

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 533 358**

51 Int. Cl.:

G10L 25/78 (2013.01)

G10L 19/22 (2013.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

96 Fecha de presentación y número de la solicitud europea: **20.06.2008** **E 08783143 (4)**

97 Fecha y número de publicación de la concesión europea: **24.12.2014** **EP 2162880**

54 Título: **Procedimiento y dispositivo para estimar la tonalidad de una señal de sonido**

30 Prioridad:

22.06.2007 US 929336 P

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

09.04.2015

73 Titular/es:

**VOICEAGE CORPORATION (100.0%)
750, chemin Lucerne Suite 250
Ville Mont-Royal, Quebec H3R 2H6, CA**

72 Inventor/es:

**MALENOWSKY, VLADIMIR;
JELINEK, MILAN;
VAILLANCOURT, TOMMY y
SALAMI, REDWAN**

74 Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

ES 2 533 358 T3

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

DESCRIPCIÓN

Procedimiento y dispositivo para estimar la tonalidad de una señal de sonido

5 Campo de la Invención

La presente invención se refiere a la detección de actividad sonora, la estimación de ruido de fondo y la clasificación de la señal de sonido, donde se entiende que sonido es una señal útil. La presente invención se refiere también al detector de actividad sonora, al estimador de ruido de fondo y al clasificador de señales de sonido correspondientes.

10 En particular, pero no de manera exclusiva:

– La detección de actividad sonora se usa para seleccionar las tramas a ser codificadas usando técnicas optimizadas para tramas inactivas.

15 – El clasificador de señales de sonido se usa para discriminar entre diferentes clases de señales de voz y música para permitir una codificación más eficiente de las señales de sonido, es decir, codificación optimizada de señales de voz sorda, codificación optimizada de señales de voz sonora estables, y codificación genérica de otras señales de sonido.

20 – Se proporciona un algoritmo y usa diversos parámetros y características pertinentes para permitir una mejor elección del modo de codificación y una estimación más robusta del ruido de fondo.

– La estimación de tonalidad se usa para mejorar el rendimiento de la detección de actividad sonora en presencia de señales musicales, y para discriminar mejor entre sonidos sordos y música. Por ejemplo, la estimación de la tonalidad puede ser usada en un códec de banda súper ancha para decidir el modelo de códec para codificar la señal por encima de 7 kHz.

25 Antecedentes de la Invención

La demanda de técnicas digitales de codificación de voz, eficientes, de banda estrecha y de banda ancha con un buen equilibrio entre la calidad subjetiva y la tasa de bits está aumentando en diversas áreas de aplicación, tales como teleconferencia, multimedia y comunicaciones inalámbricas. Hasta hace poco, el ancho de banda telefónico limitado a un intervalo de 200-3.400 Hz ha sido usado principalmente en aplicaciones de codificación de voz (señal muestreada a 8 kHz). Sin embargo, las aplicaciones de voz de banda ancha proporcionan una mayor inteligibilidad y naturalidad en la comunicación en comparación con el ancho de banda telefónico convencional. En los servicios de banda ancha, la señal de entrada es muestreada a 16 kHz y el ancho de banda codificado está en el intervalo de 50 a 7.000 Hz. Se ha encontrado que este ancho de banda es suficiente para suministrar una buena calidad dando una impresión de casi una comunicación cara a cara. Se consigue una mejora adicional de la calidad con la denominada banda súper ancha, en la que la señal es muestreada a 32 kHz y el ancho de banda codificado está comprendido en el intervalo de 50 a 15.000 Hz. Para las señales de voz, esto proporciona una calidad cara a cara, ya que casi toda la energía en la voz humana es inferior a 14.000 Hz. Este ancho de banda proporciona también una mejora de calidad considerable con respecto a las señales de audio generales incluyendo música (la banda ancha es equivalente a la radio AM y la banda súper ancha es equivalente a la radio FM). Se ha usado un ancho de banda más alto para señales de audio generales con la banda completa 20- 20.000 Hz (calidad de CD muestreada a 44,1 kHz o 48 kHz).

45 Un codificador de sonido convierte una señal de sonido (voz o audio) en un flujo de bits digital que es transmitido a través de un canal de comunicación o es almacenado en un medio de almacenamiento. La señal de sonido es digitalizada, es decir, es muestreo y cuantificada generalmente con 16 bits por muestra. El codificador de sonido desempeña el papel de representar estas muestras digitales con un número de bits más pequeño mientras mantiene una buena calidad subjetiva. El decodificador de sonido opera sobre el flujo el flujo de bits transmitido o almacenado y lo convierte de nuevo a una señal de sonido.

50 La codificación basada en predicción lineal con excitación por código (Code-Excited Linear Prediction, CELP) es una de las mejores técnicas de la técnica anterior para conseguir un buen compromiso entre la calidad subjetiva y la tasa de bits. Esta técnica de codificación es una base de varios estándares de codificación de voz, en aplicaciones tanto inalámbricas como de línea fija. En la codificación CELP, la señal de voz muestreada es procesada en bloques sucesivos de L muestras, denominadas generalmente tramas, donde L es un número predeterminado que corresponde típicamente a 10-30 ms. Se calcula y se transmite un filtro de predicción lineal (LP) cada trama. La trama de la muestra L es dividida en bloques más pequeños denominados subtramas. En cada subtrama, normalmente se obtiene una señal de excitación a partir de dos componentes, la excitación pasada y la excitación de libro de códigos fijo, innovadora. El componente formado a partir de la excitación pasada se denomina frecuentemente libro de códigos adaptativo o excitación de tono. Los parámetros que caracterizan la señal de excitación se codifican y transmiten al decodificador, donde la señal de excitación reconstruida es usada

como la entrada al filtro LP.

El uso de codificación de voz de tasa de bits variable (Variable Bit Rate, VBR) con fuente controlada mejora considerablemente la capacidad del sistema. En la codificación VBR con fuente controlada, el códec usa un módulo de clasificación de señal y se usa un modelo de codificación optimizado para codificar cada trama de voz en base a la naturaleza de la trama de voz (por ejemplo, sonora, sorda, transitoria, ruido de fondo). Además, pueden usarse diferentes tasas de bits para cada clase. La forma más simple de codificación VBR con fuente controlada es usar detección de actividad de voz (Voice Activity Detection, VAD) y codificar las tramas de voz inactivas (ruido de fondo) a una tasa de bits muy baja. Puede usarse además transmisión discontinua (Discontinuous transmission, DTX) donde no se transmiten datos en el caso de ruido de fondo estable. El decodificador usa generación de ruido de confort (Comfort Noise Generation, CNG) para generar las características de ruido de fondo. VAD/DTX/CNG resulta en una considerable reducción en la tasa de bits media y en aplicaciones de conmutación de paquetes reduce considerablemente el número de paquetes enrutados. Los algoritmos VAD funcionan bien con las señales de voz, pero pueden resultar en graves problemas en caso de señales musicales. Los segmentos de señales musicales pueden ser clasificados como señales sordas y, por consiguiente, pueden ser codificadas con un modelo optimizado para señales sordas que afecta gravemente a la calidad de la música. Además, algunos segmentos de señales musicales estables pueden ser clasificados como ruido de fondo estable y esto puede provocar la actualización de ruido de fondo en el algoritmo VAD, lo que resulta en la degradación del rendimiento del algoritmo. Por lo tanto, sería ventajoso extender el algoritmo VAD para discriminar mejor las señales musicales. En la presente descripción, este algoritmo se denominará algoritmo de Detección de actividad sonora (Sound Activity Detection, SAD) en el que el sonido podría ser habla o música o cualquier señal útil. La presente descripción describe también un procedimiento de detección de tonalidad usado para mejorar el rendimiento del algoritmo de SAD en el caso de señales musicales.

Otro aspecto en la codificación del habla y de audio es el concepto de codificación incorporada, conocida también como codificación por capas. En la codificación incorporada, la señal es codificada en una primera capa para producir un primer flujo de bits y, a continuación, el error entre la señal original y la señal codificada de la primera capa es codificado adicionalmente para producir un segundo flujo de bits. Esto puede repetirse para más capas codificando el error entre la señal original y la señal codificada de todas las capas anteriores. Los flujos de bits de todas las capas son concatenados para la transmisión. La ventaja de la codificación por capas es que pueden desecharse partes del flujo de bits (correspondientes a las capas superiores) en la red (por ejemplo, en caso de congestión) todavía pudiendo decodificarse la señal en el receptor dependiendo del número de capas recibidas. La codificación por capas es útil también en aplicaciones de multidifusión en las que el codificador produce el flujo de bits de todas las capas y la red decide enviar a diferentes tasas de bits a diferentes puntos finales en función de la tasa de bits disponible en cada enlace.

La codificación incorporada o por capas puede ser útil también para mejorar la calidad de los códecos existentes usados ampliamente, mientras todavía se mantiene la interoperabilidad con estos códecos. La adición de más capas a la capa de núcleo del códec estándar puede mejorar la calidad e incluso aumentar el ancho de banda de la señal de audio codificada. Los ejemplos son la Recomendación UIT-T G.729.1 recientemente estandarizada, en la que la capa de núcleo es interoperable con el estándar G.729 de banda estrecha usado ampliamente a 8 kbit/s y capas superiores producen tasas de bits de hasta 32 kbit/s (con una señal de banda ancha a partir de 16 kbit/s). El objetivo del trabajo de normalización actual es añadir más capas para producir un códec de banda súper ancha (ancho de banda de 14 kHz) y extensiones estéreo. Otro ejemplo es la Recomendación UIT-T G.718 para la codificación de señales de banda súper ancha a 8, 12, 16, 24 y 32 kbit/s. El códec está siendo ampliando también para codificar señales de banda súper ancha y estéreo con tasas de bits más altas.

Los requisitos para los códecos integrados normalmente exigen una buena calidad tanto en señales de audio como en señales de voz. Debido a que el habla puede ser codificada en tasas de bits relativamente bajas usando un enfoque basado en un modelo, la primera capa (o las primeras dos capas) es codificada (o son codificadas) usando una técnica específica para el habla y la señal de error para las capas superiores es codificada usando una técnica de codificación de audio más genérica. Esto proporciona una buena calidad de habla a tasas de bits bajas y una buena calidad de audio conforme se aumenta la tasa de bits. En G.718 y G.729.1, las primeras dos capas se basan en la técnica ACELP (Algebraic Code-Excited Linear Prediction) que es adecuada para la codificación de señales de voz. En las capas superiores, se usa una codificación basada en transformación adecuada para las señales de audio para codificar la señal de error (la diferencia entre la señal original y la salida de las dos primeras capas). Se usa la conocida transformada MDCT (transformada modificada discreta de coseno, Modified Discrete Cosine Transform), en la que la señal de error es transformada en el dominio de la frecuencia. En las capas de banda súper ancha, la señal superior a los 7 kHz es codificada usando un modelo de codificación genérico o un modelo de codificación tonal. La detección de tonalidad indicada anteriormente puede ser usada también para seleccionar el modelo de codificación apropiado a ser usado.

Un ejemplo de un procedimiento un aparato conocidos para determinar la tonalidad de una señal de audio de entrada se describe en el documento de patente US 2004/181393 A1.

5 Compendio de la Invención

Según un primer aspecto de la presente invención, se proporciona un procedimiento para estimar una tonalidad de una señal de sonido. El procedimiento comprende: calcular un espectro residual actual de la señal de sonido; detectar picos en el espectro residual actual; calcular un mapa de correlación entre el espectro residual actual y un espectro residual anterior para cada pico detectado; y calcular un mapa de correlación a largo plazo basado en el mapa de correlación calculado, en el que el mapa de correlación a largo plazo es indicativo de una tonalidad en la señal de sonido.

Según un segundo aspecto de la presente invención, se proporciona un dispositivo para estimar una tonalidad de una señal de sonido. El dispositivo comprende: un calculador de un espectro residual actual de la señal de sonido; un detector para detectar picos en el espectro residual actual; un calculador para calcular un mapa de correlación entre el espectro residual actual y un espectro residual anterior para cada pico detectado; y un calculador para calcular un mapa de correlación a largo plazo basado en el mapa de correlación calculado, en el que el mapa de correlación a largo plazo es indicativo de una tonalidad en la señal de sonido.

Lo indicado anteriormente y otros objetos, ventajas y características de la presente invención serán más evidentes tras la lectura de la descripción no restrictiva siguiente de una realización ilustrativa de la misma, proporcionada solamente a modo de ejemplo con referencia a los dibujos adjuntos.

Breve descripción de los dibujos

En los dibujos adjuntos:

La Figura 1 es un diagrama de bloques esquemático de una parte de un ejemplo de sistema de comunicación de sonido que incluye detección de actividad sonora, actualización de estimación de ruido de fondo y clasificación de señal de sonido;

La Figura 2 es una ilustración no limitativa de uso de ventanas en el análisis espectral;

La Figura 3 es una ilustración gráfica no restrictiva del principio de cálculo de suelo espectral y el espectro residual;

La Figura 4 es una ilustración no limitativa de cálculo de mapa de correlación espectral en una trama actual;

La Figura 5 es un ejemplo de diagrama de bloques funcional de un algoritmo de clasificación de señal; y

La Figura 6 es un ejemplo de árbol de decisión para discriminación de voz sorda.

Descripción detallada

En la realización ilustrativa, no limitativa, de la presente invención, se realiza una detección de actividad sonora (SAD) dentro de un sistema de comunicación de sonido para clasificar tramas de tiempo corto de las señales como sonido o ruido de fondo/silencio. La detección de actividad sonora se basa en una relación de señal-ruido (SNR) dependiente de la frecuencia y usa una energía de ruido de fondo estimada por cada banda crítica. Una decisión sobre la actualización del estimador de ruido de fondo se basa en diversos parámetros, incluyendo parámetros que discriminan entre el ruido de fondo/silencio y música, impidiendo, de esta manera, la actualización del estimador de ruido de fondo en las señales musicales.

SAD corresponde a una primera etapa de la clasificación de señal. Esta primera etapa se usa para discriminar tramas inactivas para una codificación optimizada de señal inactiva. En una segunda etapa, las tramas de voz sorda son discriminadas para una codificación optimizada de una señal sorda. En esta segunda etapa, se añade la detección de música con el fin de prevenir que la música sea clasificada como señal sorda. Finalmente, en una tercera etapa, las señales sonoras son discriminadas mediante un examen adicional de los parámetros de trama.

Las técnicas descritas en la presente memoria pueden implementarse con señales de sonido de banda estrecha (Narrow Band, NB) muestreadas a 8.000 muestras/s o señales de sonido de banda ancha (Wide Banda, WB) muestreadas a 16.000 muestras/s, o a cualquier otra frecuencia de muestreo. El codificador usado en la realización ilustrativa no limitativa de la presente invención se basa en los códecs AMR-WB [AMR Wideband Speech Codec: Transcoding Functions, 3GPP Technical Specification TS 26.190 (<http://www.3gpp.org>)] y VMR-WB [Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB), Service Options 62 and 63 for Spread Spectrum Systems, 3GPP2 Technical Specification C.S0052-A v1.0, Abril de 2005 (<http://www.3gpp2.org>)] que usan una conversión de muestreo interna para convertir la frecuencia de muestreo de señal a 12.800 muestras/s (funcionando en un ancho de banda de 6,4 kHz). De esta manera, la técnica de detección de actividad sonora en la realización ilustrativa no restrictiva opera sobre señales de banda estrecha o de banda ancha después

de una conversión de muestreo a 12,8 kHz.

