(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 687 249**

(51) Int. CI.:

***G10L 25/93*** (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **05.09.2014 PCT/CN2014/086058**

(87) Fecha y número de publicación internacional: **12.03.2015 WO15032351**

(96) Fecha de presentación y número de la solicitud europea: **05.09.2014 E 14842028 (4)**

(97) Fecha y número de publicación de la concesión europea: **11.07.2018 EP 3005364**

(54) Título: **Decisión no sonora/sonora para el procesamiento de la voz**

(30) Prioridad:

**09.09.2013 US 201361875198 P**
**03.09.2014 US 201414476547**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
**24.10.2018**

(73) Titular/es:

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

(72) Inventor/es:

**GAO, YANG**

(74) Agente/Representante:

**LEHMANN NOVO, María Isabel**

## DESCRIPCIÓN

Decisión no sonora/sonora para el procesamiento de la voz.

Campo técnico

La presente invención se refiere, en general, al campo del procesamiento de la voz y, en particular, a la Decisión Sonora/No Sonora para el procesamiento de la voz.

Antecedentes

La codificación de la voz se refiere a un proceso que reduce la velocidad binaria de un archivo de voz. La codificación de la voz es una aplicación de compresión de datos de señales de audio digital que contienen voz. La codificación de la voz usa una estimación de parámetros específicos para la voz mediante el uso de técnicas de procesamiento de señales de audio para modelar la señal de voz, combinadas con algoritmos de compresión de datos genéricos para representar los parámetros modelados resultantes en un tren de bits compacto. El objetivo de la codificación de la voz es lograr ahorros en el espacio de almacenamiento de memoria requerido, ancho de banda de transmisión y potencia de transmisión mediante la reducción del número de bits por muestra de modo que la voz decodificada (descomprimida) es perceptualmente indistinguible de la voz original.

Sin embargo, los codificadores de voz son codificadores con pérdidas, a saber, la señal decodificada es diferente de la original. Por lo tanto, uno de los objetivos de la codificación de la voz es minimizar la distorsión (o pérdida perceptible) a una velocidad binaria dada, o minimizar la velocidad binaria para alcanzar una distorsión dada.

La codificación de la voz difiere de otras formas de codificación de audio en que la voz es una señal mucho más simple que la mayoría de las otras señales de audio e información más estadística se encuentra disponible sobre las propiedades de la voz. Como resultado, cierta información auditiva que es relevante en la codificación de audio puede ser innecesaria en el contexto de la codificación de la voz. En la codificación de la voz, el criterio más importante es la preservación de la inteligibilidad y "agrado" de la voz, con una cantidad limitada de datos transmitidos.

La inteligibilidad de la voz incluye, además del contenido literal real, también la identidad del hablante, emociones, entonación, timbre, etc. que son todos importantes para una inteligibilidad perfecta. El concepto más abstracto de agrado de la voz degradada es una propiedad diferente de la inteligibilidad, dado que es posible que la voz degradada sea completamente inteligible pero subjetivamente desagradable para el oyente.

La redundancia de formas de onda de la voz se puede considerar con respecto a varios tipos diferentes de señal de voz como, por ejemplo, señales de voz sonora y no sonora. Los sonidos sonoros, p.ej., "a", "b", se deben, esencialmente, a las vibraciones de las cuerdas vocales, y son oscilatorias. Por lo tanto, durante períodos cortos, se modelan bien por las sumas de señales periódicas como, por ejemplo, sinusoides. En otras palabras, para la voz sonora, la señal de voz es esencialmente periódica. Sin embargo, dicha periodicidad puede ser variable a lo largo de la duración de un segmento de voz y la forma de la onda periódica cambia, en general, de forma gradual de segmento a segmento. Una codificación de la voz de baja velocidad binaria se puede beneficiar ampliamente de la exploración de dicha periodicidad. El período de voz sonora se conoce también como altura y la predicción de altura se conoce, con frecuencia, como Predicción a Largo Plazo (LTP, por sus siglas en inglés). Por el contrario, los sonidos no sonoros como, por ejemplo, "s", "sh", son más tipo ruido. Ello se debe a que la señal de voz no sonora es más como un ruido aleatorio y tiene una cantidad más pequeña de predictibilidad.

Tradicionalmente, todos los métodos paramétricos de codificación de la voz usan la redundancia inherente a la señal de la voz para reducir la cantidad de información que se debe enviar y para estimar los parámetros de muestras de voz de una señal en intervalos cortos. Dicha redundancia surge, principalmente, de la repetición de formas de onda de voz a una tasa cuasiperiódica y la baja envolvente espectral cambiante de la señal de la voz.

La redundancia de formas de onda de la voz se puede considerar con respecto a varios tipos diferentes de señal de voz como, por ejemplo, sonora y no sonora. Aunque la señal de voz es, esencialmente, periódica para la voz sonora, dicha periodicidad puede ser variable a lo largo de la duración de un segmento de voz y la forma de la onda periódica cambia, normalmente, de forma gradual de segmento a segmento. Una codificación de la voz de baja velocidad binaria se puede beneficiar ampliamente de la exploración de dicha periodicidad. El período de voz sonora se conoce también como altura y la predicción de altura se llama, con frecuencia, Predicción a Largo Plazo (LTP). En cuanto a la voz no sonora, la señal es más como un ruido aleatorio y tiene una cantidad más pequeña de predictibilidad.

En cualquier caso, la codificación paramétrica se puede usar para reducir la redundancia de los segmentos de voz mediante la separación del componente de excitación de la señal de voz del componente de envolvente espectral. La envolvente espectral que cambia lentamente se puede representar por la Codificación de Predicción Lineal (LPC, por sus siglas en inglés), también llamada Predicción a Corto Plazo (STP, por sus siglas en inglés). Una codificación

de la voz de baja velocidad binaria se puede beneficiar mucho también de la exploración de dicha Predicción a Corto Plazo. La ventaja de la codificación surge de la tasa lenta a la que cambian los parámetros. Sin embargo, es raro que los parámetros sean significativamente diferentes de los valores contenidos dentro de unos pocos milisegundos. Por consiguiente, a la velocidad de muestreo de 8 kHz, 12,8 kHz o 16 kHz, el algoritmo de codificación de la voz es tal que la duración de trama nominal se encuentra en el rango de los diez a treinta milisegundos. La duración de la trama de veinte milisegundos es la elección más común.

En estándares conocidos más recientes como, por ejemplo, G.723.1, G.729, G.718, Velocidad Total Mejorada (EFR, por sus siglas en inglés), Vocodificador de Modo Seleccionable (SMV, por sus siglas en inglés), Multivelocidad Adaptativa (AMR, por sus siglas en inglés), Banda Ancha Multimodo de Velocidad Variable (VMR-WB, por sus siglas en inglés), o Banda Ancha Multivelocidad Adaptativa (AMR-WB, por sus siglas en inglés), Técnica de Predicción Lineal Excitada por Código (CELP, por sus siglas en inglés) se han adoptado. CELP se entiende comúnmente como una combinación técnica de Excitación por Código, Predicción a Largo Plazo y Predicción a Corto Plazo. CELP se usa, principalmente, para codificar una señal de voz beneficiándose de las características específicas de la voz humana o de un modelo humano de producción de voz vocal. La Codificación de la Voz CELP es un principio de algoritmo muy popular en el área de compresión de la voz aunque los detalles de CELP para diferentes códecs pueden ser significativamente diferentes. Debido a su popularidad, el algoritmo CELP se ha usado en varios estándares ITU-T, MPEG, 3GPP y 3GPP2. Las variantes de CELP incluyen CELP algebraica, CELP relajada, CELP con bajo retardo y predicción lineal excitada por la suma del vector, y otros. CELP es un término genérico para una clase de algoritmos y no para un códec particular.

El algoritmo CELP se basa en cuatro ideas principales. Primero, se usa un modelo de filtro de la fuente de la producción de la voz a través de la predicción lineal (PL). El modelo de filtro de la fuente de la producción de la voz modela la voz como una combinación de una fuente de sonido como, por ejemplo, las cuerdas vocales, y un filtro acústico lineal, el tracto vocal (y característica de radiación). En la implementación del modelo de filtro de la fuente de la producción de la voz, la fuente de sonido, o señal de excitación, se modela, con frecuencia, como un tren de impulsos periódico, para la voz sonora, o ruido blanco para la voz no sonora. Segundo, un libro de códigos adaptativo y fijo se usa como la entrada (excitación) del modelo PL. Tercero, se lleva a cabo una búsqueda en bucle cerrado en un "dominio perceptualmente ponderado". Cuarto, se aplica la cuantificación vectorial (CV).

El documento WO2008151408 A1 describe un algoritmo de clasificación de señales según una función de mérito que se calcula como una suma ponderada de los siguientes parámetros: coherencia de altura, velocidad de cruce por cero, correlación normalizada máxima, inclinación espectral y diferencia de energía.

El documento US20050177364 A1 describe la clasificación de tramas de voz no sonora según al menos tres de los parámetros, a saber: la medida de sonoridad, la inclinación espectral, la variación de energía dentro de una trama y la energía de trama relativa.

El documento US6453285B1 describe una voz o detector de actividad de voz (VAD, por sus siglas en inglés) para detectar si las señales de voz están presentes en tramas de tiempo individuales de una señal de entrada.

El documento WO2007073604A1 describe la clasificación de señales en el decodificador mediante el uso de los siguientes parámetros: correlación normalizada, una medida de inclinación espectral, contador de estabilidad de altura, energía de trama relativa de la señal al final de la trama actual, y contador de cruce por cero.

Compendio

La invención se define en las reivindicaciones independientes anexas. Realizaciones preferidas adicionales se definen en las reivindicaciones dependientes.

En una realización alternativa, un aparato de procesamiento de la voz comprende un procesador y un medio de almacenamiento legible por ordenador que almacena la programación para la ejecución por el procesador. La programación incluye instrucciones para determinar un parámetro de no sonoridad que refleja una característica de la voz no sonora en una trama actual de una señal de voz que comprende múltiples tramas, y para determinar un parámetro de no sonoridad suavizado para incluir información del parámetro de no sonoridad en una trama anterior a la trama actual de la señal de voz. La programación además incluye instrucciones para computar una diferencia entre el parámetro de no sonoridad y el parámetro de no sonoridad suavizado, y generar un punto de decisión no sonoro/sonoro para determinar si la trama actual comprende voz no sonora o voz sonora mediante el uso de la diferencia computada como un parámetro de decisión.

En una realización alternativa, un método para el procesamiento de la voz comprende proveer múltiples tramas de una señal de voz y determinar, para una trama actual, un primer parámetro para una primera banda de frecuencia de una primera envolvente de energía de la señal de voz en el dominio temporal y un segundo parámetro para una segunda banda de frecuencia de una segunda envolvente de energía de la señal de voz en el dominio temporal. Un primer parámetro suavizado y un segundo parámetro suavizado se determinan a partir de las tramas previas de la señal de voz. El primer parámetro se compara con el primer parámetro suavizado y el segundo parámetro se

compara con el segundo parámetro suavizado. Un punto de decisión no sonoro/sonoro se genera para determinar si la trama actual comprende voz no sonora o voz sonora mediante el uso de la comparación como un parámetro de decisión.

Breve descripción de los dibujos

Para una comprensión más completa de la presente invención, y de sus ventajas, ahora se hace referencia a las siguientes descripciones tomadas en conjunto con los dibujos anexos, en los cuales:

La Figura 1 ilustra una evaluación de energía del dominio temporal de una señal de voz de banda de frecuencia baja según realizaciones de la presente invención;

la Figura 2 ilustra una evaluación de energía del dominio temporal de una señal de voz de banda de frecuencia alta según realizaciones de la presente invención;

la Figura 3 ilustra funciones llevadas a cabo durante la codificación de una voz original mediante el uso de un codificador CELP convencional mediante la implementación de una realización de la presente invención;

la Figura 4 ilustra funciones llevadas a cabo durante la decodificación de una voz original mediante el uso de un decodificador CELP convencional mediante la implementación de una realización de la presente invención;

la Figura 5 ilustra un codificador CELP convencional usado en la implementación de las realizaciones de la presente invención;

la Figura 6 ilustra un decodificador CELP básico correspondiente al codificador en la Figura 5 según una realización de la presente invención;

la Figura 7 ilustra vectores candidatos tipo ruido para construir el libro de códigos de excitación codificada o libro de códigos fijo de la codificación de voz CELP;

la Figura 8 ilustra vectores candidatos tipo pulso para construir el libro de códigos de excitación codificada o libro de códigos fijo de la codificación de voz CELP;

la Figura 9 ilustra un ejemplo de espectro de excitación para la voz sonora;

la Figura 10 ilustra un ejemplo de espectro de excitación para la voz no sonora;

la Figura 11 ilustra un ejemplo de espectro de excitación para la señal de ruido de fondo;

las Figuras 12A y 12B ilustran ejemplos de codificación/decodificación de dominio de la frecuencia con extensión de ancho de banda, en donde la Figura 12A ilustra el codificador con información conexa BWE mientras la Figura 12B ilustra el decodificador con BWE;

las Figuras 13A-13C describen funciones de procesamiento de voz según varias realizaciones descritas más arriba;

la Figura 14 ilustra un sistema de comunicación 10 según una realización de la presente invención; y

la Figura 15 ilustra un diagrama de bloques de un sistema de procesamiento que puede usarse para implementar los dispositivos y métodos descritos en la presente memoria.

Descripción detallada de realizaciones ilustrativas

En el sistema de comunicación de señales digitales de audio/voz moderno, una señal digital se comprime en un codificador y la información comprimida o tren de bits pueden paquetizarse y enviarse a un decodificador trama por trama a través de un canal de comunicación. El decodificador recibe y decodifica la información comprimida para obtener la señal digital de audio/voz.

Con el fin de codificar la señal de voz de manera más eficaz, la señal de voz se puede clasificar en diferentes clases y cada clase se codifica de manera diferente. Por ejemplo, en algunos estándares como, por ejemplo, G.718, VMR-WB o AMR-WB, una señal de voz se clasifica en NO SONORA, DE TRANSICIÓN, GENÉRICA, SONORA y RUIDO.

La señal de voz sonora es un tipo de señal cuasiperiódica que, normalmente, tiene más energía en el área de frecuencia baja que en el área de frecuencia alta. Por el contrario, la señal de voz no sonora es una señal tipo ruido que, normalmente, tiene más energía en el área de frecuencia alta que en el área de frecuencia baja. La clasificación No Sonora/Sonora o Decisión No Sonora se usa ampliamente en el campo de la codificación de señales de voz, extensión de ancho de banda (BWE, por sus siglas en inglés) de señal de voz, mejora de señal de voz y reducción de ruido (NR, por sus siglas en inglés) de fondo de señal de voz.

En la codificación de la voz, la señal de voz no sonora y la señal de voz sonora pueden codificarse/decodificarse de manera diferente. En la extensión de ancho de banda de señal de voz, la energía de señal de banda alta extendida de la señal de voz no sonora puede controlarse de manera diferente de la de la señal de voz sonora. En la reducción de ruido de fondo de señal de voz, el algoritmo NR puede ser diferente para la señal de voz no sonora y señal de voz sonora. Entonces, una Decisión No Sonora robusta es importante para los tipos de aplicaciones de más arriba.

Las realizaciones de la presente invención mejoran la exactitud de la clasificación de una señal de audio como una señal sonora o una señal no sonora anterior a la codificación de la voz, extensión de ancho de banda y/o funciones de mejora de voz. Por lo tanto, las realizaciones de la presente invención pueden aplicarse a la codificación de señal de voz, extensión de ancho de banda de señal de voz, mejora de señal de voz y reducción de ruido de fondo de señal de voz. En particular, las realizaciones de la presente invención pueden usarse para mejorar el estándar del codificador de voz ITU-T AMR-WB en la extensión de ancho de banda.

