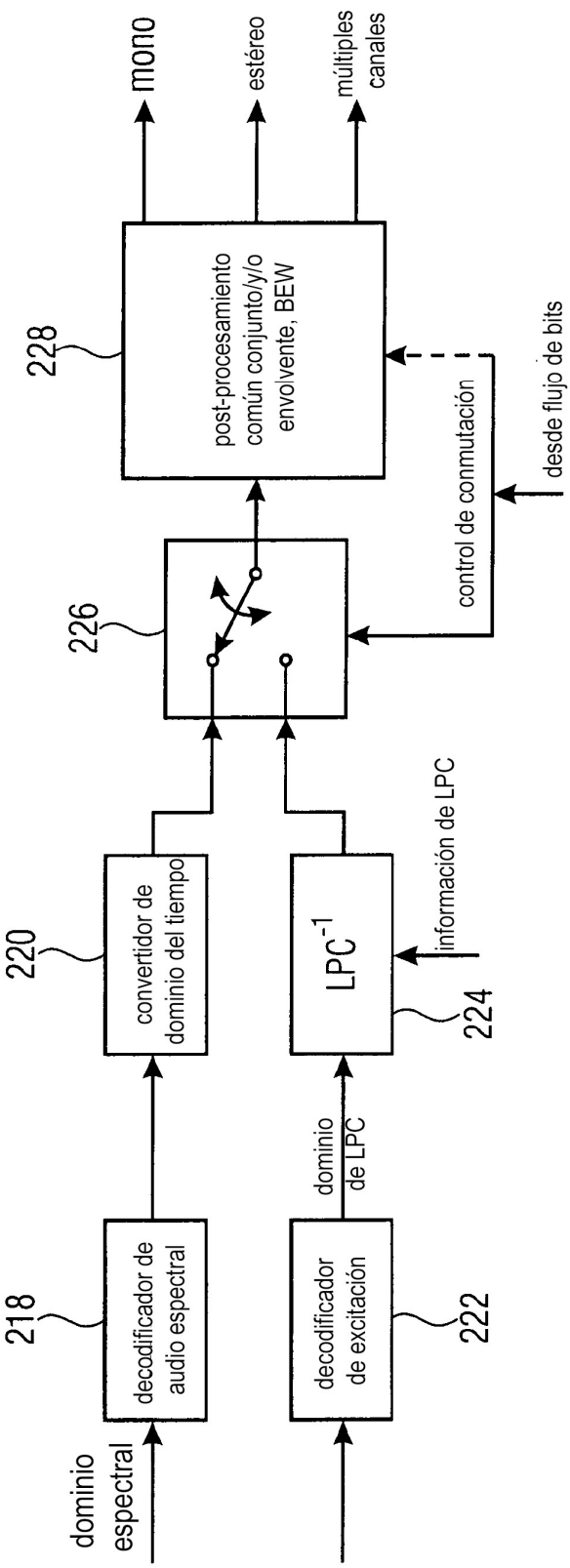

FIGURA 5
(DECODIFICADOR)