La Figura 1 es un diagrama de bloques de un sistema 100 de comunicación de sonido según la realización ilustrativa no limitativa de la invención, que incluye detección de actividad sonora.

El sistema 100 de comunicación de sonido de la Figura 1 comprende un pre-procesador 101. El preprocesamiento realizado por el módulo 101 puede ser realizado tal como se describe en el ejemplo siguiente (filtrado pasa alta, re-muestreo y pre-énfasis).

Antes de la conversión de frecuencia, la señal de sonido de entrada es filtrada con un filtro pasa alta. En esta realización ilustrativa, no restrictiva, la frecuencia de corte del filtro pasa alta es de 25 Hz para WM y de 100 Hz para NB. El filtro pasa alta sirve como precaución contra componentes de baja frecuencia no deseados. Por ejemplo, puede usarse la función de transferencia siguiente:

$$H_{h1}(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

en la que, para WB, $b_0 = 0,9930820$, $b_1 = -1,98616407$, $b_2 = 0,9930820$, $a_1 = -1,9861162$, $a_2 = 0,9862119292$ y, para NB, $b_0 = 0,945976856$, $b_1 = -1,891953712$, $b_2 = 0,945976856$, $a_1 = -1,889033079$, $a_2 = 0,894874345$. Obviamente, el filtrado de paso alto puede llevarse a cabo, de manera alternativa, después de re-muestrear a 12,8 kHz.

En el caso de WB, la señal de sonido de entrada es diezmada desde 16 kHz a 12,8 kHz. El diezmado es realizado por un sobremuestreador que sobremuestra la señal de sonido con un factor de 4. A continuación, la salida resultante es filtrada a través de un filtro pasa baja FIR (respuesta a impulso finito, Finite Impulse Response) con una frecuencia de corte de 6,4 kHz. A continuación, la señal filtrada con el filtro pasa baja es submuestreada con un factor de 5 por un submuestreador apropiado. El retardo del filtrado es de 15 muestras a una frecuencia de muestreo de 16 kHz.

En el caso de NB, la señal de sonido es sobremuestreada de 8 kHz a 12,8 kHz. Para ese propósito, una sobremuestreador realiza un sobremuestreo con un factor de 8 sobre el sonido. A continuación, la salida resultante es filtrada a través de un filtro FIR pasa baja con una frecuencia de corte a 6,4 kHz. A continuación, un submuestreador submuestra la señal filtrada de paso bajo en un factor de 5. El retardo del filtrado es de 16 muestras a una frecuencia de muestreo de 8 kHz.

Después de la conversión de muestreo, se aplica un pre-énfasis a la señal de sonido antes del procedimiento de codificación. En el pre-énfasis, se usa un filtro pasa alta de primer orden para enfatizar las frecuencias más altas. Este filtro pasa alta de primer orden forma un pre-enfatizador y usa, por ejemplo, la función de transferencia siguiente:

$$H_{\text{pre-enf}}(z) = 1 - 0,68z^{-1}$$

El pre-énfasis se usa para mejorar el rendimiento del códec a altas frecuencias y para mejorar la ponderación perceptual en el procedimiento de minimización de errores usado en el codificador.

Tal como se ha descrito anteriormente, la señal de sonido de entrada es convertida a una frecuencia de muestreo de 12,8 kHz y es preprocesada, por ejemplo, tal como se ha descrito anteriormente. Sin embargo, las técnicas descritas pueden ser aplicadas igualmente a señales a otras frecuencias de muestreo, tales como 8 kHz o 16 kHz, con un preprocesamiento diferente o sin preprocesamiento.

En la realización ilustrativa no limitativa de la presente invención, el codificador 109 (Figura 1) que usa detección de actividad sonora opera sobre tramas de 20 ms que contienen 256 muestras a la frecuencia de muestreo de 12,8 kHz. Además, el codificador 109 utiliza una anticipación de 10 ms de la trama futura para realizar su análisis (Figura 2). La detección de actividad sonora sigue la misma estructura de trama.

Con referencia a la Figura 1, el análisis espectral es realizado en el analizador 102 espectral. Se realizan dos

análisis en cada trama usando ventanas de 20 ms con una superposición del 50%. El principio de las ventanas se ilustra en la Figura 2. Se calcula la energía de la señal para los contenedores de frecuencia y las bandas críticas [J. D. Johnston, " Transform coding of audio signal using perceptual noise criteria," IEEE J. Select. Areas Commun., vol. 6, pp. 314-323, Febrero1988].

La detección de actividad sonora (primera etapa de la clasificación de señal) es realizada en el detector 103 de actividad sonora usando estimaciones de energía de ruido calculadas en la trama previa. La salida del detector 103 de actividad sonora es una variable binaria que es usada además por el codificador 109 y que determina si la trama actual está codificada como activa o inactiva.

El estimador 104 de ruido actualiza una estimación de ruido hacia abajo (primer nivel de estimación y actualización de ruido), es decir, si en una banda crítica la energía de la trama es más baja que una energía estimada del ruido de fondo, la energía de la estimación de ruido es actualizada en esa banda crítica.

La reducción de ruido es aplicada opcionalmente por un reductor 105 de ruido opcional a la señal de voz usando, por ejemplo, un procedimiento de sustracción espectral. Un ejemplo de dicho un esquema de reducción de ruido se describe en [M. Jelinek y R. Salami, " Noise Reduction Method for Wideband Speech Coding", en Proc. EUSIPCO, Viena, Austria, septiembre de 2004].

Un analizador de LP y un rastreador 106 de tono realizan un análisis de predicción lineal (LP) y un análisis de tono en bucle abierto (normalmente como parte del algoritmo de codificación de voz). En esta realización ilustrativa no restrictiva, los parámetros resultantes del analizador LP y el rastreador 106 de tono se usan en la decisión para actualizar las estimaciones de ruido en las bandas críticas tal como se realizan en el módulo 107. De manera alternativa, el detector 103 de actividad sonora puede ser usado también para tomar la decisión de actualización de ruido. Según una alternativa adicional, las funciones implementadas por el analizador LP y el rastreador 106 de tono pueden ser una parte integral del algoritmo de codificación de sonido.

Antes de actualizar las estimaciones de energía de ruido en el módulo 107, se realiza una detección de música para prevenir una falsa actualización sobre las señales musicales activas. La detección de música usa los parámetros espectrales calculados por el analizador 102 espectral.

Finalmente, las estimaciones de energía de ruido son actualizadas en el módulo 107 (segundo nivel de estimación y actualización de ruido). Este módulo 107 utiliza todos los parámetros disponibles calculados previamente en los módulos 102 a 106 para decidir acerca de la actualización de las energías de la estimación de ruido.

En el clasificador 108 de señal, la señal de sonido es clasificada adicionalmente como sorda, sonora estable o genérica. Se calculan varios parámetros para apoyar esta decisión. En este clasificador de señales, el modo de codificación de la señal de sonido de la trama actual se elige de manera que represente de la mejor manera posible la clase de señal que está siendo codificada.

El codificador 109 de sonido realiza la codificación de la señal de sonido en base al modo de codificación seleccionado en el clasificador 108 de señal de sonido. En otras aplicaciones, el clasificador 108 de señal de sonido puede ser un sistema de reconocimiento de voz automático.

Análisis espectral

El análisis espectral es realizado por el analizador 102 espectral de la Figura 1.

Se usa la transformada de Fourier para realizar el análisis espectral y la estimación de la energía espectral. El análisis espectral se realiza dos veces por cada trama usando una Transformada Rápida de Fourier (FFT) de 256 puntos con un solapamiento del 50 por ciento (como se ilustra en la Figura 2). Las ventanas de análisis se colocan de manera que se aproveche toda la anticipación. El comienzo de la primera ventana está en el comienzo de la trama actual del codificador. La segunda ventana se coloca 128 muestras más allá. Se ha usado una ventana de Hanning de raíz cuadrada (que es equivalente a una ventana sinusoidal) para ponderar la señal de sonido de entrada para el análisis espectral. Esta ventana es particularmente adecuada para procedimientos de solapamiento-suma (de esta manera, este análisis espectral particular se usa en la supresión de ruido basada en la sustracción espectral y análisis/síntesis de solapamiento-suma). La ventana de Hanning de raíz cuadrada viene dada por:

$$w_{FFT}(n) = \sqrt{0,5 - 0,5 \cos\left(\frac{2\pi n}{L_{FFT}}\right)} = \text{sen}\left(\frac{\pi n}{L_{FFT}}\right), \quad n = 0, \dots, L_{FFT} - 1 \quad (1)$$

en la que $L_{FFT} = 256$ es el tamaño del análisis FFT. Aquí, sólo se calcula y se almacena la mitad de la ventana ya que esta ventana es simétrica (de 0 a $L_{FFT}/2$).

Las señales de ventana para ambos análisis espectrales (análisis espectrales primero y segundo) se obtienen usando las dos relaciones siguientes:

$$x_w^{(1)}(n) = w_{FFT}(n)s'(n), \quad n = 0, \dots, L_{FFT} - 1$$

$$x_w^{(2)}(n) = w_{FFT}(n)s'(n + L_{FFT} / 2), \quad n = 0, \dots, L_{FFT} - 1$$

En las que $s'(0)$ es la primera muestra en la trama actual. En la realización ilustrativa no limitativa de la presente invención, el comienzo de la primera ventana es colocado al comienzo de la trama actual. La segunda ventana es colocada 128 muestras más allá.

La FFT se realiza en ambas señales de ventana para obtener los dos conjuntos de parámetros espectrales siguientes por cada trama:

$$X^{(1)}(k) = \sum_{n=0}^{N-1} x_w^{(1)}(n) e^{-j2\pi \frac{kn}{N}}, \quad k = 0, \dots, L_{FFT} - 1$$

$$X^{(2)}(k) = \sum_{n=0}^{N-1} x_w^{(2)}(n) e^{-j2\pi \frac{kn}{N}}, \quad k = 0, \dots, L_{FFT} - 1$$

en las que $N = L_{FFT}$.

La FFT proporciona las partes real e imaginaria del espectro denotado por $X_R(k)$, $k = 0$ a 128, y $X_I(k)$, $k = 1$ a 127. $X_R(0)$ corresponde al espectro a 0 Hz (DC) y $X_R(128)$ corresponde al espectro a 6.400 Hz. El espectro en estos puntos solo tiene valores reales.

Después del análisis FFT, el espectro resultante es dividido en bandas críticas usando los intervalos que tienen los siguientes límites superiores [M. Jelinek y R. Salami, " Noise Reduction Method for Wideband Speech Coding", en Proc. Eusipco, Viena, Austria, Septiembre de 2004] (20 bandas en el intervalo de frecuencias 0-6.400 Hz):

Bandas críticas = {100,0, 200,0, 300,0, 400,0, 510,0, 630,0, 770,0, 920,0, 1.080,0, 1.270,0, 1.480,0, 1.720,0, 2.000,0, 2.320,0, 2.700,0, 3.150,0, 3.700,0, 4.400,0, 5.300,0, 6.350,0} Hz.

La FFT de 256 puntos resulta en una resolución de frecuencia de 50 Hz (6.400/128). De esta manera, después de ignorar la componente DC del espectro, el número de contenedores de frecuencia por cada banda crítica es $M_{CB} = \{2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 6, 6, 8, 9, 11, 14, 18, 21\}$, respectivamente.

La energía media en una banda crítica es calculada usando la relación siguiente:

$$E_{CB}(i) = \frac{1}{(L_{FFT}/2)^2 M_{CB}(i)} \sum_{k=0}^{M_{CB}(i)-1} \left(X_R^2(k + j_i) + X_I^2(k + j_i) \right) \quad i = 0, \dots, 19 \quad (2)$$

en la que $K_R(k)$ y $X_I(k)$ son, respectivamente, las partes real e imaginaria del k -ésimo contenedor de frecuencias y j_i es el índice del primer contenedor en la i -ésima banda crítica proporcionada por $j_i = \{1, 3, 5, 7, 9, 11, 13, 16, 19, 22, 26, 30, 35, 41, 47, 55, 64, 75, 89, 107\}$.

El analizador 102 espectral calcula también la energía normalizada por cada contenedor de frecuencia, $E_{CONT}(k)$, en el intervalo de 0 a 6.400 Hz, usando la relación siguiente:

$$E_{CONT}(k) = \frac{4}{L_{FFT}^2} \left(X_R^2(k) + X_I^2(k) \right), \quad k = 1, \dots, 127 \quad (3)$$

Además, los espectros de energía por cada contenedor de frecuencia en ambos análisis se combinan entre sí para obtener el espectro logarítmico de energía media (en decibelios), es decir,

$$E_{dB}(k) = 10 \log \left[\frac{1}{2} \left(E_{CONT}^{(1)}(k) + E_{CONT}^{(2)}(k) \right) \right], \quad k = 1, \dots, 127, \quad (4)$$

en la que los superíndices (1) y (2) se usan para denotar los análisis espectrales primero y segundo, respectivamente.

Finalmente, el analizador 102 espectral calcula la energía total media para ambos análisis espectral primero y segundo en una trama de 20 ms sumando las energías medias de las bandas críticas E_{CB} . Es decir, la energía espectral para un determinado análisis espectral se calcula utilizando la relación siguiente:

$$E_{trama} = \sum_{i=0}^{19} E_{CB}(i) \quad (5)$$

y la energía de trama total se calcula como la media de las energías espectrales de ambos análisis espectrales primero y segundo en una trama. Es decir

$$E_t = 10 \log(0,5(E_{trama}(0) + E_{trama}(1))) \text{ , dB.} \quad (6)$$

Los parámetros de salida del analizador 102 espectral, es decir, la energía media por cada banda crítica, la energía por cada contenedor de frecuencias y la energía total, se usan en el detector 103 de actividad sonora, y en la selección de la tasa. El espectro logarítmico de energía media se usa en la detección de música.

En las señales de entrada de banda estrecha muestreadas a 8.000 muestras/s, después de una conversión de muestreo a 12.800 muestra/s, no hay contenido en ambos extremos del espectro, de manera que la primera banda crítica de frecuencia inferior así como las tres últimas bandas de alta frecuencia no se consideran en el cálculo de los parámetros relevantes (sólo se consideran las bandas de $i = 1$ a 16). Sin embargo, las ecuaciones (3) y (4) no se ven afectadas.

Detección de actividad sonora (SAD)

La detección de actividad sonora es realizada por el detector 103 de actividad sonora en base a la SNR de la Figura 1.