Una ilustración de las características de la señal de voz usadas para mejorar la exactitud de la clasificación de la señal de audio en señal sonora o señal no sonora según las realizaciones de la presente invención se ilustrará mediante el uso de las Figuras 1 y 2. La señal de voz se evalúa en dos regímenes: una banda de frecuencia baja y una banda de frecuencia alta en las ilustraciones de más abajo.

La Figura 1 ilustra una evaluación de energía del dominio temporal de una señal de voz de banda de frecuencia baja según realizaciones de la presente invención.

La envolvente de energía del dominio temporal 1101 de la voz de banda de frecuencia baja es una envolvente de energía suavizada con el tiempo e incluye una primera región de ruido de fondo 1102 y una segunda región de ruido de fondo 1105 separadas por regiones de voz no sonora 1103 y región de voz sonora 1104. La señal de voz sonora de baja frecuencia de la región de voz sonora 1104 tiene una energía más alta que la señal de voz no sonora de baja frecuencia en las regiones de voz no sonora 1103. Además, la señal de voz no sonora de baja frecuencia tiene una energía más alta o más cercana en comparación con la señal de ruido de fondo de baja frecuencia.

La Figura 2 ilustra una evaluación de energía del dominio temporal de la señal de voz de banda de frecuencia alta según realizaciones de la presente invención.

A diferencia de la Figura 1, la señal de voz de alta frecuencia tiene diferentes características. La envolvente de energía del dominio temporal de la señal de voz de banda alta 1201, que es la envolvente de energía suavizada con el tiempo, incluye una primera región de ruido de fondo 1202 y una segunda región de ruido de fondo 1205 separadas por regiones de voz no sonora 1203 y una región de voz sonora 1204. La señal de voz sonora de alta frecuencia tiene una energía más baja que la señal de voz no sonora de alta frecuencia. La señal de voz no sonora de alta frecuencia tiene una energía mucho más alta en comparación con la señal de ruido de fondo de alta frecuencia. Sin embargo, la señal de voz no sonora de alta frecuencia 1203 tiene una duración relativamente más corta que la voz sonora 1204.

Las realizaciones de la presente invención hacen uso de dicha diferencia en las características entre la voz sonora y no sonora en diferentes bandas de frecuencia en el dominio temporal. Por ejemplo, una señal en la trama presente puede identificarse como una señal sonora mediante la determinación de que la energía de la señal es más alta que la señal no sonora correspondiente en la banda baja pero no en la banda alta. De manera similar, una señal en la trama presente puede identificarse como una señal no sonora mediante la identificación de que la energía de la señal es más baja que la señal sonora correspondiente en la banda baja pero más alta que la señal sonora correspondiente en la banda alta.

Tradicionalmente, dos parámetros principales se usan para detectar la señal de voz No Sonora/Sonora. Un parámetro representa la periodicidad de la señal y otro parámetro indica la inclinación espectral, que es el grado en el cual la intensidad cae mientras la frecuencia aumenta.

Un parámetro de periodicidad de señal popular se provee más abajo en la Ecuación (1).

$$P^{1}_{sonoridad} = \frac{\sum_{n} s_w(n) \cdot s_w(n - Altura)}{\sqrt{\left(\sum_{n} |s_w(n)|^2\right) \left(\sum_{n} |s_w(n - Altura)|^2\right)}} = \frac{< s_w(n), s_w(n - Altura) >}{\sqrt{\|s_w(n)\|^2 \|s_w(n - Altura)\|^2}} \quad (1)$$

En la Ecuación (1), $s_w(n)$ es una señal de voz ponderada, el numerador es una correlación, y el denominador es un factor de normalización de energía El parámetro de periodicidad también se llama "correlación de altura" o "sonoridad". Otro parámetro de sonoridad a modo de ejemplo se provee más abajo en la Ecuación (2).

$$P^2_{sonoridad} = \frac{\sum_n \left|G_p \cdot e_p(n)\right|^2 - \sum_n \left|G_c \cdot e_c(n)\right|^2}{\sum_n \left|G_p \cdot e_p(n)\right|^2 + \sum_n \left|G_c \cdot e_c(n)\right|^2} \quad (2)$$

$$= \frac{\left\|G_p \cdot e_p(n)\right\|^2 - \left\|G_c \cdot e_c(n)\right\|^2}{\left\|G_p \cdot e_p(n)\right\|^2 + \left\|G_c \cdot e_c(n)\right\|^2}$$

En (2), $e_p(n)$ y $e_c(n)$ son señales de componentes de excitación y se describirán en mayor detalle más abajo. En varias aplicaciones, pueden usarse algunas variantes de las Ecuaciones (1) y (2) pero pueden aún representar la periodicidad de la señal.

El parámetro de inclinación espectral más popular se provee más abajo en la Ecuación (3).

$$P^1_{inclinación} = \frac{\sum_n s(n) \cdot s(n-1)}{\sqrt{\sum_n \left|s(n)\right|^2}} \quad (3)$$

$$= \frac{< s(n), s(n-1) >}{\sqrt{\left\|s_w(n)\right\|^2}}$$

En la Ecuación (3), $s(n)$ es una señal de voz. Si la energía del dominio de la frecuencia se encuentra disponible, el parámetro de inclinación espectral puede ser según se describe en la Ecuación (4).

$$P^2_{inclinación} = \frac{E_{LB} - E_{HB}}{E_{LB} + E_{HB}} \quad (4)$$

En la Ecuación (4), $E_{LB}$ es la energía de banda de frecuencia baja y $E_{HB}$ es la energía de banda de frecuencia alta.

Otro parámetro que puede reflejar la inclinación espectral se llama Tasa de Cruces por Cero (ZCR, por sus siglas en inglés). ZCR cuenta la tasa de cambio de señal positiva/negativa en una trama o subtrama. Normalmente, cuando la energía de banda de frecuencia alta es alta con respecto a la energía de banda de frecuencia baja, ZCR también es alta. De lo contrario, cuando la energía de banda de frecuencia alta es baja con respecto a la energía de banda de frecuencia baja, ZCR también es baja. En aplicaciones reales, pueden usarse algunas variantes de las Ecuaciones (3) y (4) pero pueden aún representar la inclinación espectral.

Según se ha mencionado previamente, la clasificación No Sonora/Sonora o Decisión No Sonora/Sonora se usa ampliamente en el campo de la codificación de señales de voz, extensión de ancho de banda (BWE) de señal de voz, mejora de señal de voz y reducción de ruido (NR) de fondo de señal de voz.

En la codificación de voz, la señal de voz no sonora puede codificarse mediante el uso de la excitación tipo ruido y la señal de voz sonora puede codificarse con excitación tipo pulso, según se ilustrará posteriormente. En la extensión de ancho de banda de señal de voz, la energía de señal de banda alta extendida de la señal de voz no sonora puede aumentarse mientras la energía de señal de banda alta extendida de la señal de voz sonora puede reducirse.

En la reducción de ruido (NR) de fondo de señal de voz, el algoritmo NR puede ser menos agresivo para la señal de voz no sonora y más agresivo para la señal de voz sonora. Entonces, una Decisión No Sonora o Sonora robusta es importante para los tipos de aplicaciones de más arriba. Según las características de la voz no sonora y voz sonora, tanto el parámetro de periodicidad $P_{sonoridad}$ como el parámetro de inclinación espectral $P_{inclinación}$ o sus parámetros variantes se usan, en mayor parte, para detectar clases No Sonora/Sonora. Sin embargo, los inventores de la presente solicitud han identificado que los valores "absolutos" del parámetro de periodicidad $P_{sonoridad}$ y el parámetro de inclinación espectral $P_{inclinación}$ o sus parámetros variantes se ven influenciados por el equipo de grabación de señales de voz, nivel de ruido de fondo y/o altavoces. Dichas influencias son difíciles de predeterminar y, posiblemente, resultan en una detección de voz No Sonora/Sonora no robusta.

Las realizaciones de la presente invención describen una detección de voz No Sonora/Sonora mejorada que usa los valores "relativos" del parámetro de periodicidad $P_{sonoridad}$ y el parámetro de inclinación espectral $P_{inclinación}$ o sus parámetros variantes en lugar de los valores "absolutos". Los valores "relativos" se ven mucho menos influenciados que los valores "absolutos" por el equipo de grabación de señales de voz, nivel de ruido de fondo y/o altavoces, lo cual resulta en una detección de voz No Sonora/Sonora más robusta.

Por ejemplo, un parámetro de no sonoridad combinado puede definirse como en la Ecuación (5) de más abajo.

$$P_{c\_\ no\ sonoridad} = (1 - P_{sonoridad}) \cdot (1 - P_{inclinación}) \cdots\cdots \quad (5)$$

Los puntos al final de la Ecuación (5) indican que pueden añadirse otros parámetros. Cuando el valor "absoluto" de $P_{c\_no\ sonoridad}$ se convierte en grande, es, probablemente, la señal de voz no sonora. Un parámetro de sonoridad combinado puede describirse como en la Ecuación (6) de más abajo.

$$P_{c\_\ sonoridad} = P_{sonoridad} \cdot P_{inclinación} \cdots\cdots \quad (6)$$

Los puntos al final de la Ecuación (6) indican, de manera similar, que pueden añadirse otros parámetros. Cuando el valor "absoluto" de $P_{c\_sonoridad}$ se convierte en grande, es, probablemente, la señal de voz sonora. Antes de que los valores "relativos" de $P_{c\_no\ sonoridad}$ o $P_{c\_sonoridad}$ se definan, un parámetro fuertemente suavizado de $P_{c\_no\ sonoridad}$ o $P_{c\_sonoridad}$ se define primero. Por ejemplo, el parámetro para la trama actual puede suavizarse a partir de una trama previa según se describe por desigualdad más abajo en la Ecuación (7).

$$\begin{aligned} &si \quad (P_{c\ \ no\ sonoridad\ \ sm} > P_{c\ \ no\ sonoridad}) \quad \{ \\ &\qquad P_{c\_\ no\ sonoridad\ \_sm} \Leftarrow 0.9\, P_{c\_\ no\ sonoridad\ \_sm} + 0.1\, P_{c\_\ no\ sonoridad} \\ &\} \\ &de\ otro\ modo \{ \\ &\qquad P_{c\ \ no\ sonoridad\ \ sm} \Leftarrow 0.99\, P_{c\ \ no\ sonoridad\ \ sm} + 0.01\, P_{c\ \ no\ sonoridad} \\ &\} \end{aligned} \quad (7)$$

En la Ecuación (7), $P_{c\_no\ sonoridad\_sm}$ es un valor fuertemente suavizado de $P_{c\_no\ sonoridad}$.

De manera similar, el parámetro de sonoridad combinado suavizado $P_{c\_sonoridad\_sm}$ puede determinarse mediante el uso de la desigualdad de más abajo mediante el uso de la Ecuación (8).

$$\begin{aligned} &si \quad (P_{c\_sonoridad\_sm} > P_{c\_sonoridad}) \quad \{ \\ &\qquad P_{c\_sonoridad\_sm} \Leftarrow (7/8)\, P_{c\_sonoridad\_sm} + (1/8)\, P_{c\_sonoridad} \\ &\} \\ &de\ otro\ modo \{ \\ &\qquad P_{c\_sonoridad\_sm} \Leftarrow (255/256)\, P_{c\_sonoridad\_sm} + (1/256)\, P_{c\_sonoridad} \\ &\} \end{aligned} \tag{8}$$

Aquí, en la Ecuación (8), $P_{c\_sonoridad\_sm}$ es un valor fuertemente suavizado de $P_{c\_sonoridad}$.

El comportamiento estadístico de la voz Sonora es diferente de aquel de la voz No Sonora y, por lo tanto, en varias realizaciones, los parámetros para decidir la desigualdad de más arriba (p.ej., 0,9, 0,99, 7/8, 255/256) pueden decidirse y además refinarse, si fuera necesario, según experimentos.

Los valores "relativos" de $P_{c\_no\ sonoridad}$ o $P_{c\_sonoridad}$ pueden definirse como en las Ecuaciones (9) y (10) descritas más abajo.

$$P_{c\_no\ sonoridad\_dif} = P_{c\_no\ sonoridad} - P_{c\_no\ sonoridad\_sm} \tag{9}$$

$P_{c\_no\ sonoridad\_dif}$ es el valor "relativo" de $P_{c\_no\ sonoridad}$; de manera similar,

$$P_{c\_sonoridad\_dif} = P_{c\_sonoridad} - P_{c\_sonoridad\_sm} \tag{10}$$

$P_{c\_sonoridad\_dif}$ es el valor "relativo" de $P_{c\_sonoridad}$.

La desigualdad de más abajo es una realización a modo de ejemplo de la aplicación de una detección No Sonora. En la presente realización a modo de ejemplo, establecer la bandera *No Sonora_bandera* para que sea *VERDADERO* indica que la señal de voz es una voz no sonora mientras que establecer la bandera *No Sonora_bandera* para que sea *FALSO* indica que la señal de voz no es una voz no sonora.

$$\begin{aligned} &si \quad (P_{c\_no\ sonoridad\_dif} > 0{,}1) \quad \{ \\ &\qquad No\ Sonora\_bandera = VERDADERO; \\ &\} \\ &de\ otro\ modo\ si \quad (P_{c\_no\ sonoridad\_dif} < 0{,}05) \quad \{ \\ &\qquad No\ Sonora\_bandera = FALSO; \\ &\} \\ &de\ otro\ modo \{ \\ &\qquad No\ Sonora\_bandera \quad no\ se\ cambia \quad (la\ anterior \quad No\ Sonora\_bandera \quad se\ mantiene). \\ &\} \end{aligned}$$

La desigualdad de más abajo es una realización alternativa a modo de ejemplo de la aplicación de una detección Sonora. En la presente realización a modo de ejemplo, establecer *Sonora_bandera* como *VERDADERO* indica que la señal de voz es una voz sonora mientras que establecer la *Sonora_bandera* para que sea *FALSO* indica que la señal de voz no es una voz sonora.

si $(P_{c\ sonoridad\ dif} > 0{,}1)$ {

Sonora _ bandera = VERDADERO;

}

de otro modo si $(P_{c\_sonoridad\_dif} < 0{,}05)$ {

Sonora _ bandera = FALSO ;

}

de otro modo{

Sonora _ bandera no se cambia ( la anterior Sonora _ bandera se mantiene).

}

Después de identificar la señal de voz como una que pertenece a una clase SONORA, la señal de voz puede entonces codificarse con el enfoque de codificación del dominio temporal como, por ejemplo, CELP. Las realizaciones de la presente invención también pueden aplicarse para reclasificar una señal NO SONORA en una señal SONORA antes de la codificación.

En varias realizaciones, el algoritmo de Detección No Sonora/Sonora mejorada puede usarse para mejorar AMR-WB-BWE y NR.

La Figura 3 ilustra funciones llevadas a cabo durante la codificación de una voz original mediante el uso de un codificador CELP convencional mediante la implementación de una realización de la presente invención.

La Figura 3 ilustra un codificador CELP inicial convencional donde un error ponderado 109 entre una voz sintetizada 102 y una voz original 101 se minimiza, con frecuencia, mediante el uso de un enfoque de análisis por síntesis, lo cual significa que la codificación (análisis) se lleva a cabo mediante la optimización perceptual de la señal decodificada (síntesis) en un bucle cerrado.

El principio básico que todos los codificadores de voz explotan es el hecho de que las señales de voz son formas de onda altamente correlacionadas. A modo de ilustración, la voz puede representarse mediante el uso de un modelo autorregresivo (AR) como en la Ecuación (11) de más abajo.

$$X_n = \sum_{i=1}^{L} a_i X_{n-1} + e_n \qquad (11)$$

En la Ecuación (11), cada muestra se representa como una combinación lineal de las $L$ muestras previas más un ruido blanco. Los coeficientes de ponderación $a_1, a_2, \ldots a_L$, se llaman Coeficientes de Predicción Lineal (LPC, por sus siglas en inglés). Para cada trama, los coeficientes de ponderación $a_1, a_2, \ldots a_L$, se eligen de modo que el espectro de $\{X_1, X_2, \ldots, X_N\}$, generado mediante el uso del modelo de más arriba, concuerda de manera cercana con el espectro de la trama de voz de entrada.