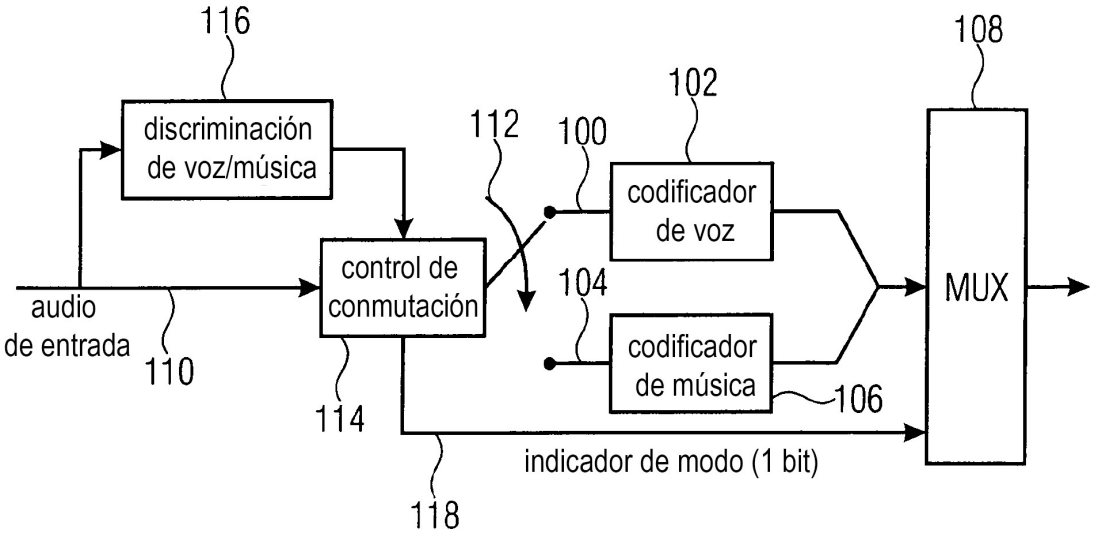

FIGURA 6

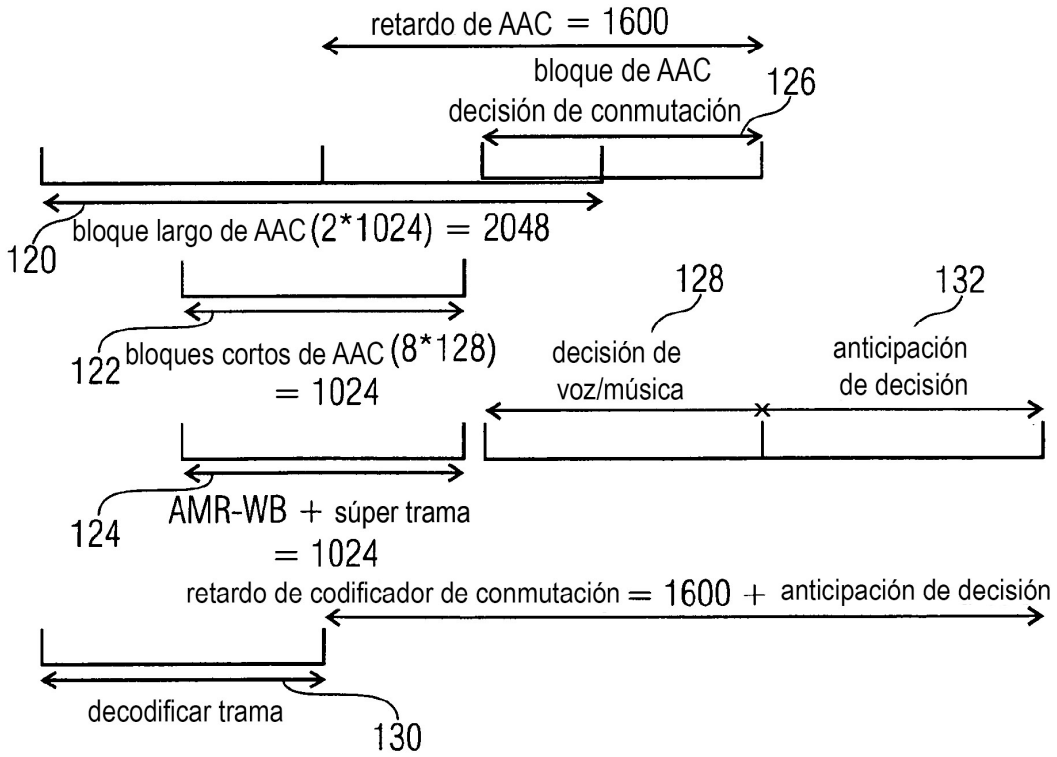

FIGURA 7