El analizador 102 realiza el análisis espectral descrito anteriormente dos veces por cada trama. Considérese que $E_{CB}^{(1)}(i)$ y $E_{CB}^{(2)}(i)$, tal como se calculan en la ecuación (2), denotan la energía media por cada información de banda crítica en los análisis espectrales primero y segundo, respectivamente. La energía media por cada banda crítica para toda la trama y parte de la trama anterior se calcula usando la relación siguiente:

$$E_{med}(i) = 0,2E_{CB}^{(0)}(i) + 0,4E_{CB}^{(1)}(i) + 0,4E_{CB}^{(2)}(i) \quad (7)$$

en la que $E_{CB}^{(0)}(i)$ denota la energía por cada información de banda crítica a partir del segundo análisis espectral de la trama anterior. A continuación, la relación señal-ruido (SNR) por cada banda crítica es calculada usando la relación siguiente:

$$SNR_{CB}(i) = E_{med}(i) / N_{CB}(i) \quad \text{con} \quad SNR_{CB} \geq 1. \quad (8)$$

en la que $N_{CB}(i)$ es la energía de ruido estimada por cada banda crítica, tal como se explicará más adelante. A continuación, la SNR media por cada trama se calcula como

$$SNR_{med} = 10 \log \left(\sum_{i=b_{min}}^{b_{max}} SNR_{CB}(i) \right), \quad (9)$$

en la que $b_{min} = 0$ y $b_{max} = 19$ en el caso de señales de banda ancha, y $b_{min} = 1$ y $b_{max} = 16$ en el caso de señales de banda estrecha.

La actividad sonora es detectada comparando la SNR media por cada trama con un umbral determinado que es una función de la SNR a largo plazo. La SNR a largo plazo viene dada por la relación siguiente:

$$SNR_{LT} = \bar{E}_f - \bar{N}_f \quad (10)$$

en la que $\bar{E}_f$ y $\bar{N}_f$ se calculan usando las ecuaciones (13) y (14), respectivamente, que se describirán más adelante. El valor inicial de $\bar{E}_f$ es 45 dB.

El umbral es una función lineal a intervalos de la SNR a largo plazo. Se usan dos funciones, una optimizada para voz limpia y una optimizada para voz ruidosa.

Para señales de banda ancha, si $SNR_{LT} < 35$ (voz ruidosa), entonces el umbral es igual a:

$$umbral_{SAD} = 0,41287 SNR_{LT} + 13,259625$$

si no (voz limpia):

$$umbral_{SAD} = 1,0333 SNR_{LT} - 18$$

Para las señales de banda estrecha, si $SNR_{LT} < 20$ (voz ruidosa), entonces el umbral es igual a:

$$umbral_{SAD} = 0,1071 SNR_{LT} + 16,5$$

si no (voz limpia):

$$umbral_{SAD} = 0,4773 SNR_{LT} - 6,1364$$

Además, se añade una histéresis en la decisión SAD para evitar una conmutación frecuente al final de un período de sonido activo. La estrategia de histéresis es diferente para las señales de banda ancha y de banda estrecha y entra en vigor sólo si la señal es ruidosa.

Para señales de banda ancha, la estrategia de histéresis se aplica en el caso en el que la trama está en un "período de mantenimiento " cuya longitud varía según la SNR a largo plazo de la manera siguiente:

$$I_{mant} = 0 \quad \text{si } SNR_{LT} \geq 35$$

$$I_{mant} = 1 \quad \text{si } 15 \leq SNR_{LT} < 35 .$$

$$I_{mant} = 2 \quad \text{si } SNR_{LT} < 15 .$$

El periodo de mantenimiento comienza en la primera trama de sonido inactiva después de tres (3) tramas de sonido activas consecutivas. Su función consiste en forzar cada trama inactiva durante el periodo de mantenimiento como una trama activa. La decisión SAD se explicará más adelante.

Para las señales de banda estrecha, la estrategia de histéresis consiste en disminuir el umbral de decisión SAD de la manera siguiente:

$$umbral_{SAD} = umbral_{SAD} - 5,2 \quad \text{si } SNR_{LT} < 19$$

$$umbral_{SAD} = umbral_{SAD} - 2 \quad \text{si } 19 \leq SNR_{LT} < 35$$

$$umbral_{SAD} = umbral_{SAD} \quad \text{si } 35 \leq SNR_{LT}$$

De esta manera, para señales ruidosas con baja SNR, el umbral se hace más pequeño para dar preferencia a la decisión de señal activa. No hay mantenimiento para las señales de banda estrecha.

Finalmente, el detector 103 de actividad sonora tiene dos salidas - un indicador SAD y un indicador SAD local. Ambos indicadores se establecen a uno si se detecta una señal activa y se establecen a cero en caso contrario. Además, el indicador SAD se establece a uno en el período de mantenimiento. La decisión SAD se realiza comparando la SNR media por cada trama con el umbral de decisión SAD (a través de un comparador, por ejemplo), es decir:

```

If SNRav > umbralSAD
    SADlocal = 1
    SAD = 1
else
    SADlocal = 1
If en periodo de mantenimiento
    SAD = 1
else
    SAD = 0
end
end.

```

Primer nivel de estimación y actualización de ruido

Un estimador 104 de ruido según se ilustra en la Figura 1 calcula la energía total de ruido, la energía relativa de la trama, actualiza la energía media del ruido a largo plazo y la energía media de trama a largo plazo, la energía media por cada banda crítica, y un factor de corrección de ruido. Además, el estimador 104 de ruido realiza una inicialización de energía de ruido y realiza una actualización a la baja.

La energía total de ruido por cada trama se calcula usando la relación siguiente:

$$N_{tot} = 10 \log \left(\sum_{i=0}^{19} N_{CB}(i) \right) \quad (11)$$

en la que $N_{CB}(i)$ es la energía de ruido estimada por cada banda crítica.

La energía relativa de la trama viene dada por la diferencia entre la energía de trama en dB y la energía media a largo plazo. La energía relativa de la trama se calcula usando la relación siguiente:

$$E_{rel} = E_t - \bar{E}_f \quad (12)$$

en la que E_t viene dada por la ecuación (6).

La energía media de ruido a largo plazo o energía media de trama a largo plazo es actualizada en cada trama. En caso de tramas de señal activa (indicador SAD = 1), la energía de trama media a largo plazo es actualizada usando la relación:

$$\bar{E}_f = 0,99\bar{E}_f + 0,01E_t \quad (13)$$

con valor inicial $\bar{E}_f = -45$ dB.

En caso de tramas de voz inactiva (indicador SAD = 0), la energía media de ruido a largo plazo es actualizada como se indica a continuación:

$$\bar{N}_f = 0,99\bar{N}_f + 0,01N_{tot} \quad (14)$$

El valor inicial de $\bar{N}_f$ se establece igual a N_{tot} para las primeras 4 tramas. Además, en las primeras cuatro (4) tramas, el valor de $\bar{E}_f$ está limitado por $\bar{E}_f \geq \bar{N}_{tot} + 10$.

La energía de trama por cada banda crítica para toda la trama se calcula promediando las energías de ambos análisis espectrales primero y segundo en la trama usando la relación siguiente:

$$\bar{E}_{CB}(i) = 0,5E_{CB}^{(1)}(i) + 0,5E_{CB}^{(2)}(i) \quad (15)$$

La energía de ruido por cada banda crítica $N_{CB}(i)$ se inicializa a 0,03.

En esta etapa, sólo se realiza una actualización a la baja de la energía de ruido para las bandas críticas de manera que la energía es menor que la energía de ruido de fondo. En primer lugar, se calcula la energía de ruido actualizada temporal usando la relación siguiente:

$$N_{imp}(i) = 0,9N_{CB}(i) + 0,1(0,25E_{CB}^{(0)}(i) + 0,75\bar{E}_{CB}(i)) \quad (18)$$

en la que $E_{CB}^{(0)}(i)$ denota la energía por cada banda crítica correspondiente al segundo análisis espectral a partir de

la trama previa.

A continuación, para $i = 0$ a 19, si $N_{tmp}(i) < N_{CB}(i)$, entonces $N_{CB}(i) = N_{tmp}(i)$.

- 5 Más tarde, se realiza un segundo nivel de la estimación y actualización de ruido estableciendo de $N_{CB}(i) = N_{tmp}(i)$ si la trama es declarada como una trama inactiva.

Segundo nivel de estimación y actualización de ruido

- 10 El módulo 107 paramétrico de detección de actividad sonora y de actualización de estimación de ruido actualiza las estimaciones de energía de ruido por cada banda crítica a ser usadas en el detector 103 de actividad sonora en la siguiente trama. La actualización se realiza durante los periodos de señal inactivos. Sin embargo, la decisión SAD realizada anteriormente, basada en la SNR por cada banda crítica, no es usada para determinar si se actualizan o no las estimaciones de energía de ruido. Se realiza otra decisión en base a otros parámetros independientes en lugar de la SNR por cada banda crítica. Los parámetros usados para la actualización de las
- 15 estimaciones de energía de ruido son: estabilidad de tono, no estacionariedad de la señal, sonoridad y la relación entre las energías de error residual LP de 2° orden y 16° y generalmente tienen una baja sensibilidad a las variaciones de nivel de ruido. La decisión para la actualización de las estimaciones de energía de ruido es optimizada para señales de voz. Para mejorar la detección de las señales musicales activa, se usan los parámetros siguientes diferentes: diversidad espectral, no estacionariedad complementaria, carácter ruido y
- 20 estabilidad tonal. La detección de música se explicará en detalle en la descripción siguiente.

- La razón para no usar la decisión SAD para la actualización de las estimaciones de energía de ruido es hacer la estimación de ruido robusta a niveles de ruido que cambian rápidamente. Si se usó la decisión SAD para la actualización de las estimaciones de energía de ruido, un aumento repentino en el nivel de ruido causaría un
- 25 aumento de la SNR incluso para tramas de señal inactivas, impidiendo la actualización de las estimaciones de energía de ruido, lo que a su vez mantendría la SNR alta en las siguientes tramas, y así sucesivamente. Por consiguiente, la actualización sería bloqueada y se necesitaría alguna otra lógica para reanudar la adaptación de ruido.

- 30 En la realización ilustrativa no restrictiva de la presente invención, se realiza un análisis de tono en bucle abierto en un módulo 106 analizador LP y rastreador de tono en la Figura 1) para calcular tres estimaciones de tono en bucle abierto por cada trama: d_0 , d_1 y d_2 correspondientes a la primera mitad de la trama, segunda mitad de la trama y la anticipación, respectivamente. Este procedimiento es bien conocido por las personas con conocimientos ordinarios en la materia y no se describirá adicionalmente en la presente descripción (por ejemplo, VMR-WB [Source-
- 35 Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB), Service Options 62 and 63 for Spread Spectrum Systems, 3GPP2 Technical Specification C.S0052-A v1.0, Abril de 2005 (<http://www.3gpp2.org>)). El módulo 106 analizador LP y rastreador de tono calcula un contador de estabilidad de tono usando la relación siguiente:

$$40 \quad pc = |d_0 - d_{-1}| + |d_1 - d_0| + |d_2 - d_1| \quad (19)$$

- en la que d_{-1} es el retardo de la segunda mitad de la trama de la trama previa. Para retardos de tono mayores que 122, el módulo 105 analizador LP y rastreador de tono establece $d_2 = d_1$. De esta manera, para dichos retardos, el
- 45 valor de pc en la ecuación (19) es multiplicado por 3/2 para compensar el tercer término faltante en la ecuación. La estabilidad de tono es verdadera si el valor de PC es menor de 14. Además, para las tramas con baja sonoridad, pc se establece a 14 para indicar inestabilidad de tono. Más específicamente:

$$50 \quad \text{Si } (C_{norm}(d_0) + C_{norm}(d_1) + C_{norm}(d_2)) / 3 + r_e < th_{Cpc} \text{ entonces } pc = 14, \quad (20)$$

- en la que $C_{norm}(d)$ es la correlación bruta normalizada y r_e es una corrección opcional añadida a la correlación normalizada para compensar la reducción de la correlación normalizada en presencia de ruido de fondo. El umbral de sonoridad $umbral_{Cpc} = 0,52$ para WB y $umbral_{Cpc} = 0,65$ para NB. El factor de corrección puede ser calculado usando la relación siguiente:
- 55

$$60 \quad r_e = 0,00024492 e^{0,1596(N_{snr}-14)} - 0,022$$

en la que N_{tot} es la energía de ruido total por cada trama calculada según la ecuación (11).

La correlación bruta normalizada puede ser calculada en base a la señal $s_{wd}(n)$ acústica ponderada diezmada usando la ecuación siguiente:

$$C_{norm}(d) = \frac{\sum_{n=0}^{L_{sec}} s_{wd}(t_{inicio}) s_{wd}(t_{inicio} - d)}{\sqrt{\sum_{n=0}^{L_{sec}} s_{wd}^2(t_{inicio}) \sum_{n=0}^{L_{sec}} s_{wd}^2(t_{inicio} - d)}},$$

en la que el límite de los sumatorios depende del propio retardo. La señal $s_{wd}(n)$ ponderada es la usada en el análisis de tono en bucle abierto y proporcionada filtrando la señal de sonido de entrada pre-procesada desde el pre-procesador 101 a través de un filtro de ponderación de la forma $A(z/\gamma)/(1-\mu z^{-1})$. La señal $s_{wd}(n)$ ponderada es diezmada en un factor de 2 y los límites de los sumatorios se proporcionan en función de:

$$\begin{aligned} L_{sec} &= 40 \text{ para } d = 10, \dots, 16 \\ L_{sec} &= 40 \text{ para } d = 10, \dots, 16 \\ L_{sec} &= 62 \text{ para } d = 32, \dots, 61 \\ L_{sec} &= 115 \text{ para } d = 62, \dots, 115 \end{aligned}$$

Estas longitudes aseguran que la longitud del vector correlacionado comprende al menos un periodo de tono que ayuda a obtener una detección de tono en bucle abierto robusta. Los instantes t_{inicio} están relacionados con el comienzo de la trama actual y vienen dados por:

$$\begin{aligned} t_{inicio} &= 0 \text{ para la primera mitad de la trama} \\ t_{inicio} &= 138 \text{ para la segunda mitad de la trama} \\ t_{inicio} &= 256 \text{ para la anticipación} \end{aligned}$$

a una frecuencia de muestreo de 12,8 kHz.

El módulo 107 paramétrico de detección de actividad sonora y de actualización de estimación de ruido realiza una estimación de no estacionariedad de la señal en base al producto de las relaciones entre la energía por cada banda crítica y la energía media a largo plazo por cada banda crítica.

La energía media a largo plazo por cada banda crítica es actualizada usando la relación siguiente:

$$E_{CB,LT}(i) = \alpha_e E_{CB,LT}(i) + (1 - \alpha_e) \bar{E}_{CB}(i), \quad \text{para } i = b_{min} \text{ to } b_{max}, \quad (21)$$

en la que $b_{min} = 0$ y $b_{max} = 19$ en el caso de señales de banda ancha, y $b_{min} = 1$ y $b_{max} = 16$ en el caso de señales de banda estrecha, y $\bar{E}_{CB}(i)$ es la energía de trama por cada banda crítica definida en la ecuación (15). El factor α_e de actualización es una función lineal de la energía de trama total, definida en la Ecuación (6), y se calcula como sigue:

Para señales de banda ancha: $\alpha_e = 0,024 E_t - 0,235$ con $0,5 \leq \alpha_e \leq 0,99$.