De manera alternativa, las señales de voz también pueden representarse por una combinación de un modelo armónico y modelo de ruido. La parte armónica del modelo es, de manera eficaz, una representación de serie de Fourier del componente periódico de la señal. En general, para las señales sonoras, el modelo de armónico más ruido de la voz está formado por una mezcla de armónicos y ruido. La proporción de armónico y ruido en una voz sonora depende de un número de factores que incluyen las características del hablante (p.ej., en qué medida la voz de un hablante es normal o entrecortada); el carácter de segmento de la voz (p. ej., en qué medida un segmento de voz es periódico) y de la frecuencia. Las frecuencias más altas de voz sonora tienen una proporción más alta de componentes tipo ruido.

El modelo de predicción lineal y el modelo de ruido armónico son los dos métodos principales para modelar y codificar señales de voz. El modelo de predicción lineal es particularmente bueno en el modelado de la envolvente espectral de la voz mientras que el modelo de ruido armónico es bueno en el modelado de la estructura fina de la voz. Los dos métodos pueden combinarse para beneficiarse de sus potencias relativas.

Según se ha indicado previamente, antes de la codificación CELP, la señal de entrada al micrófono del microteléfono se filtra y muestrea, por ejemplo, a una velocidad de 8000 muestras por segundo. Luego, cada muestra se

cuantifica, por ejemplo, con 13 bits por muestra. La velocidad de muestra se segmenta en segmentos o tramas de 20 ms (p.ej., en el presente caso, 160 muestras).

La señal de voz se analiza y su modelo PL, señales de excitación y altura se extraen. El modelo PL representa la envolvente espectral de la voz. Esta se convierte en un conjunto de coeficientes de frecuencias espectrales de línea (LSF, por sus siglas en inglés), que es una representación alternativa de parámetros de predicción lineal, dado que los coeficientes LSF tienen buenas propiedades de cuantificación. Los coeficientes LSF pueden cuantificarse por escalar o, de manera más eficaz, pueden cuantificarse por vector mediante el uso de libros de códigos de vector LSF previamente entrenados.

La excitación por código incluye un libro de códigos que comprende vectores de código, los cuales tienen componentes que se eligen, todos, de manera independiente, de modo que cada vector de código puede tener un espectro aproximadamente "blanco". Para cada subtrama de la voz de entrada, cada uno de los vectores de código se filtra a través del filtro de predicción lineal a corto plazo 103 y del filtro de predicción a largo plazo 105, y la salida se compara con las muestras de voz. En cada subtrama, el vector de código cuya salida concuerda mejor con la voz de entrada (error minimizado) se elige para representar dicha subtrama.

La excitación codificada 108 comprende, normalmente, una señal tipo pulso o señal tipo ruido, las cuales se construyen matemáticamente o se guardan en un libro de códigos. El libro de códigos se encuentra disponible tanto para el codificador como para el decodificador de recepción. La excitación codificada 108, que puede ser un libro de códigos estocástico o fijo, puede ser un diccionario de cuantificación de vector que se codifica (de forma implícita o explícita) de forma rígida en el códec. Dicho libro de códigos fijo puede ser una predicción lineal algebraica excitada por código o puede almacenarse de forma explícita.

Un vector de código del libro de códigos se escala por una ganancia apropiada para hacer que la energía sea igual a la energía de la voz de entrada. Por consiguiente, la salida de la excitación codificada 108 se escala por una ganancia $G_c$ 107 antes de atravesar los filtros lineales.

El filtro de predicción lineal a corto plazo 103 forma el espectro "blanco" del vector de código para parecerse al espectro de la voz de entrada. De manera equivalente, en el dominio temporal, el filtro de predicción lineal a corto plazo 103 incorpora correlaciones a corto plazo (correlación con muestras previas) en la secuencia blanca. El filtro que forma la excitación tiene un modelo de todos los polos de la forma 1/A(z) (filtro de predicción lineal a corto plazo 103), donde A(z) se llama el filtro de predicción y puede obtenerse mediante el uso de la predicción lineal (p.ej., algoritmo de Levinson-Durbin). En una o más realizaciones, un filtro de todos los polos puede usarse dado que es una buena representación del tracto vocal humano y dado que es fácil de computar.

El filtro de predicción lineal a corto plazo 103 se obtiene mediante el análisis de la señal original 101 y se representa por un conjunto de coeficientes:

$$A(z) = \sum_{i\ 1}^{P} 1 + a_i \cdot z^{-i} \quad , i = 1,2,....,P \qquad (12)$$

Según se ha descrito previamente, las regiones de voz sonora exhiben periodicidad a largo plazo. Dicho período, conocido como altura, se introduce en el espectro sintetizado por el filtro de altura 1/(B(z)). La salida del filtro de predicción a largo plazo 105 depende de la altura y ganancia de altura. En una o más realizaciones, la altura puede estimarse a partir de la señal original, señal residual o señal original ponderada. En una realización, la función de predicción a largo plazo ($B(z)$) puede expresarse mediante el uso de la Ecuación (13) de la siguiente manera.

$$B(z) = 1 - G_p \cdot z^{Altura} \qquad (13)$$

El filtro de ponderación 110 se relaciona con el filtro de predicción a corto plazo de más arriba. Uno de los filtros de ponderación típicos puede representarse según se describe en la Ecuación (14).

$$W(z) = \frac{A(z/\alpha)}{1-\beta \cdot z^{-1}} \qquad (14)$$

donde $\beta<\alpha$, $0<\beta<1$, $0<\alpha\leq 1$.

En otra realización, el filtro de ponderación $W(z)$ puede derivarse del filtro LPC por el uso de la expansión de ancho de banda según se ilustra en una realización en la Ecuación (15) de más abajo.

$$W(z) = \frac{A(z/\gamma 1)}{A(z/\gamma 2)} \qquad (15),$$

En la Ecuación (15), γ1>γ2, que son los factores con los cuales los polos se mueven hacia el origen.

Por consiguiente, para cada trama de voz, los LPC y la altura se computan y los filtros se actualizan. Para cada subtrama de voz, el vector de código que produce la "mejor" salida filtrada se elige para representar la subtrama. El valor cuantificado de ganancia correspondiente tiene que transmitirse al decodificador para la decodificación apropiada. Los LPC y los valores de altura también tienen que cuantificarse y enviarse en cada trama para la reconstrucción de los filtros en el decodificador. Por consiguiente, el índice de excitación codificada, índice de ganancia cuantificada, índice de parámetro de predicción a largo plazo cuantificado e índice de parámetro de predicción a corto plazo cuantificado se transmiten al decodificador.

La Figura 4 ilustra funciones llevadas a cabo durante la decodificación de una voz original mediante el uso de un decodificador CELP según una realización de la presente invención.

La señal de voz se reconstruye en el decodificador pasando los vectores de código recibidos a través de los filtros correspondientes. Como resultado, cada bloque, excepto el posprocesamiento, tiene la misma definición descrita en el codificador de la Figura 3.

El tren de bits CELP codificado se recibe y desempaqueta 80 en un dispositivo de recepción. Para cada subtrama recibida, el índice de excitación codificada recibido, índice de ganancia cuantificada, índice de parámetro de predicción a largo plazo cuantificado, e índice de parámetro de predicción a corto plazo cuantificado, se usan para encontrar los parámetros correspondientes mediante el uso de decodificadores correspondientes, por ejemplo, el decodificador de ganancia 81, decodificador de predicción a largo plazo 82 y decodificador de predicción a corto plazo 83. Por ejemplo, las posiciones y señas de amplitud de los pulsos de excitación y el vector de código algebraico de la excitación por código 402 pueden determinarse a partir del índice de excitación codificada recibido.

Con referencia a la Figura 4, el decodificador es una combinación de varios bloques que incluye excitación codificada 201, predicción a largo plazo 203 y predicción a corto plazo 205. El decodificador inicial además incluye un bloque de posprocesamiento 207 después de una voz sintetizada 206. El posprocesamiento puede además comprender posprocesamiento a corto plazo y posprocesamiento a largo plazo.

La Figura 5 ilustra un codificador CELP convencional usado en la implementación de las realizaciones de la presente invención.

La Figura 5 ilustra un codificador CELP básico mediante el uso de un libro de códigos adaptativo adicional para mejorar la predicción lineal a largo plazo. La excitación se produce mediante la suma de las contribuciones de un libro de códigos adaptativo 307 y una excitación por código 308, que puede ser un libro de códigos estocástico o fijo según se describe previamente. Las entradas en el libro de códigos adaptativo comprenden versiones retardadas de la excitación. Ello hace posible codificar, de manera eficaz, señales periódicas como, por ejemplo, sonidos sonoros.

Con referencia a la Figura 5, un libro de códigos adaptativo 307 comprende una excitación sintetizada pasada 304 o repetir el ciclo de altura de excitación pasado en el período de altura. El retardo de altura se puede codificar en un valor entero cuando es grande o largo. El retardo de altura se codifica, con frecuencia, en un valor fraccionario más preciso cuando es pequeño o corto. La información periódica de la altura se emplea para generar el componente adaptativo de la excitación. Dicho componente de excitación se escalona luego por una ganancia $G_p$ 305 (también llamada ganancia de altura).

La Predicción a Largo Plazo juega un papel muy importante para la codificación de voz sonora ya que la voz sonora tiene una fuerte periodicidad. Los ciclos de altura adyacentes de la voz sonora son similares entre sí, lo cual significa matemáticamente que la ganancia de altura $G_p$ en la siguiente excitación expresa es alta o cercana a 1. La excitación resultante puede expresarse como en la Ecuación (16) como una combinación de las excitaciones individuales.

$$e(n) = G_p \cdot e_p(n) + G_c \cdot e_c(n) \qquad (16)$$

donde $e_p(n)$ es una subtrama de una serie de muestras indexadas por $n$, que provienen del libro de códigos adaptativo 307 que comprende la excitación pasada 304 a través del bucle de realimentación (Figura 5). $e_p(n)$ puede filtrarse por paso bajo de manera adaptativa dado que el área de baja frecuencia es, con frecuencia, más periódica o más armónica que el área de alta frecuencia. $e_c(n)$ proviene del libro de códigos de excitación codificada 308 (también llamado libro de códigos fijo) que es una contribución de excitación actual. Además, $e_c(n)$ puede también mejorarse como, por ejemplo, mediante el uso de una mejora de filtrado de paso alto, mejora de altura, mejora de dispersión, mejora de formantes, y otros.

Para la voz sonora, la contribución de $e_p(n)$ del libro de códigos adaptativo 307 puede ser dominante y la ganancia de altura $G_p$ 305 puede ser un valor de alrededor de 1. La excitación se actualiza, normalmente, para cada subtrama. El tamaño de trama típico es de 20 milisegundos y el tamaño de subtrama típico es de 5 milisegundos.

Según se describe en la Figura 3, la excitación codificada fija 308 se escala por una ganancia $G_c$ 306 antes de atravesar los filtros lineales. Los dos componentes de excitación escalados de la excitación codificada fija 108 y libro de códigos adaptativo 307 se añaden juntos antes del filtrado a través del filtro de predicción lineal a corto plazo 303. Las dos ganancias ($G_p$ y $G_c$) se cuantifican y transmiten a un decodificador. Por consiguiente, el índice de excitación codificada, índice de libro de códigos adaptativo, índices de ganancia cuantificada, e índice de parámetro de predicción a corto plazo cuantificado se transmiten al dispositivo de audio de recepción.

El tren de bits CELP codificado mediante el uso de un dispositivo ilustrado en la Figura 5 se recibe en un dispositivo de recepción. La Figura 6 ilustra el decodificador correspondiente del dispositivo de recepción.

La Figura 6 ilustra un decodificador CELP básico correspondiente al codificador en la Figura 5 según una realización de la presente invención. La Figura 6 incluye un bloque de posprocesamiento 408 que recibe la voz sintetizada 407 del decodificador principal. Dicho decodificador es similar a la Figura 2 excepto por el libro de códigos adaptativo 307.

Para cada subtrama recibida, el índice de excitación codificada recibido, índice de ganancia de excitación codificada cuantificada, índice de altura cuantificada, índice de ganancia de libro de códigos adaptativo cuantificada, e índice de parámetro de predicción a corto plazo cuantificado, se usan para encontrar los parámetros correspondientes mediante el uso de decodificadores correspondientes, por ejemplo, el decodificador de ganancia 81, decodificador de altura 84, decodificador de ganancia de libro de códigos adaptativo 85, y decodificador de predicción a corto plazo 83.

En varias realizaciones, el decodificador CELP es una combinación de varios bloques y comprende excitación codificada 402, libro de códigos adaptativo 401, predicción a corto plazo 406 y posprocesamiento 408. Cada bloque, excepto el posprocesamiento, tiene la misma definición descrita en el codificador de la Figura 5. El posprocesamiento puede además incluir posprocesamiento a corto plazo y posprocesamiento a largo plazo.

Como ya se ha mencionado, CELP se usa, principalmente, para codificar una señal de voz beneficiándose de las características específicas de la voz humana o de un modelo humano de producción de voz vocal. Con el fin de codificar la señal de voz de manera más eficaz, la señal de voz se puede clasificar en diferentes clases y cada clase se codifica de manera diferente. La clasificación Sonora/No Sonora o Decisión No Sonora puede ser una clasificación importante y básica entre todas las clasificaciones de diferentes clases. Para cada clase, el filtro LPC o STP se usa siempre para representar la envolvente espectral. Pero la excitación para el filtro LPC puede ser diferente. Las señales no sonoras pueden codificarse con una excitación tipo ruido. Por otro lado, las señales sonoras pueden codificarse con una excitación tipo pulso.

El bloque de excitación por código (al que se hace referencia con la etiqueta 308 en la Figura 5 y 402 en la Figura 6) ilustra la ubicación del Libro de Códigos Fijo (FCB) para una codificación CELP general. Un vector de código seleccionado de FCB se escalona por una ganancia que con frecuencia se nota como $G_c$ 306.

La Figura 7 ilustra vectores candidatos tipo ruido para construir el libro de códigos de excitación codificada o libro de códigos fijo de la codificación de voz CELP.

Un FCB que contiene vectores tipo ruido puede ser la mejor estructura para señales no sonoras desde el punto de vista de la calidad perceptual. Ello se debe a que la contribución del libro de códigos adaptativo o contribución LTP será pequeña o no existente, y la principal contribución de excitación depende del componente FCB para la señal de clase no sonora. En el presente caso, si se usa un FCB tipo pulso, la señal de voz sintetizada de salida puede sonar filosa dado que existen muchos ceros en el vector de código seleccionado del FCB tipo pulso diseñado para la codificación de bajas velocidades binarias.

Con referencia a la Figura 7, se ilustra una estructura FCB que incluye vectores candidatos tipo ruido para construir una excitación codificada. El FCB tipo ruido 501 selecciona un vector de código tipo ruido 502 particular, el cual se escala por la ganancia 503.

La Figura 8 ilustra vectores candidatos tipo pulso para construir el libro de códigos de excitación codificada o libro de códigos fijo de la codificación de voz CELP.

Un FCB tipo pulso provee una mejor calidad que un FCB tipo ruido para la señal de clase sonora desde el punto de vista perceptual. Ello se debe a que la contribución del libro de códigos adaptativo o contribución LTP será dominante para la señal de clase sonora altamente periódica y la principal contribución de excitación no depende del componente FCB para la señal de clase sonora. Si se usa un FCB tipo ruido, la señal de voz sintetizada de salida puede sonar ruidosa o menos periódica dado que es más difícil tener una buena concordancia de forma de onda

mediante el uso del vector de código seleccionado del FCB tipo ruido diseñado para la codificación de bajas velocidades binarias.