Para señales de banda estrecha: $\alpha_e = 0,00091 E_t + 0,3185$ con $0,5 \leq \alpha_e \leq 0,999$.

E_t se calcula usando la ecuación (6).

La no estacionariedad de la trama viene dada por el producto de las relaciones entre la energía de trama y la energía media a largo plazo por cada banda crítica. Más específicamente:

$$no_estac = \prod_{i=b_{min}}^{b_{max}} \frac{\max(\bar{E}_{CB}(i), E_{CB,LT}(i))}{\min(\bar{E}_{CB}(i), E_{CB,LT}(i))} \quad (22)$$

El módulo 107 paramétrico de detección de actividad sonora y de actualización de estimación de ruido produce además un factor de sonoridad para la actualización de ruido usando la relación siguiente:

$$sonoridad = (C_{norm}(d_0) + C_{norm}(d_1)) / 2 + r_e \quad (23)$$

Finalmente, el módulo 107 paramétrico de detección de actividad sonora y de actualización de estimación de ruido calcula una relación entre la energía residual LP después del análisis LP de 2º orden y 16º orden usando la relación:

$$relación_resid = E(2) / E(16) \quad (24)$$

en la que E(2) y E(16) son las energías residuales LP después de un análisis LP de 2º orden y 16º orden calculado en el módulo 106 analizador LP y rastreador de tono usando recursión de Levinson-Durbin, que es un procedimiento bien conocido por las personas con conocimientos ordinarios en la materia. Esta relación refleja el hecho de que para representar una envolvente espectral de la señal, generalmente se necesita un orden superior de LP para una señal de voz que para el ruido. En otras palabras, se supone que la diferencia entre E(2) y E(16) es inferior para el ruido que para la voz activa.

La decisión de actualización realizada por el módulo 107 paramétrico de detección de actividad sonora y de actualización de estimación de ruido se determina en base a una variable *actualiza_ruido* que se establece inicialmente a 6 y se decrementa en 1 si se detecta una trama inactiva y se incrementa en 2 si se detecta una trama activa. Además, la variable *actualiza_ruido* está limitada entre 0 y 6. Las estimaciones de energía de ruido sólo se actualizan cuando *actualiza_ruido* = 0.

El valor de la variable *actualiza_ruido* es actualizado en cada trama de la manera siguiente:

```

If (noestac > umbralestac) OR (pc < 14) OR (sonoridad > umbralCnorm) OR (relación_resid > umbralresid)
    actualiza_ruido = actualiza_ruido + 2
else
    actualiza_ruido = actualiza_ruido - 2

```

donde para señales de banda ancha, $umbral_{estac} = umbral_{Cnorm} = 0,85$ y $umbral_{resid} = 1,6$, y para señales de banda estrecha, $umbral_{estac} = 500.000$, $umbral_{Cnorm} = 0,7$ y $umbral_{resid} = 10,4$.

En otras palabras, las tramas se declaran inactivas para la actualización de ruido cuando

(noestac ≤ umbral_{estac}) AND (pc ≥ 14) AND (sonoridad ≤ umbral_{Cnorm}) AND (relación_resid ≤ umbral_{resid})

y se usa un mantenimiento de 6 tramas antes de realizar la actualización de ruido.

De esta manera, si *actualiza_ruido* = 0 entonces para $i = 0$ a 19 $N_{CB}(i) = N_{tmp}(i)$ donde $N_{tmp}(i)$ es la energía de ruido actualizada temporal ya calculada en la Ecuación (18).

Mejora de la detección de ruido por las señales musicales

La estimación del ruido descrita anteriormente tiene sus limitaciones para ciertas señales musicales, tales como conciertos de piano o rock y pop instrumental, ya que se desarrolló y optimizó principalmente para detección del habla. Para mejorar la detección de señales musicales en general, el módulo 107 paramétrico de detección de actividad sonora y de actualización de estimación de ruido usa otros parámetros o técnicas en conjunción con las existentes. Estos otros parámetros o técnicas comprenden, tal como se ha descrito anteriormente en la presente memoria, la diversidad espectral, la no estacionariedad complementaria, el carácter del ruido y la estabilidad tonal, calculados por un calculador de diversidad espectral, un calculador de no estacionariedad complementaria, un calculador de carácter de ruido y un estimador de tonalidad, respectivamente. Se describirán en detalle en la presente memoria, a continuación.

Diversidad espectral

La diversidad espectral proporciona información acerca de los cambios significativos de la señal en el dominio de la frecuencia. Los cambios son rastreados en las bandas críticas comparando las energías en el primer análisis espectral de la trama actual y el segundo análisis espectral hace dos tramas. La energía en una banda i crítica del primer análisis espectral en la trama actual se denota como $E_{CB}^{(1)}(i)$. Denótese la energía en la misma banda crítica calculada en el segundo análisis espectral hace dos tramas como $E_{CB}^{(-2)}(i)$. Estas dos energías se inicializan a 0,0001. A continuación, para todas las bandas críticas superiores a 9, el máximo y el mínimo de las dos energías se calculan como se indica a continuación:

$$\begin{aligned} E_{\max}(i) &= \max \{ E_{CB}^{(1)}(i), E_{CB}^{(-2)}(i) \} \\ E_{\min}(i) &= \min \{ E_{CB}^{(1)}(i), E_{CB}^{(-2)}(i) \} \end{aligned}, \quad \text{para } i=10, \dots, b_{\max}.$$

Posteriormente, se calcula una relación entre la energía máxima y mínima en una banda crítica específica como

$$E_{rat}(i) = \frac{E_{\max}(i)}{E_{\min}(i)}, \quad \text{para } i=10, \dots, b_{\max}.$$

Finalmente, el módulo 107 paramétrico de detección de actividad sonora y de actualización de estimación de ruido calcula un parámetro de diversidad espectral como una suma ponderada normalizada de las relaciones en la que el propio peso es la energía máxima $E_{\max}(i)$. Este parámetro de diversidad espectral viene determinado por la relación siguiente:

$$div_espec = \frac{\sum_{i=10}^{b_{\max}} E_{\max}(i) E_{rat}(i)}{\sum_{i=10}^{b_{\max}} E_{\max}(i)} \quad (25)$$

El parámetro div_espec se usa en la decisión final acerca de la actividad musical y la actualización de energía de ruido. El parámetro div_espec se usa también como un parámetro auxiliar para el cálculo de un parámetro no estacionariedad complementaria que se describe más adelante.

No estacionariedad complementaria

La inclusión de un parámetro no estacionariedad complementaria está motivada por el hecho de que el parámetro no estacionariedad, definido en la ecuación (22), falla cuando un ataque agudo de energía en una señal musical es seguido por una disminución lenta de la energía. En este caso, la energía media a largo plazo por cada banda crítica, $E_{CB,LT}(i)$, definida en la Ecuación (21), aumenta lentamente durante el ataque, mientras que la energía de la trama por banda crítica, definida en la Ecuación (15), disminuye lentamente. En una cierta trama después del ataque, estos dos valores de energía se encuentran y el parámetro $noestac$ resulta en un valor pequeño que indica una ausencia de señal activa. Esto conduce a una actualización de ruido falsa y, posteriormente, a una decisión SAD falsa.

Para superar este problema, se calcula una energía media a largo plazo alternativa por cada banda crítica usando la relación siguiente:

$$E2_{CB,LT}(i) = \beta_e E2_{CB,LT}(i) + (1 - \beta_e) \bar{E}_{CB}(i), \quad \text{para } i=b_{min} \text{ to } b_{max}. \quad (26)$$

La variable $E2_{CB,LT}(i)$ es inicializada a 0,03 para todo i . La Ecuación (26) se asemeja mucho a la Ecuación (21) siendo la única diferencia el factor β_e de actualización que se calcula como sigue:

```
if (div_espec > umbraldiv_espec)
     $\beta_e = 0$ 
else
     $\beta_e = \alpha_e$ 
end,
```

en la que $umbral_{div_espec} = 5$. De esta manera, cuando se detecta un ataque de energía ($div_espec > 5$) la energía media alternativa a largo plazo es establecida inmediatamente a la energía media de trama, es decir, $E2_{CB,LT}(i) = \bar{E}_{CB}(i)$. De lo contrario, esta energía media alternativa a largo plazo es actualizada en la misma forma que la no estacionariedad convencional, es decir, usando el filtro exponencial con el factor α_e de actualización. El parámetro no estacionariedad complementaria se calcula de la misma manera que $noestac$, pero usando $E2_{CB,LT}(i)$, es decir,

$$no_estac2 = \prod_{i=b_{min}}^{b_{max}} \frac{\max(\bar{E}_{CB}(i), E2_{CB,LT}(i))}{\min(\bar{E}_{CB}(i), E2_{CB,LT}(i))} \quad (27)$$

El parámetro no estacionariedad complementaria, $noestac2$, puede fallar unos pocos tramas justo después de un ataque de energía, pero no debería fallar durante los pasos caracterizados por una energía lentamente decreciente. Debido a que el parámetro $noestac$ funciona bien en ataques de energía y unas pocas tramas después, por lo tanto, una disyunción lógica de $noestac$ y $noestac2$ resuelve el problema de la detección de señal inactiva en ciertas señales musicales. Sin embargo, la disyunción se aplica sólo en pasajes que es "probable que sean activos". La probabilidad se calcula como sigue:

```

    If (noestac > umbralestac) OR (estabilidad_tonal = 1))
        pred_act_LT = ka pred_act_LT + (1 - ka) . 1
    else
        pred_act_LT = ka pred_act_LT + (1 - ka) . 0
    end.

```

El coeficiente ka se establece a 0,99. El parámetro $pred_act_LT$ que está comprendido en el intervalo $<0:1>$ puede ser interpretado como un predictor de actividad. Cuando es cercano a 1, es probable que la señal sea activa, y cuando es cercano a 0, es probable que sea inactiva. El parámetro $pred_act_LT$ se inicializa a uno. En la condición anterior, $estabilidad_tonal$ es un parámetro binario que se usa para detectar una señal tonal estable. Este parámetro $estabilidad_tonal$ se describirá en la descripción siguiente.

El parámetro $noestac2$ se toma en consideración (en disyunción con $noestac$) en la actualización de la energía de ruido sólo si $pred_act_LT$ es mayor que cierto umbral, el cual ha sido establecido a 0,8. La lógica de la actualización de energía de ruido se explica en detalle al final de la presente sección.

Carácter de ruido

Carácter de ruido es otro parámetro que se usa en la detección de ciertas señales musicales similares al ruido, tales como los platillos o los tambores de baja frecuencia. Este parámetro se calcula usando la relación siguiente:

$$caract_ruido = \frac{\sum_{i=10}^{b_{max}} E_{CB}(i)}{\sum_{i=b_{min}}^9 E_{CB}(i)} \quad (28)$$

El parámetro $caract_ruido$ se calcula sólo para las tramas cuyo contenido espectral tiene por lo menos una energía mínima, lo cual se cumple cuando el numerador y el denominador de la Ecuación (28) son mayores de 100. El parámetro $caract_ruido$ tiene un límite superior de 10 y su valor a largo plazo se actualiza usando la relación siguiente:

$$caract_ruido_LT = \alpha_n caract_ruido_LT + (1 - \alpha_n) caract_ruido \quad (29)$$

El valor inicial de $caract_ruido_LT$ es 0 y α_n se establece a un valor de 0,9. Este parámetro $caract_ruido_LT$ se usa en la decisión acerca de la actualización de energía de ruido que se explica al final de la presente sección.

Estabilidad tonal

Estabilidad tonal es el último parámetro usado para prevenir una actualización falsa de las estimaciones de energía de ruido. Estabilidad tonal se usa también para evitar declarar algunos segmentos de música como tramas sordas. Estabilidad tonal se usa además en un códec de banda súper ancha incorporado para decidir qué modelo de codificación se usará para codificar la señal de sonido por encima de 7 kHz. La detección de la estabilidad tonal hace uso de la naturaleza tonal de las señales musicales. En una señal musical típica hay tonos que son estables durante varias tramas consecutivas. Para hacer uso de esta característica, es necesario realizar un seguimiento de las posiciones y las formas de los fuertes picos espectrales, ya que estos pueden corresponder a los tonos. La detección de la estabilidad tonal se basa en un análisis de correlación entre los picos espectrales en la trama actual y los de la trama pasada. La entrada es el espectro logarítmico de energía media definido en la Ecuación (4). El número de contenedores espectrales se denota como N_{ESPEC} (el contenedor 0 es el componente CC y $N_{ESPEC} = L_{FFT}/2$). En la descripción siguiente, el término "espectro" se referirá al espectro logarítmico de energía media, definido por la Ecuación (4).

La detección de estabilidad tonal se realiza en tres etapas. Además, la detección de estabilidad tonal usa un calculador de un espectro residual actual, un detector de picos en el espectro residual actual y un calculador de un mapa de correlación y un mapa de correlación a largo plazo, que se describirán más adelante en la presente memoria.

En la primera etapa, se buscan los índices de los mínimos locales del espectro (por parte de un localizador de mínimos de espectro por ejemplo), en un bucle descrito por la fórmula siguiente y se almacenan en una memoria intermedia i_{min} que puede ser expresada como sigue:

$$i_{min} = (\forall i : (E_{dB}(i-1) > E_{dB}(i)) \wedge (E_{dB}(i) < E_{dB}(i+1))) \quad i = 1, \dots, N_{ESPEC} - 2 \quad (30)$$

en la que el símbolo $\wedge$ significa AND lógico.

En la Ecuación (30), $E_{dB}(i)$ denota el espectro logarítmico de energía media calculado mediante la Ecuación (4). El primer índice en i_{min} es 0, si $E_{dB}(0) < E_{dB}(1)$. Por consiguiente, el último índice en i_{min} es $N_{ESPEC}-1$, si $E_{dB}(N_{ESPEC}-1) < E_{dB}(N_{ESPEC}-2)$. Denótese el número de mínimos encontrados como N_{min} .

La segunda etapa consiste en calcular un suelo espectral (mediante un estimador de suelo espectral, por ejemplo) y restarlo del espectro (mediante un sustractor adecuado, por ejemplo). El suelo espectral es una función lineal por tramos que se extiende a través de los mínimos locales detectados. Cada tramo lineal entre dos mínimos $i_{min}(x)$ e $i_{min}(x+1)$ consecutivo puede describirse como:

$$fl(j) = k.(j - i_{min}(x)) + q \quad j = i_{min}(x), \dots, i_{min}(x+1),$$

en la que k es la pendiente de la línea y $q = E_{dB}(i_{min}(x))$. La pendiente k puede ser calculada usando la relación siguiente:

$$k = \frac{E_{dB}(i_{min}(x+1)) - E_{dB}(i_{min}(x))}{i_{min}(x+1) - i_{min}(x)}.$$

De esta manera, el suelo espectral es una conexión lógica de todos los tramos:

$$\begin{aligned} \text{suelo_espect}(j) &= E_{dB}(j) & j &= 0, \dots, i_{min}(0) - 1 \\ \text{suelo_espect}(j) &= fl(j) & j &= i_{min}(0), \dots, i_{min}(N_{min} - 1) - 1. \\ \text{suelo_espect}(j) &= E_{dB}(j) & j &= i_{min}(N_{min} - 1), \dots, N_{ESPEC} - 1 \end{aligned} \quad (31)$$

Los principales contenedores hasta $i_{min}(0)$ y los contenedores de terminación desde $i_{min}(N_{min}-1)$ del suelo espectral se establecen al propio espectro. Finalmente, el suelo espectral se resta del espectro usando la relación siguiente:

$$E_{dB,res}(j) = E_{dB}(j) - suelo_espect(j) \quad j = 0, \dots, N_{ESPEC} - 1 \quad (32)$$

y el resultado se denomina espectro residual. El cálculo del suelo espectral se ilustra en la Figura 3.

En la tercera etapa, se calculan un mapa de correlación y un mapa de correlación a largo plazo a partir del espectro residual de la trama actual y la trama anterior. Una vez más, esta es una operación por tramos. De esta manera, el mapa de correlación se calcula pico a pico ya que los mínimos delimitan los picos. En la descripción siguiente, se usará el término "pico" para denotar un tramo entre dos mínimos en el espectro residual $E_{dB,res}$.

Denótese el espectro residual de la trama anterior como $E_{dB,res}^{(-1)}(j)$. Para cada pico en el espectro residual actual se calcula una correlación normalizada en el que la forma en el espectro residual anterior corresponde a la posición de este pico. Si la señal era estable, los picos no deberían moverse considerablemente de trama a trama y sus posiciones y formas deberían ser aproximadamente la misma. De esta manera, la operación de correlación tiene en cuenta todos los índices (contenedores) de un pico específico, que está delimitada por dos mínimos consecutivos. Más específicamente, la correlación normalizada se calcula usando la relación siguiente:

$$map_cor(i_{min}(x) : i_{min}(x+1)) = \frac{\left(\sum_{j=i_{min}(x)}^{i_{min}(x+1)-1} E_{dB,res}(j) E_{dB,res}^{(-1)}(j) \right)^2}{\sum_{j=i_{min}(x)}^{i_{min}(x+1)-1} (E_{dB,res}(j))^2 \sum_{j=i_{min}(x)}^{i_{min}(x+1)} (E_{dB,res}^{(-1)}(j))^2},$$

$$x = 0, \dots, N_{min} - 2 \quad (33)$$

Los principales contenedores de map_cor hasta $i_{min}(0)$ y los contenedores de terminación map_cor desde $i_{min}(N_{min}-1)$ se establecen a cero. El mapa de correlación se muestra en la Figura 4.

El mapa de correlación de la trama actual se usa para actualizar su valor a largo plazo que viene descrito por:

$$map_cor_LT(k) = \alpha_{map} map_cor_LT(k) + (1 - \alpha_{map}) map_cor(k),$$

$$k = 0, \dots, N_{ESPEC} - 1, \quad (34)$$

en la que $\alpha_{map} = 0,9$. El map_cor_LT se inicializa a cero para todo k .

Finalmente, todos los valores de map_cor_LT se suman entre sí (mediante un sumador, por ejemplo) como sigue:

$$sum_map_cor = \sum_{j=0}^{N_{ESPEC}-1} map_cor_LT(j). \quad (35)$$

Si algún valor de $map_cor_LT(j)$, $j = 0, \dots, N_{ESPEC}-1$, supera un umbral de 0,95, se establece un indicador cor_fuerte (que puede ser considerado como un detector) a uno, de lo contrario, se establece a cero.

La decisión acerca de la estabilidad tonal se calcula sometiendo sum_map_cor a un umbral adaptativo, $umbral_tonal$. Este umbral se inicializa a 56 y se actualiza cada trama de la manera siguiente:

```

        if (sum_map_cor > 56)
            umbral_tonal = umbral_tonal - 0,2
        else
5         umbral_tonal = umbral_tonal + 0,2
        end.

```

10 El umbral *umbral_tonal* adaptativo tiene un límite superior de 60 y un límite inferior de 49. De esta manera, el umbral *umbral_tonal* adaptativo disminuye cuando la correlación es relativamente buena, indicando un segmento de señal activa, y si no aumenta. Cuando el umbral es más bajo, es más probable que más tramas sean clasificadas como activas, especialmente al final de los períodos activos. Por lo tanto, el umbral adaptativo puede considerarse como un mantenimiento.

15 El parámetro *estabilidad_tonal* se establece a uno cada vez que *sum_map_cor* es mayor que *umbral_tonal* o cuando el indicador *cor_fuerte* se establece a uno. Más específicamente:

```

        If ((sum_map_cor > umbral_tonal) OR (cor_fuerte = 1))
            estabilidad_tonal = 1
        else
20         estabilidad_tonal = 0
        end.

```

Uso de los parámetros de detección de música en la actualización de energía de ruido

25 Todos los parámetros de detección de música se incorporan en la decisión definitiva realizada en el módulo 107 paramétrico de detección de actividad sonora y actualización de estimación de ruido (Act) acerca de la actualización de las estimaciones de energía de ruido. Las estimaciones de energía de ruido se actualizan siempre que el valor *actualiza_ruido* es igual a cero. Inicialmente, se establece a 6 y se actualiza cada trama de la manera siguiente:

```

30 if (noestac > umbral_estac) OR (pc < 14) OR (sonoridad > umbralCnorm) OR (relación_resid > umbral_resid) OR
    (estabilidad_tonal = 1) OR (car_ruido_LT > 0,3) OR ((pred_act_LT > 0,8) AND (noestac2 > umbral_estac))
        actualiza_ruido = actualiza_ruido + 2
    else
35         actualiza_ruido = actualiza_ruido - 1
    end.

```

40 Si la condición combinada tiene un resultado positivo, la señal es activa y el parámetro *actualiza_ruido* se incrementa. De lo contrario, la señal es inactiva y el parámetro se decrementa. Cuando llega a 0, la energía de ruido es actualizada con la energía de la señal actual.

Además de en la actualización de energía de ruido, el parámetro *estabilidad_tonal* se usa también en el algoritmo de clasificación de señal de sonido sorda. Específicamente, el parámetro se usa para mejorar la robustez de la clasificación de señal sorda sobre música, tal como se describirá en la siguiente sección.

45 *Clasificación de señal de sonido (clasificador 108 de señal de sonido)*

La filosofía general subyacente en el clasificador 108 de señal de sonido (Figura 1) se representa en la Figura 5. El enfoque puede ser descrito como sigue. La clasificación de señal de sonido se realiza en tres etapas en los módulos 501, 502 y 503 lógicos, cada uno de los cuales discrimina una clase de señal específica. En primer lugar, un detector 501 de actividad de señal (SAD) discrimina entre tramas de señal activas e inactivas. Este detector 501 de actividad de señal es el mismo que el del detector 103 de actividad de señal en la Figura 1. El detector de actividad de señal ya ha sido descrito en la descripción anterior.

55 Si el detector 501 de actividad de señal detecta una trama inactiva (señal de ruido de fondo), entonces la cadena de clasificación termina y, si hay compatibilidad con transmisión discontinua (Discontinuous Transmission, DTX), un módulo 541 de codificación que puede ser incorporado en el codificador 109 (Figura 1) codifica la trama con generación de ruido de confort (CNG). Si no hay compatibilidad con DTX, la trama continúa en la clasificación de señal activa, y más frecuentemente es clasificada como trama de voz sorda.

60 Si una trama de señal activa es detectada por el detector 501 de actividad sonora, la trama es sometida a un segundo clasificador 502 dedicado a discriminar tramas de voz sordas. Si el clasificador 502 clasifica la trama

como señal de voz sorda, la cadena de clasificación termina, un módulo 542 de codificación que puede ser incorporado en el codificador 109 (Figura 1) codifica la trama con un procedimiento de codificación optimizado para señales de voz sordas.

5 De lo contrario, la trama de señal es procesada a través de un clasificador 503 de "voz sonora estable". Si la trama es clasificada como una trama sonora estable por el clasificador 503, a continuación, un módulo 543 de codificación que puede ser incorporado en el codificador 109 (Figura 1) codifica la trama usando un procedimiento de codificación optimizado para señales sonoras estables o cuasi periódicas.

10 De lo contrario, es probable que la trama contenga un segmento de señal no estacionario, tal como un inicio de voz sonora o voz sonora que evoluciona rápidamente o una señal musical. Típicamente, estas tramas requieren un módulo 544 de codificación de propósito general que puede ser incorporado en el codificador 109 (Figura 1) para codificar la trama a una alta tasa de bits para mantener una buena calidad subjetiva.

15 A continuación, se describirá la clasificación de tramas de señal sordas y sonoras. El detector 501 SAD (o 103 en la Figura 1) usado para discriminar las tramas inactivas ha sido ya descrito en la descripción anterior.

20 Las partes sordas de la señal de voz se caracterizan por carecer del componente periódico y pueden ser divididas en tramas inestables, en las que la energía y el espectro cambian rápidamente, y tramas estables, en las que estas características se mantienen relativamente estables. La realización ilustrativa no restrictiva de la presente invención propone un procedimiento para la clasificación de las tramas sordas usando los parámetros siguientes:

- medida de sonoridad, calculada como una correlación normalizada media ($\bar{r}_x$);
- medida de inclinación espectral media ($\bar{e}$);
- 25 • aumento máximo de energía en tiempo corto desde nivel bajo ($dE0$) diseñado para detectar de manera eficiente las oclusivas de voz en una señal;
- la estabilidad tonal para discriminar la música de una señal sorda (descrita en la descripción anterior); y
- energía relativa de trama (E_{rel}) para detectar las señales de muy baja energía.

30 *Medida de sonoridad*

La correlación normalizada, usada para determinar la medida de sonoridad, se calcula como parte del análisis de tono en bucle abierto realizado en el módulo 106 analizador LP y rastreador de tono de la Figura 1. Pueden usarse, por ejemplo, tramas de 20 ms. El módulo 106 analizador LP y rastreador de tono normalmente emite una estimación de tono en bucle abierto cada 10 ms (dos veces por cada trama). Aquí, el módulo 106 analizador LP y rastreador de tono se usa también para producir y emitir las medidas de correlación normalizadas. Estas correlaciones normalizadas se calculan sobre una señal ponderada y una señal ponderada pasada en el retardo de tono de bucle abierto. La señal $s_w(n)$ de voz ponderada es calculada usando un filtro de ponderación perceptual. Por ejemplo, puede usarse un filtro de ponderación perceptual con denominador fijo, adecuado para señales de banda ancha. Un ejemplo de una función de transferencia para el filtro de ponderación perceptual viene determinado por la relación siguiente:

$$W(z) = \frac{A(z/\gamma_1)}{1 - \gamma_2 z^{-1}},$$

45 en la que $0 < \gamma_2 < \gamma_1 \leq 1$

en la que $A(z)$ es la función de transferencia de un filtro de predicción lineal (LP) calculada en el módulo 106 analizador LP y rastreador de tono, que viene determinada por la relación siguiente:

$$50 \quad A(z) = 1 + \sum_{i=1}^p a_i z^{-i}.$$

Los detalles del análisis LP y el análisis de tono en bucle abierto no se describirán adicionalmente en la presente memoria, ya que se cree que son bien conocidos por las personas con conocimientos ordinarios en la materia.

55 La medida de sonoridad viene dada por la correlación media $\bar{C}_{norm}$ que se define como:

$$\bar{C}_{norm} = \frac{1}{3}(C_{norm}(d_0) + C_{norm}(d_1) + C_{norm}(d_2)) + r_e \quad (36)$$

en la que $C_{norm}(d_0)$, $C_{norm}(d_1)$ y $C_{norm}(d_2)$ son, respectivamente, la correlación normalizada de la primera mitad de la trama actual, la correlación normalizada de la segunda mitad de la trama actual, y la correlación normalizada de la anticipación (el principio de la trama siguiente). Los argumentos a las correlaciones son los retardos de tono en bucle abierto indicados anteriormente calculados en el módulo 106 analizador LP y rastreador de tono de la Figura 1. Por ejemplo, puede usarse una anticipación de 10 ms. Se añade un factor r_e de corrección a la correlación media con el fin de compensar el ruido de fondo (en presencia de ruido de fondo el valor de correlación disminuye). El factor de corrección es calculado usando la relación siguiente:

$$r_e = 0,00024492 e^{0,1596(N_{tot}-14)} - 0,022 \quad (37)$$

en la que N_{tot} es la energía de ruido total por cada trama calculada según la Ecuación (11).

15 *Inclinación espectral*

El parámetro inclinación espectral contiene información acerca de la distribución de frecuencias de la energía. La inclinación espectral puede ser estimada en el dominio de la frecuencia como una relación entre la energía concentrada en las frecuencias bajas y la energía se concentra en las frecuencias altas. Sin embargo, puede ser estimada también usando otros procedimientos, tales como una relación entre los dos primeros coeficientes de autocorrelación de la señal.

El analizador 102 espectral en la Figura 1 se usa para realizar dos análisis espectrales por cada trama, tal como se ha descrito en la descripción anterior. La energía en las frecuencias altas y en las frecuencias bajas es calculada siguiendo las bandas críticas perceptivas [M. Jelinek y R. Salami, " Noise Reduction Method for Wideband Speech Coding", en Proc. Eusipco, Viena, Austria, Septiembre de 2004], repetido aquí por conveniencia

$$\text{Bandas críticas} = \{100,0, 200,0, 300,0, 400,0, 510,0, 630,0, 770,0, 920,0, \\ 1.080,0, 1.270,0, 1.480,0, 1.720,0, 2.000,0, 2.320,0, 2.700,0, 3.150,0, \\ 3.700,0, 4.400,0, 5.300,0, 6.350,0\} \text{ Hz}$$

La energía en las frecuencias altas es calculada como el promedio de las energías de las dos últimas bandas críticas usando las relaciones siguientes:

$$\bar{E}_h = 0,5 \left[E_{CB}(b_{\max} - 1) + E_{CB}(b_{\max}) \right] \quad (39)$$

en la que las energías de las bandas críticas $E_{CB}(i)$ se calculan según la Ecuación (2). El cálculo se realiza dos veces para ambos análisis espectrales.

La energía en las frecuencias bajas se calcula como el promedio de las energías en las primeras 10 bandas críticas (para NB señales, no se incluye la primera banda), usando la relación siguiente:

$$\bar{E}_l = \frac{1}{10 - b_{\min}} \sum_{i=b_{\min}}^9 E_{CB}(i) \quad (40)$$

Las bandas críticas intermedias han sido excluidas del cálculo para mejorar la discriminación entre las tramas con alta concentración de energía en las frecuencias bajas (generalmente sonoras) y con alta concentración de energía en las frecuencias altas (generalmente sordas). En la parte media, el contenido de energía no es característico de ninguna de las clases y aumenta la confusión de la decisión.