Con referencia a la Figura 8, una estructura FCB puede incluir múltiples vectores candidatos tipo pulso para construir una excitación codificada. Un vector de código tipo pulso 602 se selecciona del FCB tipo pulso 601 y se escala por la ganancia 603.

La Figura 9 ilustra un ejemplo de espectro de excitación para la voz sonora. Después de eliminar la envolvente espectral LPC 704, el espectro de excitación 702 es casi plano. El espectro de excitación de banda baja 701 es, normalmente, más armónico que el espectro de banda alta 703. En teoría, el espectro de excitación de banda alta ideal o no cuantificado puede tener casi el mismo nivel de energía que el espectro de excitación de banda baja. En la práctica, si tanto la banda baja como la banda alta se codifican con tecnología CELP, el espectro de banda alta sintetizado o cuantificado puede tener un nivel de energía más bajo que el espectro de banda baja sintetizado o cuantificado por al menos dos motivos. Primero, la codificación CELP en bucle cerrado enfatiza más la banda baja que la banda alta. Segundo, la concordancia de forma de onda para la señal de banda baja es más fácil que la señal de banda alta, no solo debido al cambio más rápido de la señal de banda alta sino también debido a la característica más tipo ruido de la señal de banda alta.

En la codificación CELP de velocidad binaria baja como, por ejemplo, AMR-WB, la banda alta no se codifica, normalmente, sino que se genera en el decodificador con una tecnología de extensión de ancho de banda (BWE). En el presente caso, el espectro de excitación de banda alta puede simplemente copiarse del espectro de excitación de banda baja mientras se añade cierto ruido aleatorio. La envolvente de energía espectral de banda alta puede predecirse o estimarse a partir de la envolvente de energía espectral de banda baja. El control apropiado de la energía de señal de banda alta se convierte en importante cuando se usa BWE. A diferencia de la señal de voz no sonora, la energía de la señal de voz sonora de banda alta generada tiene que reducirse de manera apropiada para lograr la mejor calidad perceptual.

La Figura 10 ilustra un ejemplo de un espectro de excitación para la voz no sonora.

En el caso de la voz no sonora, el espectro de excitación 802 es casi plano después de eliminar la envolvente espectral LPC 804. Tanto el espectro de excitación de banda baja 801 como el espectro de banda alta 803 son tipo ruido. En teoría, el espectro de excitación de banda alta ideal o no cuantificado puede tener casi el mismo nivel de energía que el espectro de excitación de banda baja. En la práctica, si tanto la banda baja como la banda alta se codifican con tecnología CELP, el espectro de banda alta sintetizado o cuantificado puede tener un nivel de energía igual o ligeramente más alto que el espectro de banda baja sintetizado o cuantificado por dos motivos. Primero, la codificación CELP en bucle cerrado enfatiza más el área de energía más alta. Segundo, aunque la concordancia de forma de onda para la señal de banda baja es más fácil que la señal de banda alta, siempre es difícil tener una buena concordancia de forma de onda para señales tipo ruido.

De manera similar a la codificación de voz sonora, para la codificación CELP de velocidad binaria baja no sonora como, por ejemplo, AMR-WB, la banda alta no se codifica, normalmente, sino que se genera en el decodificador con una tecnología BWE. En el presente caso, el espectro de excitación de banda alta no sonora puede simplemente copiarse del espectro de excitación de banda baja no sonora mientras se añade cierto ruido aleatorio. La envolvente de energía espectral de banda alta de señal de voz no sonora puede predecirse o estimarse a partir de la envolvente de energía espectral de banda baja. El control de la energía de la señal de banda alta no sonora de manera apropiada es especialmente importante cuando se usa la BWE. A diferencia de la señal de voz sonora, es mejor que la energía de la señal de voz no sonora de banda alta generada aumente de manera apropiada para lograr una mejor calidad perceptual.

La Figura 11 ilustra un ejemplo de espectro de excitación para la señal de ruido de fondo.

El espectro de excitación 902 es casi plano después de eliminar la envolvente espectral LPC 904. El espectro de excitación de banda baja 901 es, normalmente, tipo ruido como el espectro de banda alta 903. En teoría, el espectro de excitación de banda alta ideal o no cuantificado de la señal de ruido de fondo puede tener casi el mismo nivel de energía que el espectro de excitación de banda baja. En la práctica, si tanto la banda baja como la banda alta se codifican con tecnología CELP, el espectro de banda alta sintetizado o cuantificado de la señal de ruido de fondo puede tener un nivel de energía más bajo que el espectro de banda baja sintetizado o cuantificado por dos motivos. Primero, la codificación CELP en bucle cerrado enfatiza más la banda baja que tiene energía más alta que la banda alta. Segundo, la concordancia de forma de onda para la señal de banda baja es más fácil que la señal de banda alta. De manera similar a la codificación de voz, para la codificación CELP de velocidad binaria baja de la señal de ruido de fondo, la banda alta no se codifica, normalmente, sino que se genera en el decodificador con una tecnología BWE. En el presente caso, el espectro de excitación de banda alta de la señal de ruido de fondo puede simplemente copiarse del espectro de excitación de banda baja mientras se añade cierto ruido aleatorio; la envolvente de energía espectral de banda alta de la señal de ruido de fondo puede predecirse o estimarse a partir de la envolvente de energía espectral de banda baja. El control de la señal de ruido de fondo de banda alta puede ser diferente de la

señal de voz cuando se usa la BWE. A diferencia de la señal de voz, es mejor que la energía de la señal de voz de ruido de fondo de banda alta generada sea estable con el tiempo para lograr una mejor calidad perceptual.

Las Figuras 12A y 12B ilustran ejemplos de codificación/decodificación de dominio de la frecuencia con extensión de ancho de banda. La Figura 12A ilustra el codificador con información conexa BWE, mientras la Figura 12B ilustra el decodificador con BWE.

Con referencia, primero, a la Figura 12A, la señal de banda baja 1001 se codifica en el dominio de la frecuencia mediante el uso de parámetros de banda baja 1002. Los parámetros de banda baja 1002 se cuantifican y el índice de cuantificación se transmite a un dispositivo de acceso a audio de recepción a través del canal de tren de bits 1003. La señal de banda alta extraída de la señal de audio 1004 se codifica con una pequeña cantidad de bits mediante el uso de los parámetros de lado de banda alta 1005. Los parámetros de lado de banda alta cuantificados (índice de información conexa HB) se transmiten al dispositivo de acceso a audio de recepción a través del canal de tren de bits 1006.

Con referencia a la Figura 12B, en el decodificador, el tren de bits de banda baja 1007 se usa para producir una señal de banda baja decodificada 1008. El tren de bits de lado de banda alta 1010 se usa para decodificar y generar los parámetros de lado de banda alta 1011. La señal de banda alta 1012 se genera a partir de la señal de banda baja 1008 con ayuda de los parámetros de lado de banda alta 1011. La señal de audio final 1009 se produce mediante la combinación de la señal de banda baja y la señal de banda alta. La BWE de dominio de la frecuencia también necesita un control de energía apropiado de la señal de banda alta generada. Los niveles de energía pueden establecerse de manera diferente para señales No Sonoras, Sonoras y de Ruido. Entonces, la clasificación de alta calidad de la señal de voz también se necesita para la BWE del dominio de la frecuencia.

Detalles relevantes del algoritmo de reducción de ruido de fondo se describen más abajo. En general, dado que la señal de voz no sonora es tipo ruido, la reducción de ruido de fondo (NR) en una área no sonora debe ser menos agresiva que en el área sonora, beneficiándose del efecto de enmascaramiento por ruido. En otras palabras, un ruido de fondo de mismo nivel es más audible en el área sonora que en el área no sonora de modo que NR debe ser más agresiva en el área sonora que en el área no sonora. En dicho caso, se necesita una decisión No Sonora/Sonora de alta calidad.

En general, la señal de voz no sonora es una señal tipo ruido que no tiene periodicidad. Además, la señal de voz no sonora tiene más energía en el área de frecuencia alta que en el área de frecuencia baja. Por el contrario, la señal de voz sonora tiene características opuestas. Por ejemplo, la señal de voz sonora es un tipo de señal cuasiperiódica que, normalmente, tiene más energía en el área de frecuencia baja que en el área de frecuencia alta (es preciso ver también las Figuras 9 y 10).

Las Figuras 13A-13C son ilustraciones esquemáticas de procesamiento de voz mediante el uso de varias realizaciones de procesamiento de voz descritas más arriba.

Con referencia a la Figura 13A, un método para el procesamiento de voz incluye recibir múltiples tramas de una señal de voz que se procesarán (casilla 1310). En varias realizaciones, las múltiples tramas de una señal de voz pueden generarse dentro del mismo dispositivo de audio, p.ej., que comprende un micrófono. En una realización alternativa, la señal de voz puede recibirse en un dispositivo de audio como un ejemplo. Por ejemplo, la señal de voz puede codificarse o decodificarse posteriormente. Para cada trama, se determina un parámetro de no sonoridad/sonoridad que refleja una característica de voz no sonora/sonora en la trama actual (casilla 1312). En varias realizaciones, el parámetro de no sonoridad/sonoridad puede incluir un parámetro de periodicidad, un parámetro de inclinación espectral, u otras variantes. El método además incluye determinar un parámetro de no sonoridad suavizado para incluir información del parámetro de no sonoridad/sonoridad en tramas previas de la señal de voz (casilla 1314). Se obtiene una diferencia entre el parámetro de no sonoridad/sonoridad y el parámetro de no sonoridad/sonoridad suavizado (casilla 1316). De manera alternativa, un valor relativo (p.ej., relación) entre el parámetro de no sonoridad/sonoridad y el parámetro de no sonoridad/sonoridad suavizado puede obtenerse. Cuando se decide si una trama actual es más apropiada para que se maneje como una voz no sonora/sonora, la decisión no sonora/sonora se lleva a cabo mediante el uso de la diferencia determinada como un parámetro de decisión (casilla 1318).

Con referencia a la Figura 13B, un método para el procesamiento de voz incluye recibir múltiples tramas de una señal de voz (casilla 1320). La realización se describe mediante el uso de un parámetro de sonoridad pero se aplica igualmente al uso de un parámetro de no sonoridad. Un parámetro de sonoridad combinado se determina para cada trama (casilla 1322). En una o más realizaciones, el parámetro de sonoridad combinado puede ser un parámetro de periodicidad y un parámetro de inclinación y un parámetro de sonoridad combinado suavizado. El parámetro de sonoridad combinado suavizado puede obtenerse mediante el suavizado del parámetro de sonoridad combinado en una o más tramas previas de la señal de voz. El parámetro de sonoridad combinado se compara con el parámetro de sonoridad combinado suavizado (casilla 1324). La trama actual se clasifica como una señal de voz SONORA o una señal de voz NO SONORA mediante el uso de la comparación en la toma de decisiones (casilla 1326). La señal

de voz puede procesarse, por ejemplo, codificarse o decodificarse, según la clasificación determinada de la señal de voz (casilla 1328).

Con referencia, a continuación, a la Figura 13C, en otra realización a modo de ejemplo, un método para el procesamiento de voz comprende recibir múltiples tramas de una señal de voz (casilla 1330). Se determina una primera envolvente de energía de la señal de voz en el dominio temporal (casilla 1332). La primera envolvente de energía puede determinarse dentro de una primera banda de frecuencia, por ejemplo, una banda de frecuencia baja como, por ejemplo, hasta 4000 Hz. Una energía de banda de frecuencia baja suavizada puede determinarse a partir de la primera envolvente de energía mediante el uso de las tramas previas. Una diferencia o una primera relación de la energía de banda de frecuencia baja de la señal de voz con respecto a la energía de banda de frecuencia baja suavizada se computa (casilla 1334). Una segunda envolvente de energía de la señal de voz se determina en el dominio temporal (casilla 1336). La segunda envolvente de energía se determina dentro de una segunda banda de frecuencia. La segunda banda de frecuencia es una banda de frecuencia diferente de la primera banda de frecuencia. Por ejemplo, la segunda frecuencia puede ser una banda de frecuencia alta. En un ejemplo, la segunda banda de frecuencia puede ser de entre 4000 Hz y 8000 Hz. Una energía de banda de frecuencia alta suavizada en una o más de las tramas previas de la señal de voz se computa. Una diferencia o una segunda relación se determina mediante el uso de la segunda envolvente de energía para cada trama (casilla 1338). La segunda relación puede computarse como la relación entre la energía de banda de frecuencia alta de la señal de voz en la trama actual con respecto a la energía de banda de frecuencia alta suavizada. La trama actual se clasifica como una señal de voz SONORA o una señal de voz NO SONORA mediante el uso de la primera relación y la segunda relación en la toma de decisiones (casilla 1340). La señal de voz clasificada se procesa, p.ej., se codifica, decodifica, y otras, según la clasificación determinada de la señal de voz (casilla 1342).

En una o más realizaciones, la señal de voz puede codificarse/decodificarse mediante el uso de la excitación tipo ruido cuando se determina que la señal de voz es una señal de voz NO SONORA, y en donde la señal de voz se codifica/decodifica con excitación tipo pulso cuando se determina que la señal de voz es una señal SONORA.

En realizaciones adicionales, la señal de voz puede codificarse/decodificarse en el dominio de la frecuencia cuando se determina que la señal de voz es una señal NO SONORA, y en donde la señal de voz se codifica/decodifica en el dominio temporal cuando se determina que la señal de voz es una señal SONORA.

Por consiguiente, las realizaciones de la presente invención pueden usarse para mejorar la decisión No Sonora/Sonora para la codificación de voz, extensión de ancho de banda y/o mejora de voz.

La Figura 14 ilustra un sistema de comunicación 10 según una realización de la presente invención.

El sistema de comunicación 10 tiene dispositivos de acceso a audio 7 y 8 acoplados a una red 36 mediante enlaces de comunicación 38 y 40. En una realización, los dispositivos de acceso a audio 7 y 8 son dispositivos de protocolo de transmisión de la voz por internet (VOIP, por sus siglas en inglés) y la red 36 es una red de área amplia (WAN, por sus siglas en inglés), red telefónica pública conmutada (PTSN, por sus siglas en inglés) y/o Internet. En otra realización, los enlaces de comunicación 38 y 40 son conexiones de banda ancha alámbrica y/o inalámbrica. En una realización alternativa, los dispositivos de acceso de audio 7 y 8 son teléfonos celulares o móviles, los enlaces 38 y 40 son canales telefónicos móviles inalámbricos y la red 36 representa una red telefónica móvil.

El dispositivo de acceso a audio 7 usa un micrófono 12 para convertir sonido, como, por ejemplo, música o la voz de una persona, en una señal de entrada de audio analógico 28. Una interfaz de micrófono 16 convierte la señal de entrada de audio analógico 28 en una señal de audio digital 33 para la entrada en un codificador 22 de CÓDEC 20. El codificador 22 produce la señal de audio codificada TX para la transmisión a una red 26 mediante una interfaz de red 26 según las realizaciones de la presente invención. Un decodificador 24 dentro del CÓDEC 20 recibe la señal de audio codificada RX de la red 36 mediante la interfaz de red 26 y convierte la señal de audio codificada RX en una señal de audio digital 34. La interfaz de altavoz 18 convierte la señal de audio digital 34 en la señal de audio 30 apropiada para dirigir los altavoces 14.