Sin embargo, la energía en las frecuencias bajas se calcula de manera diferente para las señales sordas armónicas con alto contenido de energía en las frecuencias bajas. Esto es debido al hecho de que para los segmentos de voz sonora-femenina, la estructura armónica del espectro puede ser explotada para aumentar la discriminación sonora-sorda. Las señales afectadas son aquellas cuyo período de tono es más corto que 128 o aquellas que no se consideran a priori como sordas. Las señales de sonido consideradas a priori como sordas deben cumplir la condición siguiente:

$$\frac{1}{2}(C_{norm}(d_0) + C_{norm}(d_1)) + r_e < 0,6. \quad (41)$$

De esta manera, para las señales discriminadas por la condición anterior, la energía en las frecuencias bajas se calcula por contenedor sólo se tienen en cuenta en la suma los contenedores cuyas frecuencias son suficientemente cercanas a los armónicos. Más específicamente, se usa la relación siguiente:

$$\bar{E}_l = \frac{1}{cnt} \sum_{i=K_{min}}^{25} E_{BIN}(i) w_h(i). \quad (42)$$

en la que K_{min} es el primer contenedor ($K_{min} = 1$ para WB y $K_{min} = 3$ para NB) y $E_{CONT}(k)$ son las energías de contenedor, tal como se define en la Ecuación (3), en los primeros 25 contenedores de frecuencia (el componente CC se omite). Estos 25 contenedores corresponden a las primeras 10 bandas críticas. En la suma anterior, sólo se consideran los términos cercanos los armónicos del tono; $w_h(i)$ se establece a 1 si la distancia entre los armónicos más cercanos no es mayor que un cierto umbral de frecuencia (por ejemplo, 50 Hz) y se establece en 0 en caso contrario; por lo tanto, sólo papeleras se tienen en cuenta los contenedores alejados menos que 50 Hz de los armónicos más cercanos. El contador cnt es igual al número de términos distintos de cero en la suma. Por lo tanto, si la estructura es armónica en las frecuencias bajas, solo se incluirán términos de energía alta en la suma. Por otra parte, si la estructura no es armónica, la selección de los términos será aleatoria y la suma será menor. De esta manera, pueden detectarse incluso las señales de sonido sordas con alto contenido de energía en las frecuencias bajas.

La inclinación espectral viene dada por la relación siguiente:

$$e_l = \frac{\bar{E}_l - \bar{N}_l}{\bar{E}_h - \bar{N}_h} \quad (43)$$

en la que $\bar{N}_h$ y $\bar{N}_l$ son las energías medias de ruido en las dos (2) últimas bandas críticas y las primeras 10 bandas críticas (o las primeras 9 bandas críticas para NB), respectivamente, calculadas de la misma manera que $\bar{E}_h$ y $\bar{E}_l$ en las Ecuaciones (39) y (40). Las energías de ruido estimadas se han incluido en el cálculo de inclinación para tener en cuenta la presencia de ruido de fondo. Para las señales de NB, las bandas faltantes son compensadas multiplicando e_l por 6. El cálculo de la inclinación espectral se realiza dos veces por cada trama para obtener $e_l(0)$ y $e_l(1)$ correspondientes a ambos análisis espectrales primero y segundo por cada trama. La inclinación espectral media usada en la clasificación de tramas sordas viene dada por

$$\bar{e}_l = \frac{1}{3}(e_{old} + e_l(0) + e_l(1)), \quad (44)$$

en la que $e_{antigua}$ es la inclinación en la segunda mitad de la trama previa.

Aumento máximo de energía de corta duración a bajo nivel

El aumento máximo de energía de corta duración en el nivel bajo dEO se evalúa sobre la señal de sonido $s(n)$, donde $n=0$ corresponde al comienzo de la trama actual. Por ejemplo, se usan tramas de voz de 20 ms y cada trama se divide en 4 subtramas para fines de codificación de voz. La energía de la señal se evalúa dos veces por cada subtrama, es decir, 8 veces por cada trama, en base a segmentos de corta duración de una longitud de 32 muestras (a una tasa de muestreo de 12,8 kHz). Además, también se calculan las energías a corto plazo de las últimas 32 muestras de la trama previa. Las energías de corta duración se calculan usando la relación siguiente:

$$E_{st}^{(l)}(j) = \max_{i=0}^{31} (s^2(i + 32j)), \quad j=-1, \dots, 7, \quad (45)$$

en la que $j = -1$ y $j = 0, \dots, 7$ corresponden al final de la trama previa y la trama actual, respectivamente. Se calcula otro conjunto de 9 energías máximas desplazando los índices de señal en la Ecuación (45) en 16 muestras. Es

decir

$$E_{st}^{(2)}(j) = \max_{i=0}^{31} (s^2(i + 32j + 16)), \quad j = -1, \dots, 7. \quad (46)$$

Para aquellas energías que son suficientemente bajas, es decir, que cumplen la condición $10\log(E_{st}(j)) < 37$, se calcula la relación siguiente:

$$rat^{(1)}(j) = \frac{E_{st}^{(1)}(j+1)}{E_{st}^{(1)}(j) + 100}, \quad \text{para } j = -1, \dots, 6, \quad (47)$$

para el primer conjunto de índices y el mismo cálculo se repite para $E_{st}^{(2)}(j)$ para obtener dos conjuntos de relaciones $rat^{(1)}(j)$ y de $rat^{(2)}(j)$. El único máximo en estos dos conjuntos se busca como sigue:

$$dEO = \max(rat^{(1)}(j), rat^{(2)}(j)) \quad (48)$$

que es el aumento máximo de energía de corta duración en el nivel bajo.

Medida de la planitud del espectro de ruido

En este ejemplo, las tramas inactivas son codificadas normalmente con un modo de codificación diseñado para voz sorda en ausencia de operación DTX. Sin embargo, en el caso de un ruido de fondo cuasi periódico, tal como algunos ruidos de coche, se consigue una reproducción más fiel de ruido si en su lugar se usa una codificación genérica para WB.

Para detectar este tipo de ruido de fondo, se calcula una medida de la planitud del espectro de ruido de fondo y se promedia en el tiempo. En primer lugar, se calcula la energía media de ruido para las cuatro primeras y últimas bandas críticas de la manera siguiente:

$$\bar{N}_{l4} = \frac{1}{4} \sum_{i=0}^3 N_{CB}(i)$$

$$\bar{N}_{h4} = \frac{1}{4} \sum_{i=15}^{19} N_{CB}(i)$$

A continuación, la medida de planitud es calculada usando la relación siguiente:

$$f_{ruido_plano} = (\bar{N}_{l4} - \bar{N}_{h4}) / \bar{N}_{l4} + 0,5 [N_{CB}(1) + N_{CB}(2)] / N_{CB}(0)$$

y es promediada en el tiempo usando la relación siguiente:

$$\bar{f}_{ruido_plano}^{[0]} = 0,99 \bar{f}_{ruido_plano}^{[-1]} + 0,01 f_{ruido_plano}$$

en la que $\bar{f}_{ruido_plano}^{[-1]}$ es la medida de planitud promediada de la trama pasada y $\bar{f}_{ruido_plano}^{[0]}$ es el valor actualizado de la medida de planitud promediada de la trama actual.

Clasificación señal sorda

La clasificación de tramas de señal sordas se basa en los parámetros descritos anteriormente, concretamente: la medida $\bar{C}_{norm}$ de sonoridad, la inclinación espectral media $\bar{e}_t$, el aumento máximo de energía de corta duración a baja nivel $dE0$ y la medida de la planitud de espectro de ruido de fondo, $\bar{f}_{ruido_plano}^{[0]}$. La clasificación es soportada por el parámetro de estabilidad tonal y la energía relativa de trama calculada durante la fase de actualización de energía de ruido (módulo 107 en la Figura 1). La energía relativa de trama es calculada usando la relación siguiente:

$$E_{rel} = E_t - \bar{E}_f \quad (50)$$

en la que E_t es la energía de trama total (en dB) calculada en la Ecuación (6) y $\bar{E}_f$ es la energía media de trama largo plazo, actualizada en cada trama activa usando la relación siguiente:

$$\bar{E}_f = 0,99\bar{E}_f - 0,01E_t.$$

La actualización tiene lugar sólo cuando el indicador SAD está establecido (variable SAD igual a 1).

Las reglas para la clasificación como sordas para las señales de WM se resumen a continuación:

[[$(\bar{C}_{norm} < 0,695)$ AND $(\bar{e}_t < 4,0)$] OR $(E_{rel} < -14)$] AND
[Última trama INACTIVA o SORDA OR $(e_{antiguo} < 2,4)$ AND
 $(C_{norm}(d_0) + r_e < 0,66)$] AND
 $[dE0 < 250]$ AND
 $[e(1) < 2,7]$ AND

$[(\text{indicador SAD local} = 1)$ OR $(\bar{f}_{ruido_plano}^{[0]} < 1,45)$ OR $(\bar{N}_f < 20)]$ AND

NOT $[(\text{estabilidad_tonal})$ AND $(\bar{C}_{norm} > 0,52)$ AND $(\bar{e}_t > 0,5)$] OR $(\bar{e}_t > 0,85)$] AND $(E_{rel} > -14)$ AND
indicador SAD establecido a 1]

La primera línea de la condición está relacionada con las señales de baja energía y las señales con baja correlación que concentran su energía en las frecuencias altas. La segunda línea cubre los desplazamientos sonoros, la tercera línea cubre los segmentos explosivos de una señal y la cuarta línea es para los inicios sonoros. La quinta línea asegura un espectro plano en caso de tramas inactivas ruidosas. La última línea discrimina las señales musicales que si no serían declaradas como sordas.

Para las señales de NB, la condición de clasificación sorda tiene la forma siguiente:

$[(\text{indicador SAD local establecido a } 0)$ OR $(E_{rel} < -25)$ OR
 $(\bar{C}_{norm} < 0,61)$ AND $\bar{e}_t < 7,0$] AND (última trama INACTIVA OR
SORDA OR $(e_{antiguo} < 7,0)$ AND $(C_{norm}(d_0) + r_e < 0,52)$)] AND
 $[dE0 < 250]$ AND
 $[\bar{e}_t < 390]$ AND
NOT $[(\text{estabilidad_tonal})$ AND $(\bar{C}_{norm} > 0,52)$ AND $(\bar{e}_t > 0,5)$] OR $(\bar{e}_t > 0,75)$] AND
 $(E_{rel} > -10)$ AND indicador SAD establecido a 1]

Los árboles de decisión para el caso WB y el caso NB se muestran en la Figura 6. Si se cumplen las condiciones combinadas, la clasificación termina seleccionando el modo de codificación sorda.

Clasificación de señal sonora

Si una trama no es clasificada como trama inactiva o como trama sorda, entonces se comprueba si es una trama sonora estable. La regla de decisión se basa en la correlación normalizada en cada subtrama (con una resolución

de $\frac{1}{4}$ de submuestra), la inclinación espectral media y las estimaciones de tono de bucle abierto en todas las subtramas (con una resolución de $\frac{1}{4}$ de submuestra).

El procedimiento de estimación de tono en bucle abierto es realizado por el módulo 106 analizador LP y rastreador de tono de la Figura 1. En la Ecuación (19), se usan tres estimaciones de tono en bucle abierto: d_0 , d_1 y d_2 , correspondientes a la primera mitad de la trama, la segunda mitad de la trama y la anticipación. Con el fin de obtener información de tono precisa en todas las cuatro subtramas, se calcula una mejora de tono fraccionario con resolución de $\frac{1}{4}$ de muestra. Esta mejora es calculada sobre la señal sonora ponderada $s_{wd}(n)$. En esta realización ejemplar, la señal ponderada $s_{wd}(n)$ no es diezmada por la mejora de estimación de tono en bucle abierto. Al principio de cada subtrama, se realiza un análisis de correlación corto (64 muestras a una frecuencia de muestreo de 12,8 kHz) con una resolución de 1 muestra en el intervalo $(-7, +7)$ usando los retardos siguientes: para las subtramas primera y segunda y d_1 para las subtramas tercera y cuarta. A continuación, las correlaciones son interpoladas alrededor de sus máximos en las posiciones fraccionarias $d_{max} - 3/4$, $d_{max} - 1/2$, $d_{max} - 1/4$, d_{max} , $d_{max} + 1/4$, $d_{max} + 1/2$, $d_{max} + 3/4$. El valor que produce la correlación máxima es seleccionado como el retardo de tono mejorado.

Denótese los retardos de tono de bucle abierto mejorados en las cuatro subtramas como $T(0)$, $T(1)$, $T(2)$ y $T(3)$ y sus correspondientes correlaciones normalizadas como $C(0)$, $C(1)$, $C(2)$ y $C(3)$. Entonces, la condición de clasificación de la señal sonora viene determinada por:

$$\begin{aligned} & [C(0) > 0,605] \text{ AND} \\ & [C(1) > 0,605] \text{ AND} \\ & [C(2) > 0,605] \text{ AND} \\ & [C(3) > 0,605] \text{ AND} \\ & [\bar{e}_t > 4] \text{ AND} \\ & [|T(1) - T(0)| < 3] \text{ AND} \\ & [|T(2) - T(1)| < 3] \text{ AND} \\ & [|T(3) - T(2)| < 3] \end{aligned}$$

La condición dice que la correlación normalizada es suficientemente alta en todas las subtramas, las estimaciones de tono no divergen en toda la trama y la energía se concentra en las frecuencias bajas. Si esta condición se cumple, la clasificación termina seleccionando el modo de codificación de señal sonora, de lo contrario, la señal es codificada mediante un modo de codificación de señal genérica. La condición se aplica tanto a las señales de WB como a las señales de NB.

Estimación de tonalidad en el contenido de banda súper ancha

En la codificación de señales de banda súper ancha, se usa un modo de codificación específico para las señales de sonido con estructura tonal. El intervalo de frecuencias de interés es principalmente 7.000-14.000 Hz, pero también puede ser diferente. El objetivo es detectar tramas que tienen un fuerte contenido tonal en el intervalo de interés de manera que el modo de codificación específico de tono puede ser usado de manera eficiente. Esto se realiza usando el análisis de estabilidad tonal descrito anteriormente en la presente descripción. Sin embargo, hay algunas aberraciones que se describen en esta sección.

En primer lugar, el suelo espectral que es restado del espectro logarítmico de energía se calcula de la manera siguiente. El espectro logarítmico de energía es filtrado usando un filtro de media móvil (Moving Average, MA), o un filtro FIR, cuya longitud es $L_{MA} = 15$ muestras. El espectro filtrado viene determinado por:

$$suelo_espec(j) = \frac{1}{2L_{MA} + 1} \sum_{k=-L_{MA}}^{L_{MA}} E_{dB}(j+k), \text{ para } j=L_{MA}, \dots, N_{ESPEC}-L_{MA}-1.$$

Para evitar la complejidad computacional, la operación de filtrado es realizada sólo para $j = L_{MA}$ y para los otros retardos, se calcula como:

$$\begin{aligned} & suelo_espec(j) = suelo_espec(j-1) + \frac{1}{2L_{MA} + 1} [E_{dB}(j+L_{MA}) - E_{dB}(j-L_{MA}-1)], \\ & \text{para } j=L_{MA}+1, \dots, N_{ESPEC}-L_{MA}-1. \end{aligned}$$

Para los retardos $0, \dots, L_{MA}-1$ y $N_{SPEC}-L_{MA}, \dots, N_{SPEC}-1$, el suelo espectral se calcula mediante extrapolación. Más específicamente, se usa la relación siguiente:

$$suelo_espec(j) = 0,9suelo_espec(j+1) + 0,1E_{dB}(j), \text{ para } j=L_{MA}-1, \dots, 0,$$

$$suelo_espec(j) = 0,9suelo_espec(j-1) + 0,1E_{dB}(j), \text{ para } j=N_{SPEC}-L_{MA}, \dots, N_{SPEC}-1.$$

En la primera ecuación anterior la actualización continúa desde $L_{MA}-1$ hacia abajo hasta 0.

A continuación, el suelo espectral es restado del espectro logarítmico de energía de la misma manera que se ha descrito anteriormente en la presente descripción.

A continuación, el espectro residual, denotado como $E_{res,dB}(j)$, es suavizado sobre 3 muestras como sigue usando un filtro de media móvil de tiempo corto:

$$E'_{res,dB}(j) = 0,33[E_{res,dB}(j-1) + E_{res,dB}(j) + E_{res,dB}(j+1)], \text{ para } j=1, \dots, N_{SPEC}-1.$$

La búsqueda de mínimos espectrales y sus índices, el cálculo del mapa de correlación y el mapa de correlación a largo plazo son los mismos que en el procedimiento descrito anteriormente en la presente descripción, usando el espectro suavizado $E'_{res,dB}(j)$.

La decisión acerca de la tonalidad de la señal en el contenido de banda súper ancha también es el mismo que el descrito anteriormente en la presente descripción, es decir, basado en un umbral adaptativo. Sin embargo, en este caso se usan un umbral fijo y etapa diferentes. El umbral *umbral_tonal* se inicializa a 130 y se actualiza en cada trama de la manera siguiente:

```

if (sum_map_cor > 130)
    umbral_tonal = umbral_tonal - 1,0
else
    umbral_tonal = umbral_tonal + 1,0
end.
```

El umbral *umbral_tonal* adaptativo tiene un límite superior de 140 y un límite inferior de 120. El umbral fijo se ha establecido con respecto al intervalo de frecuencias de 7.000 a 14.000 Hz. Para un intervalo diferente, tendrá que ser ajustado. Como regla general, se puede aplicar la relación siguiente $umbral_tonal = N_{SPEC}/2$.

La última diferencia con el procedimiento descrito anteriormente en la presente descripción es que no se usa la detección de tonos fuertes en el contenido de banda súper ancha. Esto está motivado por el hecho de que los tonos fuertes no son perceptualmente adecuados para el propósito de codificar la señal tonal en el contenido de banda súper ancha.

La presente invención ha sido descrita en la descripción anterior por medio de una realización ilustrativa, no restrictiva de la misma. El alcance de la presente invención está definido por las reivindicaciones adjuntas.

REIVINDICACIONES

1. Un procedimiento para estimar una tonalidad de una señal de sonido, en el que el procedimiento comprende:

- 5 calcular un espectro residual actual de la señal de sonido;
 detectar los picos en el espectro residual actual;
 calcular un mapa de correlación entre el espectro residual actual y un espectro residual previo para cada
 pico detectado; y
 10 calcular un mapa de correlación a largo plazo basado en el mapa de correlación calculado, en el que el
 mapa de correlación a largo plazo es indicativo de una tonalidad en la señal de sonido.

2. Procedimiento según la reivindicación 1, en el que el cálculo del espectro residual actual comprende:

- 15 buscar los mínimos en el espectro de la señal de sonido en una trama actual;
 estimar un suelo espectral conectando los mínimos entre sí; y
 restar el suelo espectral estimado del espectro de la señal de sonido en la trama actual para producir el
 espectro residual actual.

3. Procedimiento según la reivindicación 1 o 2, en el que la detección de los picos en el espectro residual actual
 20 comprende localizar un máximo entre cada par de dos mínimos consecutivos.

4. Procedimiento según la reivindicación 1, 2 o 3, en el que el cálculo del mapa de correlación comprende:

- 25 para cada pico detectado en el espectro residual actual, calcular un valor de correlación normalizado con el
 espectro residual anterior, sobre los contenedores de frecuencia entre dos mínimos consecutivos en el
 espectro residual actual que delimitan el pico; y
 asignar una puntuación a cada pico detectado, en el que la puntuación corresponde al valor de correlación
 normalizado; y
 30 para cada pico detectado, asignar el valor de correlación normalizado del pico sobre los contenedores de
 frecuencia entre los dos mínimos consecutivos que delimitan el pico para formar el mapa de correlación.

5. Procedimiento según cualquiera de las reivindicaciones anteriores, en el que el cálculo del mapa de correlación
 a largo plazo comprende:

- 35 filtrar el mapa de correlación a través de un filtro de un polo de contenedor de frecuencias en contenedor de
 frecuencias; y
 sumar el mapa de correlación filtrado sobre los contenedores de frecuencia para producir un mapa de
 correlación sumado a largo plazo.

40 6. Procedimiento para detectar actividad sonora en una señal de sonido, en el que la señal de sonido es clasificada
 como una de entre una señal de sonido inactiva y una señal de sonido activa según la actividad sonora detectada
 en la señal de sonido, en el que el procedimiento comprende:

- 45 estimar un parámetro relacionado con una tonalidad de la señal de sonido usada para distinguir una señal
 musical de una señal de ruido de fondo; en el que la estimación del parámetro relacionado con la tonalidad
 de la señal de sonido previene la actualización de las estimaciones de energía de sonido cuando se detecta
 una señal musical;
 en el que la estimación de tonalidad es realizada según una cualquiera de las reivindicaciones 1 a 5.

50 7. Procedimiento según la reivindicación 6, que comprende además calcular un parámetro no estacionariedad
 complementaria y un parámetro carácter de ruido con el fin de distinguir una señal musical de una señal de ruido
 de fondo y evitar la actualización de las estimaciones de energía de ruido en la señal musical.

55 8. Procedimiento según la reivindicación 7, en el que el cálculo del parámetro no estacionariedad complementaria
 comprende calcular un parámetro similar a un no estacionariedad convencional con restablecimiento de energía a
 largo plazo cuando se detecta un ataque espectral.

60 9. Procedimiento según la reivindicación 8, en el que la detección del ataque espectral y el restablecimiento de la
 energía a largo plazo comprende calcular un parámetro diversidad espectral y en el que el cálculo del parámetro
 diversidad espectral comprende:

calcular una relación entre una energía de la señal de sonido en una trama actual y una energía de la señal de sonido en una trama previa, para las bandas de frecuencia más altas que un número determinado; y calcular la diversidad espectral como una suma ponderada de la relación calculada sobre todas las bandas de frecuencia más altas que el número determinado.

- 5
10. Procedimiento según la reivindicación 8 o 9, en el que el cálculo del parámetro carácter de ruido comprende:

dividir una pluralidad de bandas de frecuencia en un primer grupo de un cierto número de primeras bandas de frecuencia y un segundo grupo de un resto de las bandas de frecuencia;
10 calcular un primer valor de energía para el primer grupo de bandas de frecuencia y un segundo valor de energía del segundo grupo de bandas de frecuencias;
calcular una relación entre los valores de energía primero y segundo para producir el parámetro carácter de ruido; y
calcular un valor a largo plazo del parámetro carácter ruido en base al parámetro carácter de ruido
15 calculado;
en el que la actualización de las estimaciones de energía de ruido se evita si el parámetro carácter de ruido es menor que un umbral fijo determinado.
- 20 11. Un procedimiento de clasificación de una señal de sonido con el fin de optimizar la codificación de la señal de sonido usando la clasificación de la señal de sonido, en el que el procedimiento comprende:

detectar una actividad sonora en la señal de sonido;
clasificar la señal de sonido como una de entre una señal de sonido inactivo y una señal de sonido activo según la actividad sonora detectada en la señal de sonido; y
25 en respuesta a la clasificación de la señal de sonido como una señal de sonido activo, clasificar adicionalmente la señal de sonido activo como una de entre una señal de voz sorda y una señal de voz no sorda;
en el que la clasificación de la señal de sonido activo como una señal de voz sorda comprende la estimación de una tonalidad de la señal de sonido con el fin de evitar la clasificación de las señales
30 musicales como señales de voz sorda, en el que la estimación de tonalidad es realizada según una cualquiera de las reivindicaciones 1 a 5.
- 35 12. Procedimiento según la reivindicación 11, que comprende además codificar la señal de sonido según la clasificación de la señal de sonido, en el que la codificación de la señal de sonido según la clasificación de la señal de sonido comprende codificar la señal de sonido inactivo usando generación de ruido de confort.
13. Procedimiento según la reivindicación 11 o 12, en el que la clasificación de la señal de sonido activo como una señal de voz sorda comprende calcular una regla de decisión en base a al menos una de entre una medida de sonoridad, una medida de inclinación espectral media, un aumento máximo de energía de corto tiempo a bajo
40 nivel, una estabilidad tonal y una energía relativa de trama.
14. Un procedimiento para codificar una banda superior de una señal de sonido usando una clasificación de la señal de sonido, en el que el procedimiento comprende:

45 clasificar la señal de sonido como una de entre una señal de sonido tonal y una señal de sonido no tonal;
en el que la clasificación de la señal de sonido como una señal tonal comprende estimar una tonalidad de la señal de sonido según una cualquiera de las reivindicaciones 1 a 5.
- 50 15. Procedimiento según la reivindicación 14, en el que la estimación de la tonalidad de la señal de sonido según una cualquiera de las reivindicaciones 1 a 5 comprende además el uso de un procedimiento alternativo para calcular un suelo espectral, en el que el uso del procedimiento alternativo para calcular el suelo espectral comprende filtrar un espectro logarítmico de energía de la señal de sonido en una trama actual usando un filtro de media móvil.
- 55 16. Procedimiento según la reivindicación 14 o 15, en el que la estimación de la tonalidad de la señal de sonido según una cualquiera de las reivindicaciones 1 a 5 comprende además suavizar el espectro residual por medio de un filtro de media móvil de tiempo corto.
- 60 17. Procedimiento según la reivindicación 14 o 16, que comprende además codificar la banda superior de la señal de sonido según la clasificación de dicha señal de sonido.

18. Procedimiento según cualquiera de las reivindicaciones 14 a 17, en el que la banda superior de la señal de sonido comprende un intervalo de frecuencias por encima de 7 kHz.

19. Un dispositivo para estimar una tonalidad de una señal de sonido, en el que el dispositivo comprende:

un calculador para calcular un espectro residual actual de la señal de sonido;
un detector para detectar los picos en el espectro residual actual;
un calculador para calcular un mapa de correlación entre el espectro residual actual y un espectro residual previo para cada pico detectado; y
un calculador para calcular un mapa de correlación a largo plazo en base al mapa de correlación calculado, en el que el mapa de correlación a largo plazo es indicativo de una tonalidad en la señal de sonido.

20. Un dispositivo según la reivindicación 19, en el que el calculador del espectro residual actual comprende:

un localizador de mínimos en el espectro de la señal de sonido en una trama actual;
un estimador de un suelo espectral que conecta los mínimos entre sí; y
un restador del suelo espectral estimado del espectro para producir el espectro residual actual.

21. Dispositivo según la reivindicación 19 o 20, en el que el calculador del mapa de correlación a largo plazo comprende:

un filtro para filtrar el mapa de correlación de contenedor de frecuencias en contenedor de frecuencias; y
un sumador para sumar el mapa de correlación filtrado sobre los contenedores de frecuencia con el fin de producir un mapa sumado de correlación a largo plazo.

22. Un dispositivo para detectar la actividad sonora en una señal de sonido, en el que la señal de sonido es clasificada como una de entre una señal de sonido inactivo y una señal de sonido activo según la actividad sonora detectada en la señal de sonido, en el que el dispositivo comprende:

un estimador de tonalidad para la señal de sonido, usado para distinguir una señal musical de una señal de ruido de fondo;
en el que el estimador de tonalidad comprende un dispositivo según una cualquiera de las reivindicaciones 19 a 21.

23. Un dispositivo para clasificar una señal de sonido con el fin de optimizar la codificación de la señal de sonido usando la clasificación de la señal de sonido, en el que el dispositivo comprende:

un detector para detectar una actividad sonora en la señal de sonido;
un primer clasificador de señal de sonido para clasificar la señal de sonido como una de entre una señal de sonido inactivo y una señal de sonido activo según la actividad sonora detectada en la señal de sonido;
un segundo clasificador de señal de sonido en conexión con el primer clasificador de sonido para clasificar la señal de sonido activo como una de entre una señal de voz sorda y una señal de voz no sorda;
en el que el detector de actividad sonora comprende un estimador de tonalidad para estimar una tonalidad de la señal de sonido con el fin de evitar la clasificación de las señales musicales como señales de voz sorda en el que el estimador de tonalidad comprende un dispositivo según una cualquiera de las reivindicaciones 19 a 21.

24. Dispositivo según la reivindicación 23, que comprende además un codificador de sonido para codificar la señal de sonido según la clasificación de la señal de sonido, en el que el codificador de sonido es seleccionado de entre el grupo que consiste en: un codificador de ruido para codificar las señales de sonido inactivas, un codificador optimizado para voz sorda, un codificador optimizado para voz sonora para codificar señales sonoras estables, y un codificador de señal de sonido genérico para codificar señales sonoras de evolución rápida.

25. Un dispositivo para codificar una banda superior de una señal de sonido usando una clasificación de la señal de sonido, en el que el dispositivo comprende:

un clasificador de señal de sonido para clasificar la señal de sonido como una de entre una señal de sonido tonal y una señal de sonido no tonal; y
un codificador de sonido para codificar la banda superior de la señal de sonido clasificada; en el que el clasificador de señal de sonido comprende un dispositivo para estimar una tonalidad de la señal de sonido según una cualquiera de las reivindicaciones 19 a 21.

26. Dispositivo según la reivindicación 25, que comprende además un filtro de media móvil para calcular un suelo espectral derivado de la señal de sonido, en el que el suelo espectral se usa en la estimación de la tonalidad de la señal de sonido.

5

27. Dispositivo según la reivindicación 25 o 26, que comprende además un filtro de media móvil de tiempo corto para suavizar un espectro residual de la señal de sonido, en el que el espectro residual se usa en la estimación de la tonalidad de la señal de sonido.