En las realizaciones de la presente invención, donde el dispositivo de acceso a audio 7 es un dispositivo VOIP, algunos o todos los componentes dentro del dispositivo de acceso a audio 7 se implementan dentro de un microteléfono. En algunas realizaciones, sin embargo, el micrófono 12 y el altavoz 14 son unidades separadas y la interfaz de micrófono 16, interfaz de altavoz 18, CÓDEC 20 e interfaz de red 26 se implementan dentro de un ordenador personal. El CÓDEC 20 se puede implementar en software que se ejecuta en un ordenador o un procesador dedicado o mediante hardware dedicado, por ejemplo, en un circuito integrado para aplicaciones específicas (ASIC, por sus siglas en inglés). La interfaz de micrófono 16 se implementa por un convertidor analógico digital (A/D), así como otros circuitos de interfaz ubicados dentro del microteléfono y/o dentro del ordenador. Asimismo, la interfaz de altavoz 18 se implementa por un convertidor digital analógico y otros circuitos de interfaz ubicados dentro del microteléfono y/o dentro del ordenador. En realizaciones adicionales, el dispositivo de acceso a audio 7 se puede implementar y particionar de otras maneras conocidas en la técnica.

En las realizaciones de la presente invención donde el dispositivo de acceso a audio 7 es un teléfono celular o móvil, los elementos dentro del dispositivo de acceso a audio 7 se implementan dentro de un microteléfono celular. El CÓDEC 20 se implementa por software que se ejecuta en un procesador dentro del microteléfono o por hardware dedicado. En realizaciones adicionales de la presente invención, el dispositivo de acceso a audio se puede implementar en otros dispositivos como, por ejemplo, sistemas de comunicaciones digitales alámbricos e inalámbricos entre pares como, por ejemplo, intercomunicaciones y aparatos de radio. En aplicaciones como, por ejemplo, dispositivos de audio para el consumidor, el dispositivo de acceso a audio puede contener un CÓDEC con un codificador 22 o decodificador 24 solamente, por ejemplo, en un sistema de micrófono digital o dispositivo de reproducción musical. En otras realizaciones de la presente invención, el CÓDEC 20 se puede usar sin micrófono 12 y altavoz 14, por ejemplo, en estaciones base celulares que acceden a la PTSN.

El procesamiento de voz para mejorar la clasificación no sonora/sonora descrita en varias realizaciones de la presente invención puede implementarse en el codificador 22 o decodificador 24, por ejemplo. El procesamiento de voz para mejorar la clasificación no sonora/sonora puede implementarse en hardware o software en varias realizaciones. Por ejemplo, el codificador 22 o decodificador 24 pueden ser parte de un chip de procesamiento de señales digitales (DSP, por sus siglas en inglés).

La Figura 15 ilustra un diagrama de bloques de un sistema de procesamiento que puede usarse para implementar los dispositivos y métodos descritos en la presente memoria. Dispositivos específicos pueden utilizar todos los componentes que se muestran, o solamente un subconjunto de los componentes, y los niveles de integración pueden variar de dispositivo a dispositivo. Además, un dispositivo puede contener múltiples instancias de un componente como, por ejemplo, múltiples unidades de procesamiento, procesadores, memorias, transmisores, receptores, etc. El sistema de procesamiento puede comprender una unidad de procesamiento equipada con uno o más dispositivos de entrada/salida como, por ejemplo, un altavoz, micrófono, ratón, pantalla táctil, teclado, impresora, visualización, y similares. La unidad de procesamiento puede incluir una unidad de procesamiento central (CPU, por sus siglas en inglés), memoria, un dispositivo de almacenamiento masivo, un adaptador de vídeo, y una interfaz E/S conectada a un bus.

El bus puede ser uno o más de cualquier tipo de varias arquitecturas de bus que incluyen un bus de memoria o controlador de memoria, un bus periférico, bus de vídeo, o similares. La CPU puede comprender cualquier tipo de procesador electrónico de datos. La memoria puede comprender cualquier tipo de memoria de sistema como, por ejemplo, memoria estática de acceso aleatorio (SRAM, por sus siglas en inglés), memoria dinámica de acceso aleatorio (DRAM, por sus siglas en inglés), DRAM síncrona (SDRAM, por sus siglas en inglés), memoria de solo lectura (ROM, por sus siglas en inglés), una combinación de ellas, o similares. En una realización, la memoria puede incluir ROM para su uso en el arranque, y DRAM para el almacenamiento de programas y datos para su uso mientras se ejecutan programas.

El dispositivo de almacenamiento masivo puede comprender cualquier tipo de dispositivo de almacenamiento configurado para almacenar datos, programas y otra información y para hacer que los datos, programas y otra información sean accesibles mediante el bus. El dispositivo de almacenamiento masivo puede comprender, por ejemplo, una o más de una unidad en estado sólido, unidad de disco duro, una unidad de disco magnético, una unidad de disco óptico, o similares.

El adaptador de vídeo y la interfaz E/S proveen interfaces para acoplar dispositivos de entrada y salida externos a la unidad de procesamiento. Según se ilustra, ejemplos de dispositivos de entrada y salida incluyen la visualización acoplada al adaptador de vídeo y el ratón/teclado/impresora acoplados a la interfaz E/S. Otros dispositivos pueden acoplarse a la unidad de procesamiento, y pueden utilizarse menos tarjetas de interfaz o tarjetas de interfaz adicionales. Por ejemplo, una interfaz serial como, por ejemplo, un Bus Serial Universal (USB, por sus siglas en inglés) (no se muestra) puede usarse para proveer una interfaz para una impresora.

La unidad de procesamiento también incluye una o más interfaces de red, que pueden comprender enlaces cableados como, por ejemplo, un cable Ethernet o similares, y/o enlaces inalámbricos para acceder a nodos o diferentes redes. La interfaz de red permite a la unidad de procesamiento comunicarse con unidades remotas mediante las redes. Por ejemplo, la interfaz de red puede proveer una comunicación inalámbrica mediante uno o más transmisores/antenas de transmisión y uno o más receptores/antenas de recepción. En una realización, la unidad de procesamiento se acopla a una red de área local o red de área amplia para el procesamiento de datos y comunicaciones con dispositivos remotos como, por ejemplo, otras unidades de procesamiento, Internet, instalaciones de almacenamiento remoto, o similares.

Mientras la presente invención se ha descrito con referencia a realizaciones ilustrativas, la presente descripción no pretende interpretarse en un sentido restrictivo. Varias modificaciones y combinaciones de las realizaciones ilustrativas, así como otras realizaciones de la invención, serán aparentes para las personas con experiencia en la técnica con referencia a la descripción. Por ejemplo, varias realizaciones descritas más arriba pueden combinarse entre sí.

Aunque la presente invención y sus ventajas se han descrito en detalle, debe comprenderse que varios cambios, reemplazos y alteraciones pueden llevarse a cabo en la presente memoria sin apartarse del alcance de la invención según se define por las reivindicaciones anexas. Por ejemplo, muchas de las características y funciones descritas más arriba pueden implementarse en software, hardware, o firmware, o una combinación de ellos. Además, el alcance de la presente solicitud no pretende limitarse a las realizaciones particulares del proceso, máquina, fabricación, composición química, medios, métodos y etapas descritas en la memoria descriptiva. Como una persona con experiencia ordinaria en la técnica apreciará inmediatamente a partir de la descripción de la presente invención, los procesos, máquinas, fabricación, composiciones químicas, medios, métodos, o etapas, actualmente existentes o que se desarrollarán más tarde, que llevan a cabo sustancialmente la misma función o logran sustancialmente el mismo resultado que las realizaciones correspondientes descritas en la presente memoria pueden utilizarse según la presente invención. Por consiguiente, las reivindicaciones anexas pretenden incluir dentro de su alcance dichos procesos, máquinas, fabricación, composiciones químicas, medios, métodos o etapas.

**REIVINDICACIONES**

1. Un aparato de procesamiento de voz que comprende:

un procesador; y

un medio de almacenamiento legible por ordenador que almacena la programación para la ejecución por el procesador, la programación incluyendo instrucciones adaptadas para:

determinar un parámetro de no sonoridad que refleja una característica de voz no sonora en una trama actual de una señal de voz que comprende múltiples tramas,

determinar un parámetro de no sonoridad suavizado para incluir información del parámetro de no sonoridad en una trama anterior a la trama actual de la señal de voz,

computar una diferencia entre el parámetro de no sonoridad y el parámetro de no sonoridad suavizado, y

determinar si la trama actual comprende voz no sonora o voz sonora mediante el uso de la diferencia computada como un parámetro de decisión;

en donde el parámetro de no sonoridad es un parámetro combinado que refleja un producto de un parámetro de periodicidad y un parámetro de inclinación espectral.

2. El aparato de la reivindicación 1, en donde cuando la diferencia entre el parámetro de no sonoridad y el parámetro de no sonoridad suavizado es mayor que 0,1, determinar la trama actual de la señal de voz que será una voz no sonora, en donde cuando la diferencia entre el parámetro de no sonoridad y el parámetro de no sonoridad suavizado es menor que 0,05, determinar la trama actual de la señal de voz que no será una voz no sonora.

3. El aparato de las reivindicaciones 1 o 2, en donde la trama comprende una subtrama.

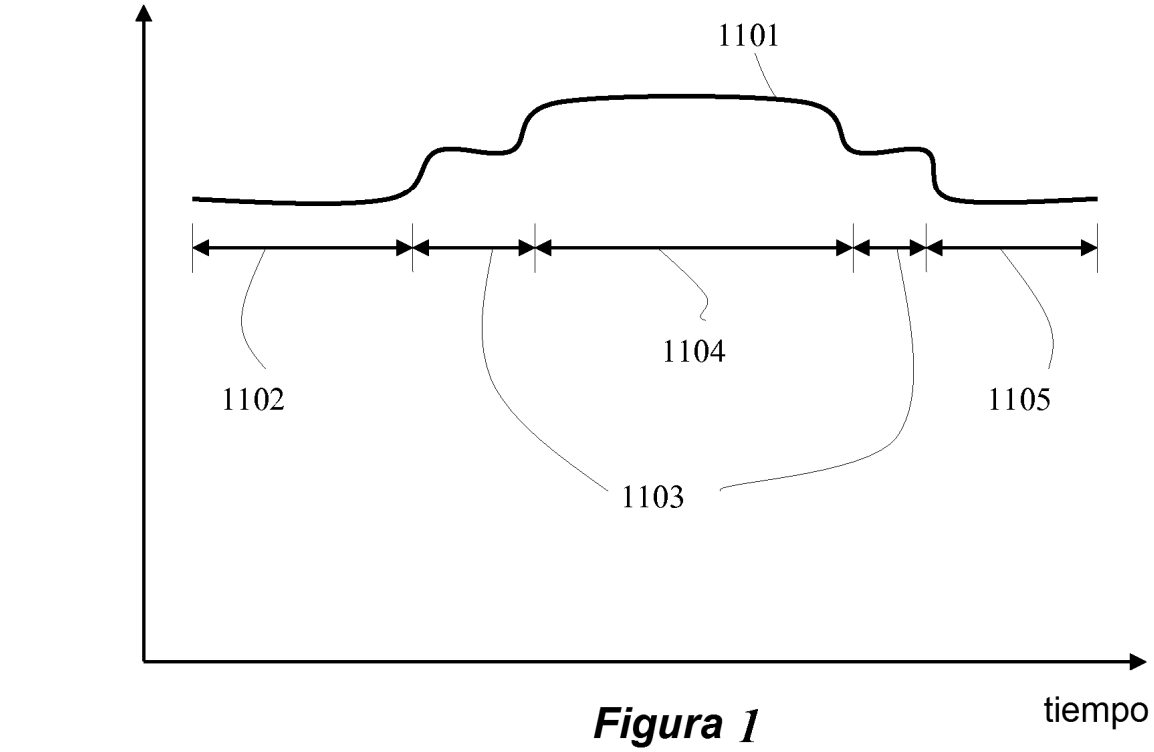
Envolvente de Energía de Dominio Temporal
1101
1102
1103
1104
1105
tiempo

***Figura 1***

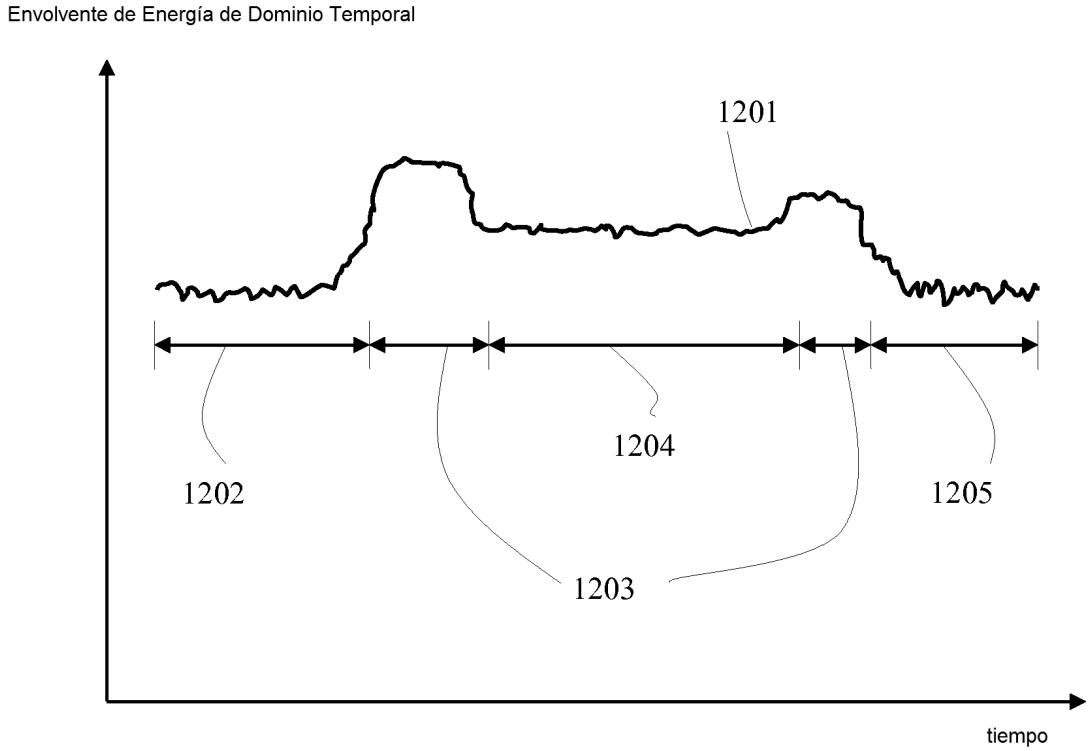
Envolvente de Energía de Dominio Temporal
1201
1202
1203
1204
1205
tiempo

***Figura 2***

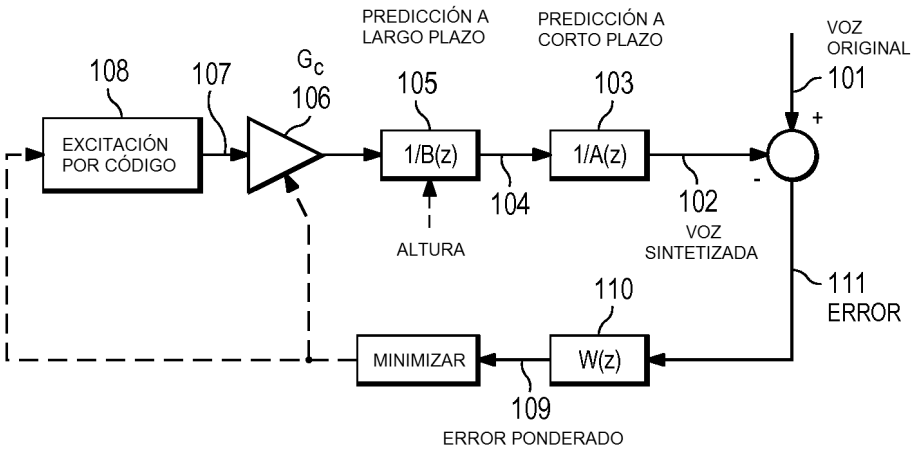


***Figura 3***

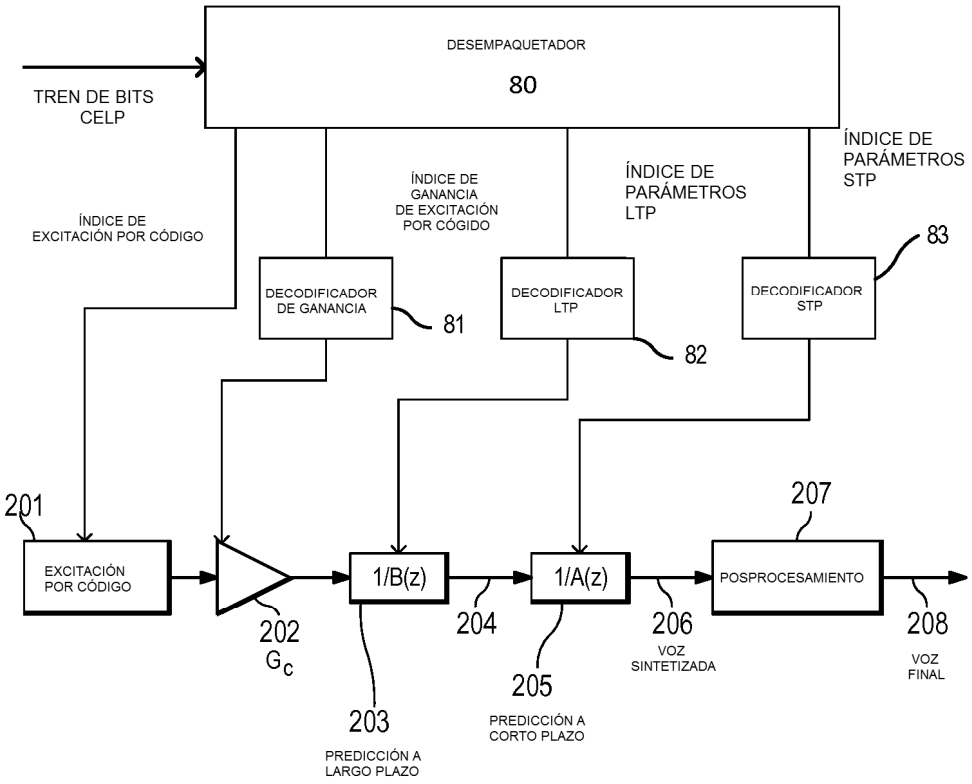


***Figura 4***

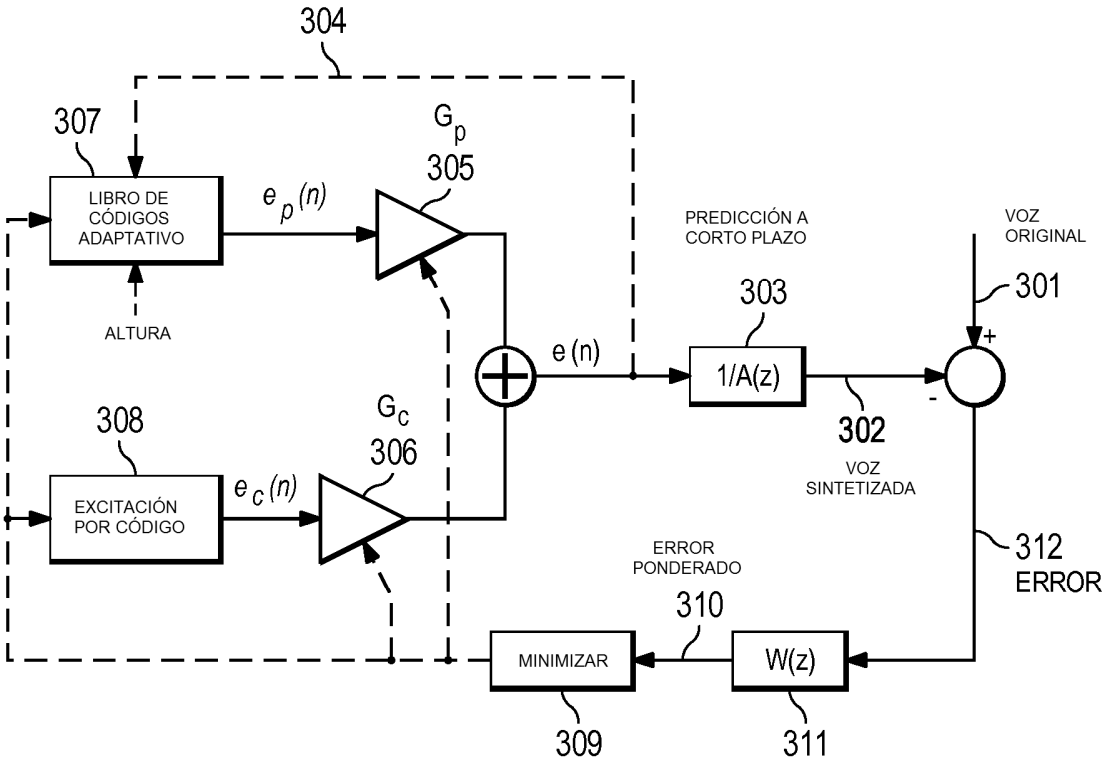
304
307
LIBRO DE CÓDIGOS ADAPTATIVO
ALTURA
$e_p(n)$
$G_p$
305
308
EXCITACIÓN POR CÓDIGO
$e_c(n)$
$G_c$
306
e (n)
PREDICCIÓN A CORTO PLAZO
303
1/A(z)
302
VOZ SINTETIZADA
VOZ ORIGINAL
301
+
-
312
ERROR
ERROR PONDERADO
310
MINIMIZAR
W(z)
309
311

***Figura 5***

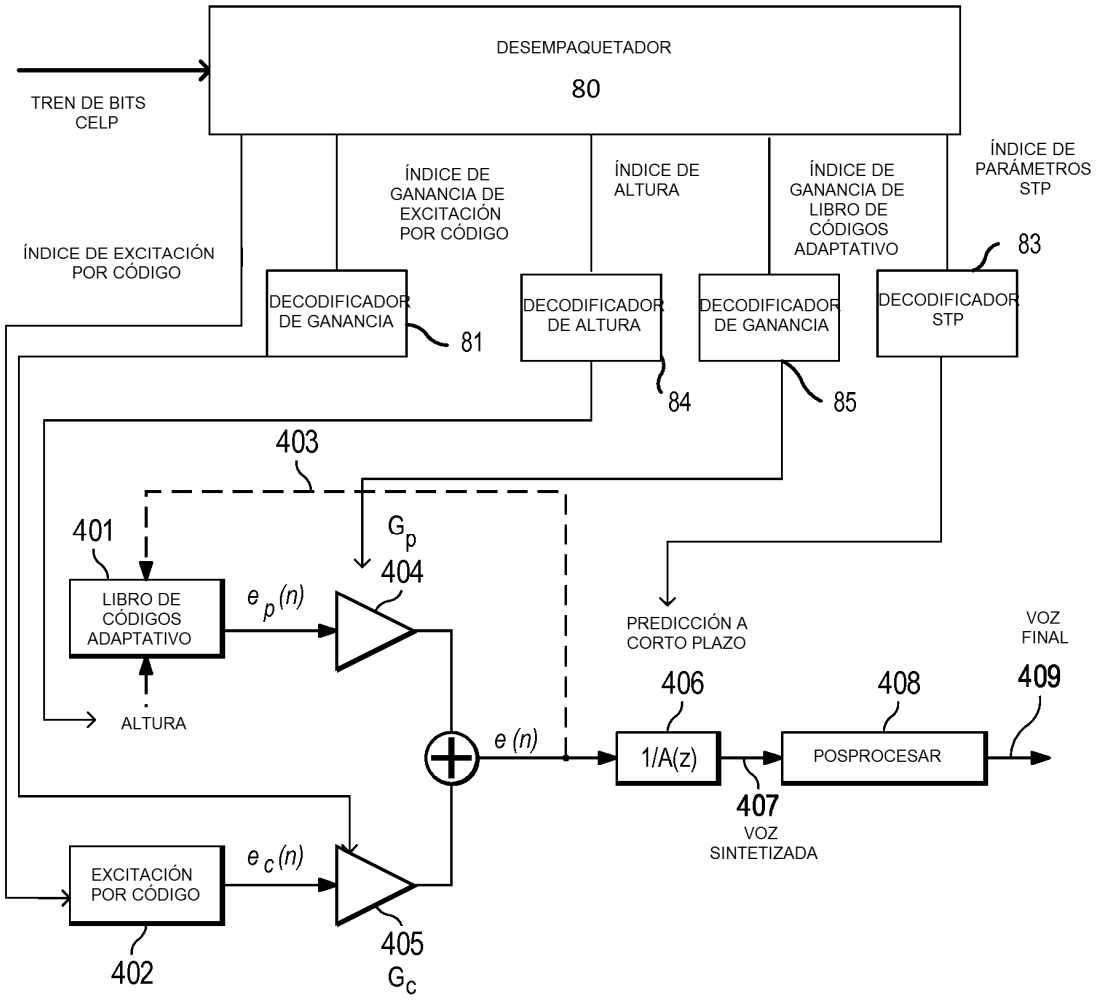
DESEMPAQUETADOR
80
TREN DE BITS CELP
ÍNDICE DE GANANCIA DE EXCITACIÓN POR CÓDIGO
ÍNDICE DE ALTURA
ÍNDICE DE GANANCIA DE LIBRO DE CÓDIGOS ADAPTATIVO
ÍNDICE DE PARÁMETROS STP
ÍNDICE DE EXCITACIÓN POR CÓDIGO
DECODIFICADOR DE GANANCIA
81
DECODIFICADOR DE ALTURA
84
DECODIFICADOR DE GANANCIA
85
DECODIFICADOR STP
83
403
401
$G_p$
404
LIBRO DE CÓDIGOS ADAPTATIVO
$e_p(n)$
ALTURA
PREDICCIÓN A CORTO PLAZO
VOZ FINAL
406
408
409
$e(n)$
1/A(z)
POSPROCESAR
407
VOZ SINTETIZADA
EXCITACIÓN POR CÓDIGO
$e_c(n)$
405
$G_c$
402