10

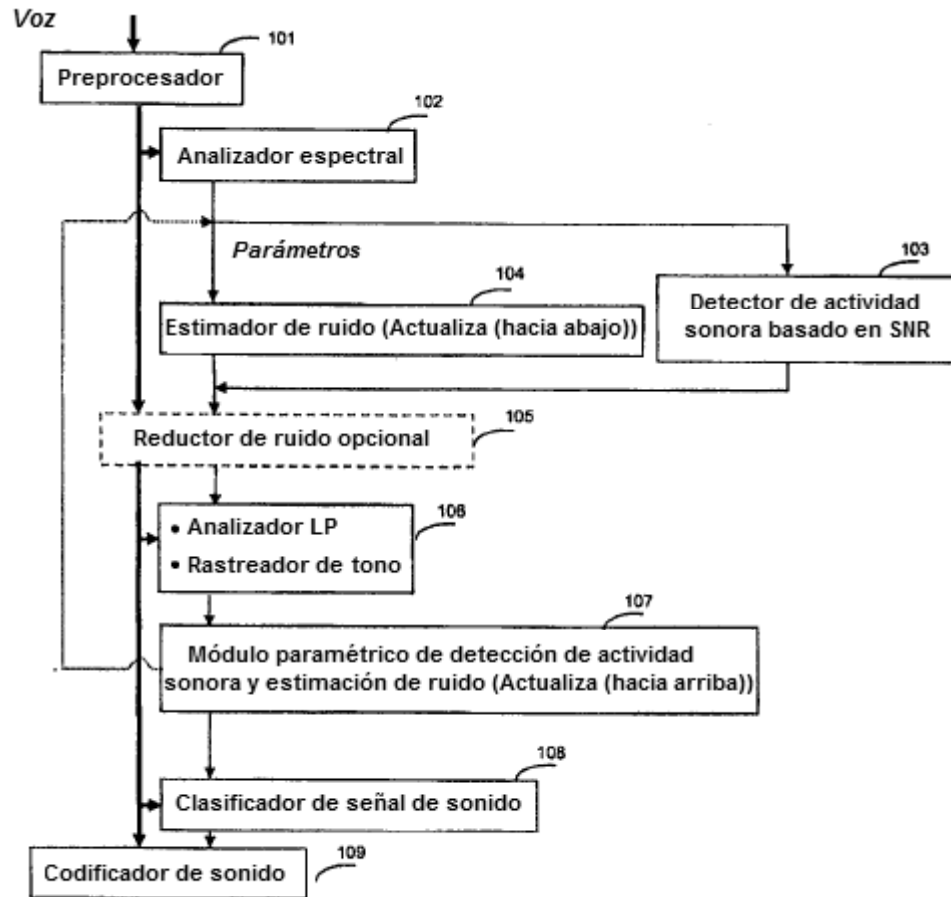


Figura 1

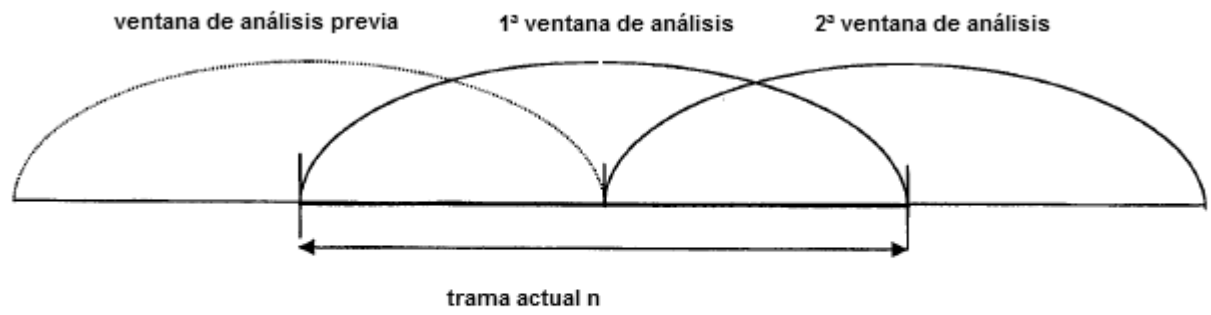


Figura 2

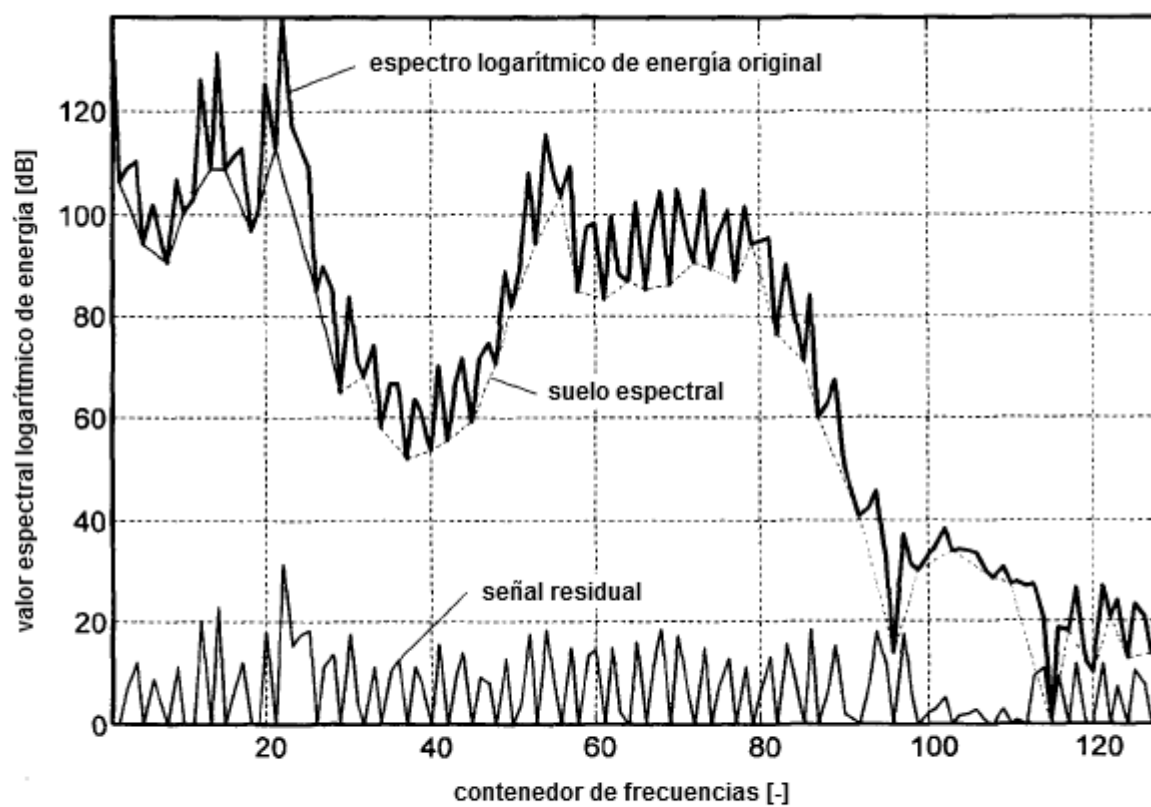


Figura 3

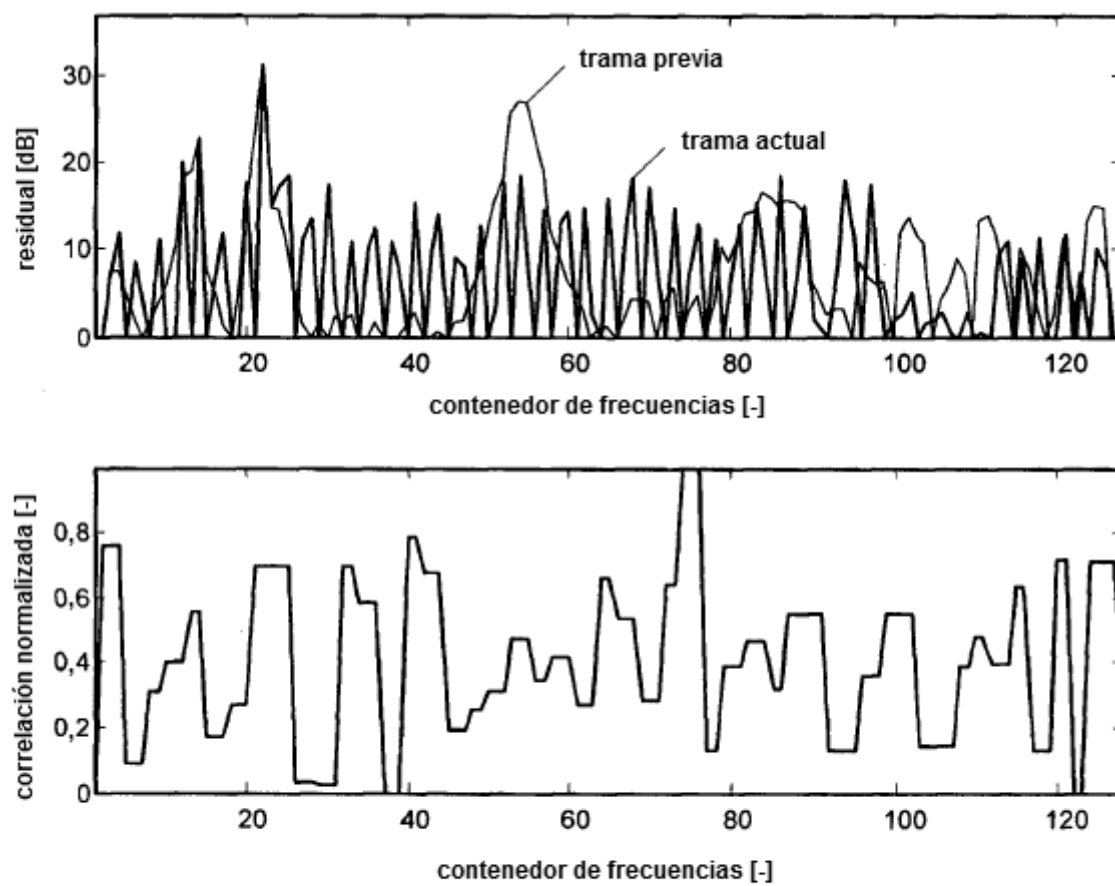


Figura 4

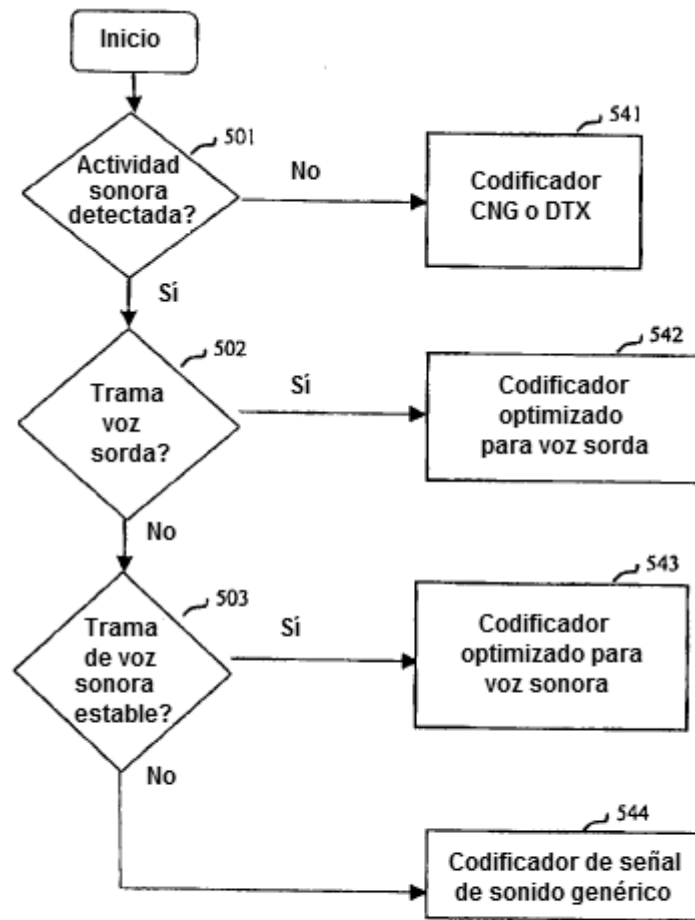


Figura 5

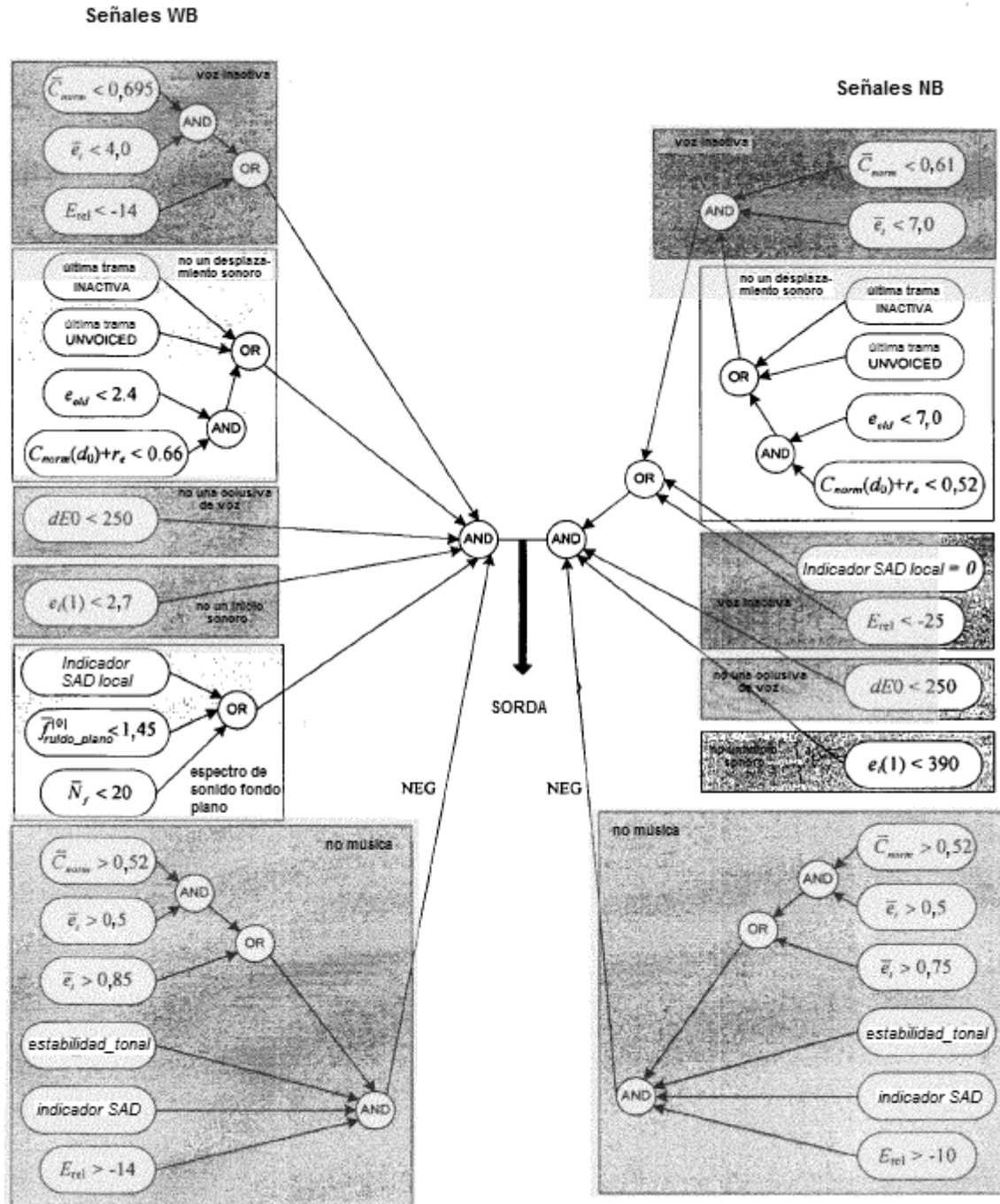


Figura 6