***Figura 6***

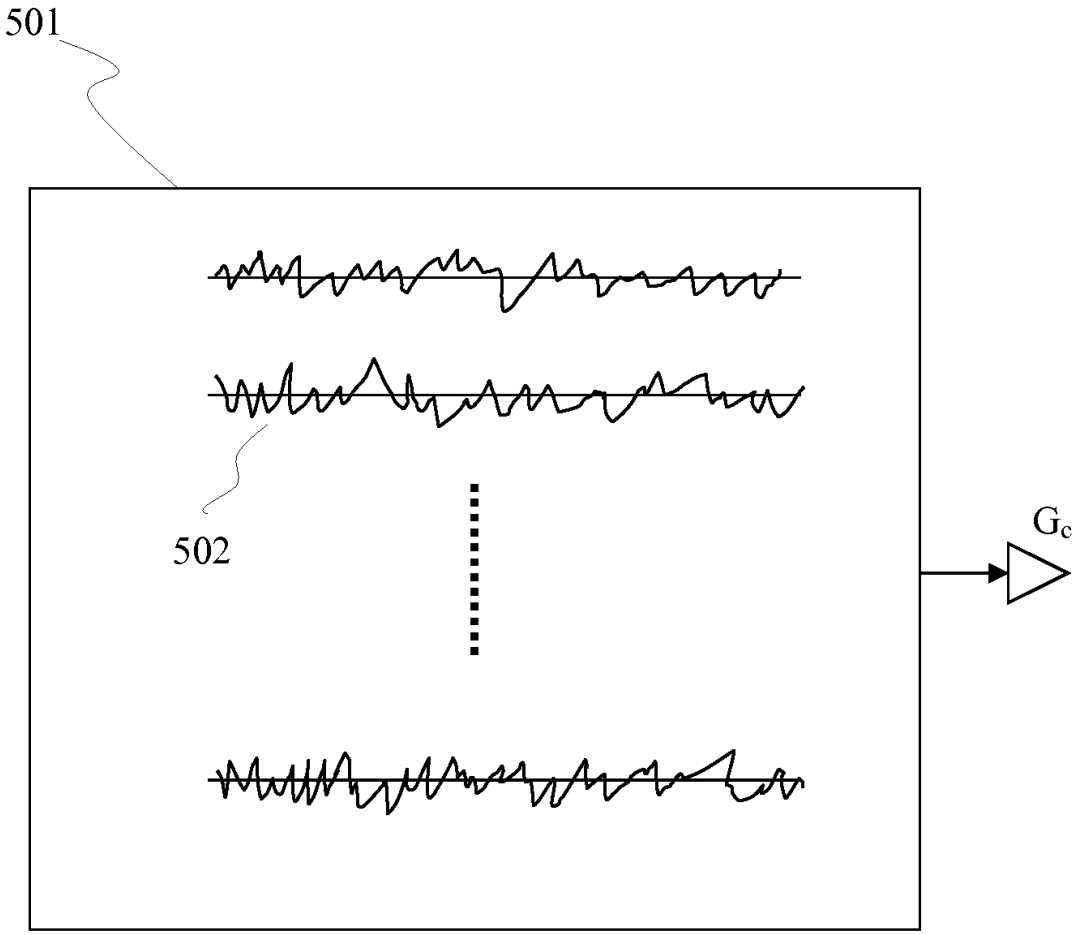


*Figura 7*

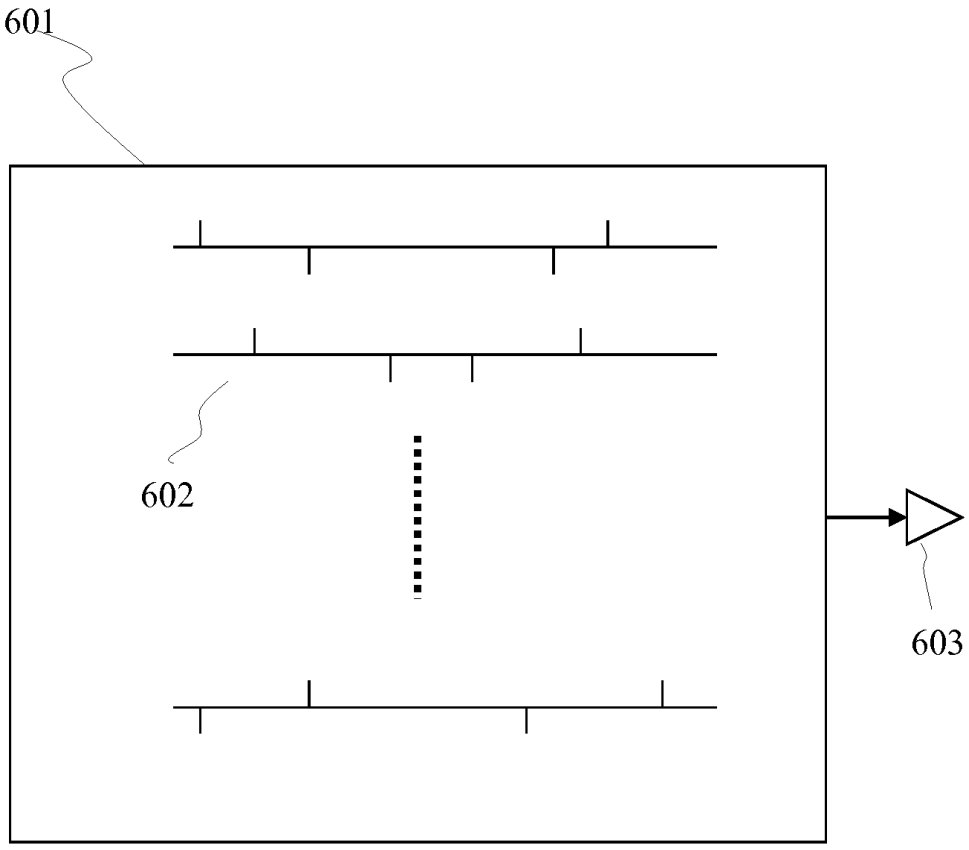


*Figura 8*

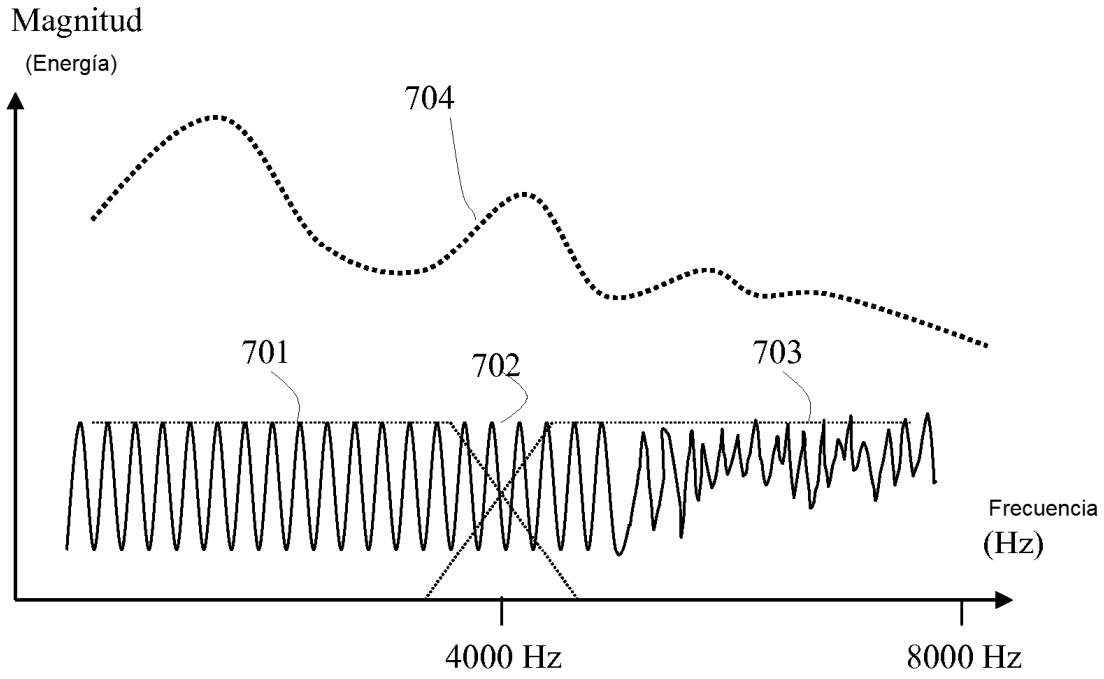
Magnitud
(Energía)
704
701
702
703
Frecuencia
(Hz)
4000 Hz
8000 Hz

***Figura 9***

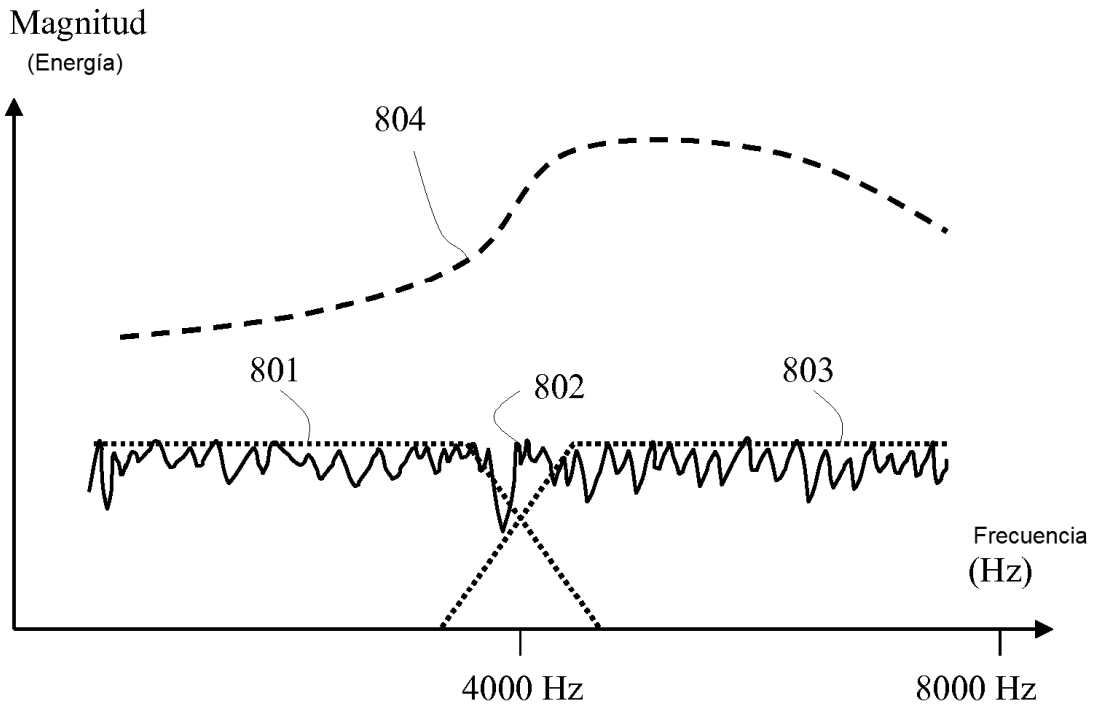
Magnitud
(Energía)
804
801
802
803
Frecuencia
(Hz)
4000 Hz
8000 Hz

***Figura 10***

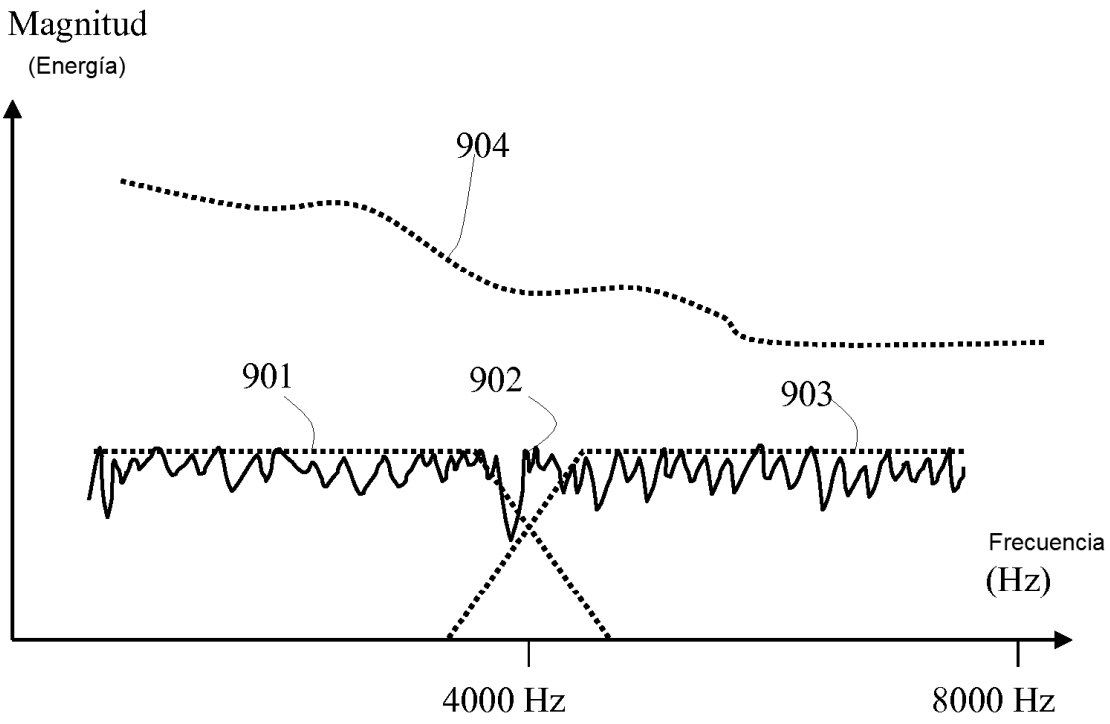
Magnitud
(Energía)
904
901
902
903
Frecuencia
(Hz)
4000 Hz
8000 Hz

***Figura 11***

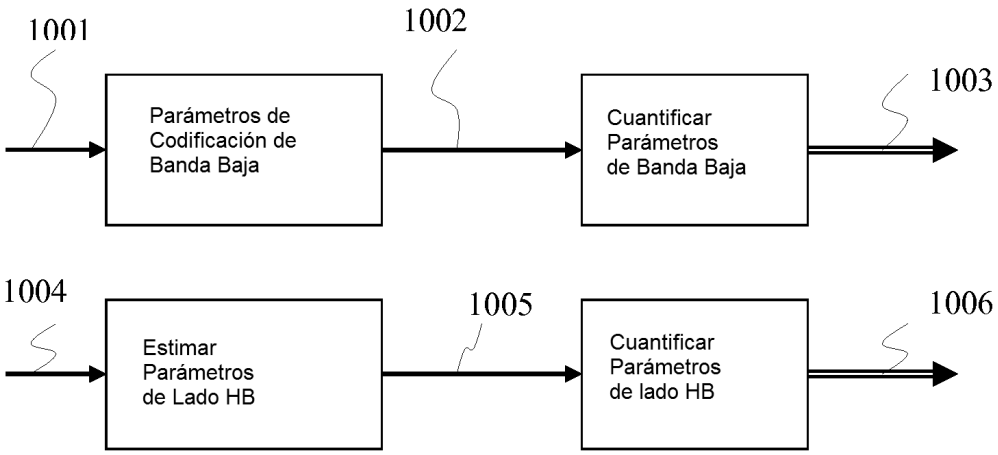


*Figura 12A*

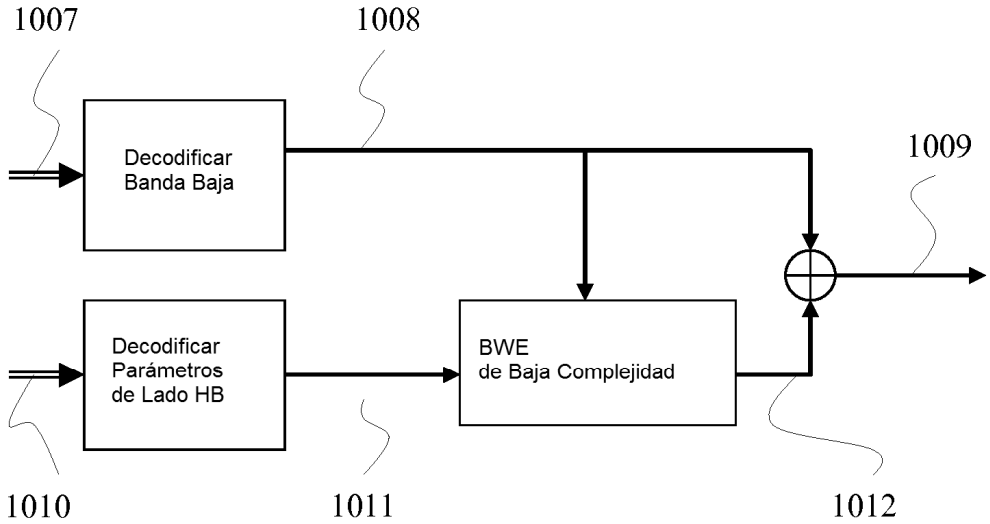


*Figura 12B*

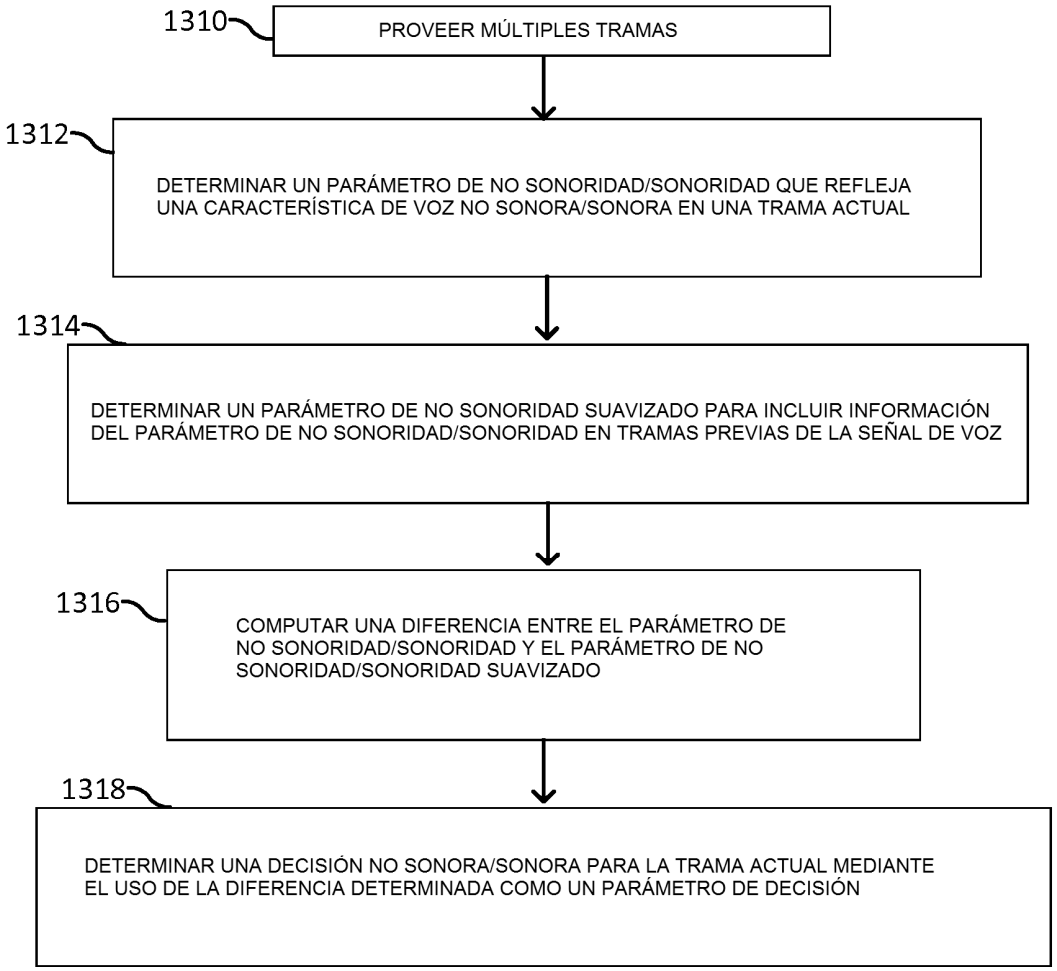
1310
PROVEER MÚLTIPLES TRAMAS
1312
DETERMINAR UN PARÁMETRO DE NO SONORIDAD/SONORIDAD QUE REFLEJA UNA CARACTERÍSTICA DE VOZ NO SONORA/SONORA EN UNA TRAMA ACTUAL
1314
DETERMINAR UN PARÁMETRO DE NO SONORIDAD SUAVIZADO PARA INCLUIR INFORMACIÓN DEL PARÁMETRO DE NO SONORIDAD/SONORIDAD EN TRAMAS PREVIAS DE LA SEÑAL DE VOZ
1316
COMPUTAR UNA DIFERENCIA ENTRE EL PARÁMETRO DE NO SONORIDAD/SONORIDAD Y EL PARÁMETRO DE NO SONORIDAD/SONORIDAD SUAVIZADO
1318
DETERMINAR UNA DECISIÓN NO SONORA/SONORA PARA LA TRAMA ACTUAL MEDIANTE EL USO DE LA DIFERENCIA DETERMINADA COMO UN PARÁMETRO DE DECISIÓN

*Fig. 13A*

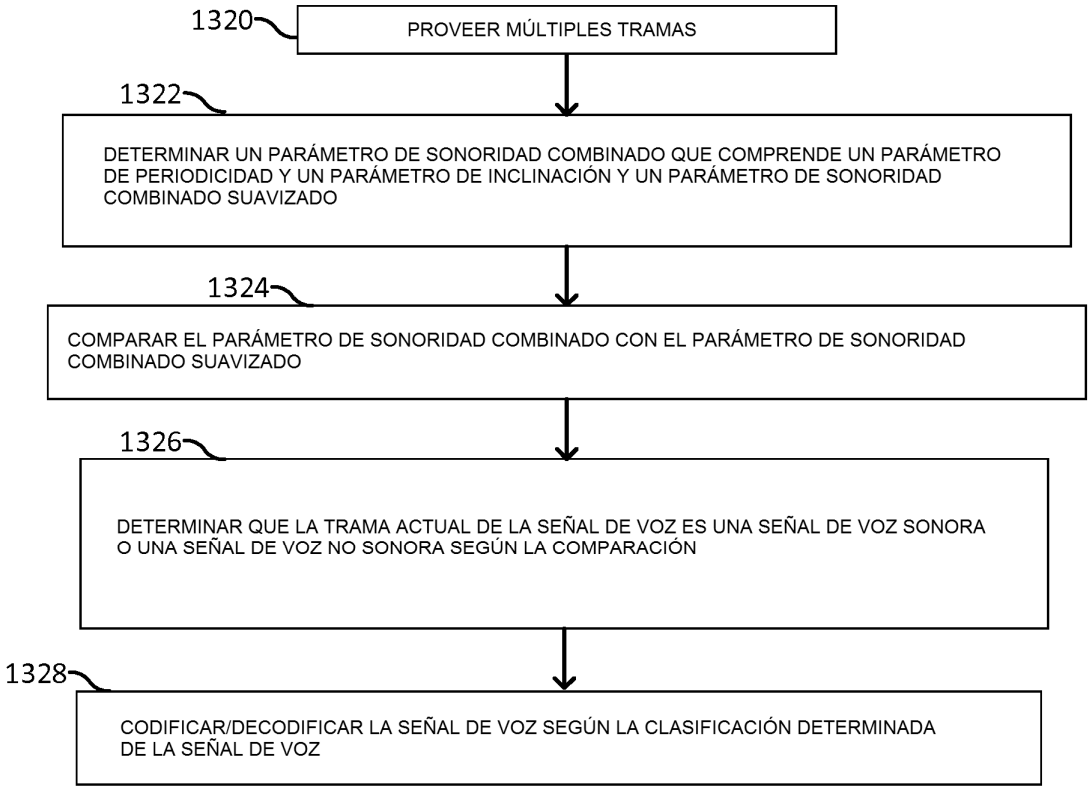
1320
PROVEER MÚLTIPLES TRAMAS
1322
DETERMINAR UN PARÁMETRO DE SONORIDAD COMBINADO QUE COMPRENDE UN PARÁMETRO DE PERIODICIDAD Y UN PARÁMETRO DE INCLINACIÓN Y UN PARÁMETRO DE SONORIDAD COMBINADO SUAVIZADO
1324
COMPARAR EL PARÁMETRO DE SONORIDAD COMBINADO CON EL PARÁMETRO DE SONORIDAD COMBINADO SUAVIZADO
1326
DETERMINAR QUE LA TRAMA ACTUAL DE LA SEÑAL DE VOZ ES UNA SEÑAL DE VOZ SONORA O UNA SEÑAL DE VOZ NO SONORA SEGÚN LA COMPARACIÓN
1328
CODIFICAR/DECODIFICAR LA SEÑAL DE VOZ SEGÚN LA CLASIFICACIÓN DETERMINADA DE LA SEÑAL DE VOZ

***Fig. 13B***

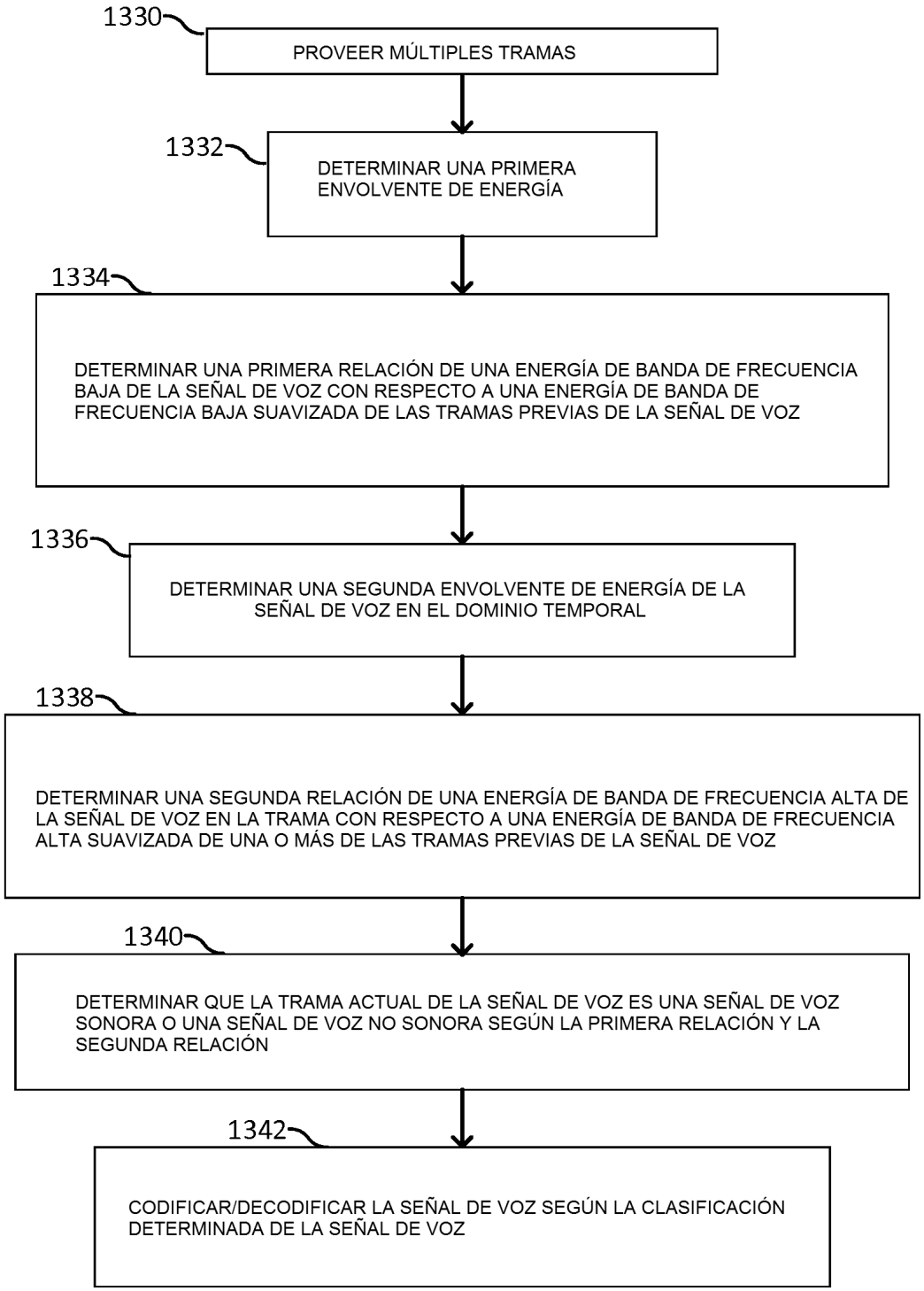
1330
PROVEER MÚLTIPLES TRAMAS
1332
DETERMINAR UNA PRIMERA ENVOLVENTE DE ENERGÍA
1334
DETERMINAR UNA PRIMERA RELACIÓN DE UNA ENERGÍA DE BANDA DE FRECUENCIA BAJA DE LA SEÑAL DE VOZ CON RESPECTO A UNA ENERGÍA DE BANDA DE FRECUENCIA BAJA SUAVIZADA DE LAS TRAMAS PREVIAS DE LA SEÑAL DE VOZ
1336
DETERMINAR UNA SEGUNDA ENVOLVENTE DE ENERGÍA DE LA SEÑAL DE VOZ EN EL DOMINIO TEMPORAL
1338
DETERMINAR UNA SEGUNDA RELACIÓN DE UNA ENERGÍA DE BANDA DE FRECUENCIA ALTA DE LA SEÑAL DE VOZ EN LA TRAMA CON RESPECTO A UNA ENERGÍA DE BANDA DE FRECUENCIA ALTA SUAVIZADA DE UNA O MÁS DE LAS TRAMAS PREVIAS DE LA SEÑAL DE VOZ
1340
DETERMINAR QUE LA TRAMA ACTUAL DE LA SEÑAL DE VOZ ES UNA SEÑAL DE VOZ SONORA O UNA SEÑAL DE VOZ NO SONORA SEGÚN LA PRIMERA RELACIÓN Y LA SEGUNDA RELACIÓN
1342
CODIFICAR/DECODIFICAR LA SEÑAL DE VOZ SEGÚN LA CLASIFICACIÓN DETERMINADA DE LA SEÑAL DE VOZ

*Fig. 13C*

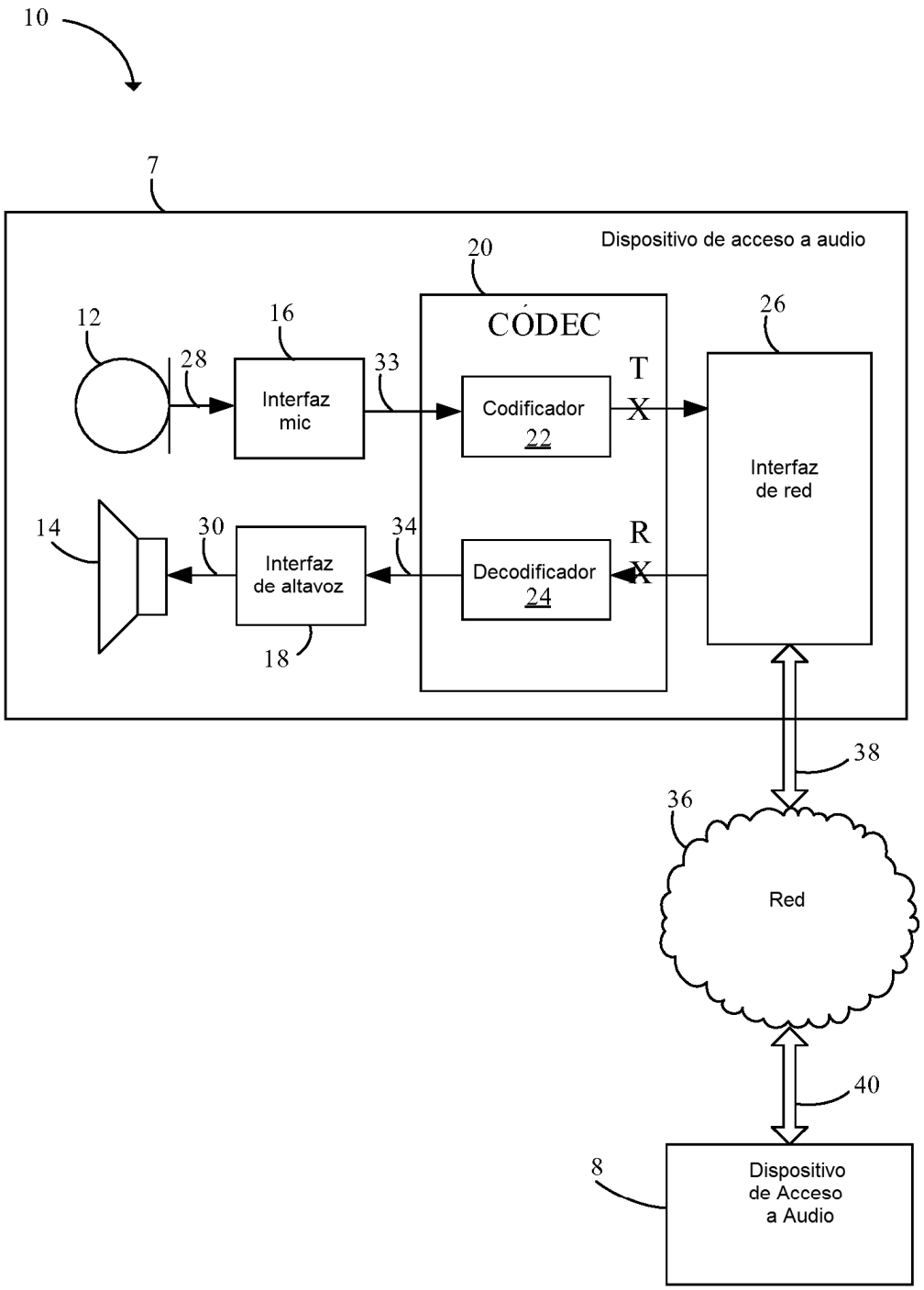
10
7
Dispositivo de acceso a audio
20
CÓDEC
12
16
26
28
33
Interfaz mic
Codificador 22
T X
Interfaz de red
14
30
34
Interfaz de altavoz
Decodificador 24
R X
18
38
36
Red
40
8
Dispositivo de Acceso a Audio

*Figura 14*

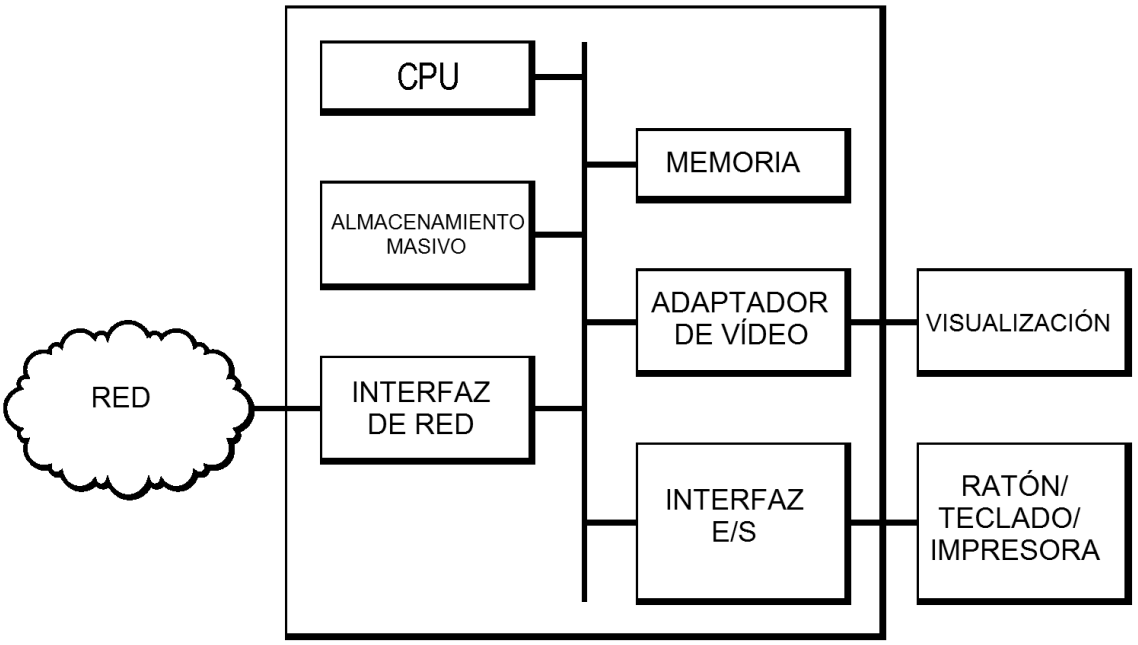
CPU
MEMORIA
ALMACENAMIENTO MASIVO
ADAPTADOR DE VÍDEO
VISUALIZACIÓN
RED
INTERFAZ DE RED
INTERFAZ E/S
RATÓN/ TECLADO/ IMPRESORA

***Figura 15***