(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 913 979**

(51) Int. CI.:

G10L 25/90 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **02.10.2017 PCT/EP2017/074984**

(87) Fecha y número de publicación internacional: **12.04.2018 WO18065366**

(96) Fecha de presentación y número de la solicitud europea: **02.10.2017 E 17772748 (4)**

(97) Fecha y número de publicación de la concesión europea: **23.03.2022 EP 3523802**

(54) Título: **Aparato y método para determinar una información de la altura del sonido**

(30) Prioridad:

04.10.2016 EP 16192253

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
07.06.2022

(73) Titular/es:

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

(72) Inventor/es:

LECOMTE, JÉRÉMIE y
TOMASEK, ADRIAN

(74) Agente/Representante:

ARIZTI ACHA, Monica

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

Aparato y método para determinar una información de la altura del sonido

La presente invención se refiere al procesamiento de señales de audio, de manera más específica, se refiere a obtener una información de la altura del sonido de una señal de audio.

Antecedentes de la invención

En algunos algoritmos, la determinación de la altura del sonido se realiza basándose en una autocorrelación de una señal de audio. Sin embargo, estos algoritmos emplean una cantidad estática de muestras de la señal para grandes intervalos de desfases de la altura del sonido.

En consecuencia, un problema de soluciones conocidas es que se obtiene información inexacta de la altura del sonido, debido a una consideración insuficientemente flexible de las muestras de la señal de la señal de audio para la determinación de la información de la altura del sonido.

Por lo tanto, existe el deseo de un concepto que proporcione un mejor compromiso entre la complejidad del cálculo y la exactitud de una determinación del valor de la altura del sonido.

Medan *et al*. en “Super resolution pitch determination of speech signals”, 1991, propone el cálculo de la correlación cruzada con ventanas cuyas longitudes dependen del desfase de la altura del sonido candidato.

Sumario de la invención

Una realización de acuerdo con la invención crea un aparato para determinar una información de la altura del sonido basándose en una señal de audio. El aparato se configura para obtener un valor de la similitud que está asociado con un par dado de porciones de la señal de audio que tienen un desplazamiento temporal dado. Además, el aparato se configura para elegir una longitud de las porciones de la señal de la señal de audio utilizada para obtener un valor de la similitud para el desplazamiento temporal dado, dependiendo del desplazamiento temporal dado. Además, el aparato se configura para elegir la longitud de las porciones de la señal para ser linealmente dependiente del desplazamiento temporal dado, con una tolerancia de ± 1 muestra.

El aparato descrito permite una determinación exacta de una información de la altura del sonido, mientras que evita una evaluación de porciones innecesariamente grandes de la señal de audio. Se logra una determinación razonablemente exacta de la altura del sonido utilizando una longitud suficiente de las porciones de la señal y se logra una baja complejidad del cálculo utilizando una longitud razonablemente pequeña de las porciones de la señal consideradas. Por lo tanto, la dependencia lineal de la longitud de la porción de la señal con el desplazamiento temporal dado proporciona un buen intercambio, puesto que evita longitudes excesivas de las porciones de la señal, mientras que proporciona todavía porciones de la señal suficientemente largas para obtener una información exacta de la altura del sonido. Puesto que una información de la altura del sonido es una información sobre la frecuencia, una periodicidad se asocia con la misma. La longitud del periodo de la altura del sonido que corresponde a una altura del sonido se caracteriza por un desplazamiento temporal que resulta en un alto valor de la similitud. Por lo tanto, es beneficioso emplear porciones de la señal de una longitud que sea linealmente dependiente del desplazamiento temporal dado. En otras palabras, por ejemplo, para verificar si una señal tiene una altura del sonido baja que corresponde a un periodo largo de la altura del sonido, se utiliza un desplazamiento temporal grande. En este caso, cuando se emplea una dependencia lineal con una pendiente positiva, se elige una longitud de la porción de la señal apropiadamente más larga, para la determinación de la información de la altura del sonido, en comparación con cuando se verifica una altura del sonido más alta que corresponde a un periodo de la altura del sonido comparativamente más corto. Así, el concepto permite ajustar la longitud de las porciones, de manera que se utiliza una porción razonable de una señal bajo consideración, tanto cuando se evalúa un desplazamiento temporal más pequeño como cuando se evalúa un desplazamiento temporal más grande.

De acuerdo con una realización preferida de la invención, el aparato se configura para obtener una información de la altura del sonido, basándose en una secuencia de valores de la similitud. La consideración de más de un valor de la similitud mejora la exactitud de la altura del sonido determinada.

De acuerdo con una realización preferida de la invención, el aparato se configura para obtener la secuencia de valores de la similitud, basándose en los valores de la similitud para los desplazamientos temporales en un intervalo que empieza entre 1 ms y 4 ms y que se extiende hasta desplazamientos temporales de entre 15 ms a 25 ms. La realización descrita es beneficiosa, puesto que el intervalo considerado de desplazamientos temporales es un intervalo característico para el habla humana, que corresponde a las frecuencias fundamentales del habla.

Además, la restricción del intervalo de los desplazamientos temporales a los valores descritos reduce la complejidad del cálculo para determinar las secuencias de los valores de la similitud, puesto que limita la cantidad de valores de la similitud que necesitan determinarse.

De acuerdo con una realización preferida adicional de la invención, el aparato se configura para incrementar paso a paso la longitud de las porciones de la señal, en etapas de una muestra con un desplazamiento temporal que se incrementa, cuando se obtienen valores de la similitud para diferentes pares de porciones que tienen diferentes desplazamientos temporales. La realización descrita es especialmente útil debido a su capacidad para proporcionar porciones de la señal con una diferencia de la longitud mínima. En otras palabras, se logra una granularidad fina de las longitudes, permitiendo una elección flexible de las longitudes de la porción de la señal, permitiendo por lo tanto un buen intercambio entre la exactitud y la complejidad del cálculo. De acuerdo con una realización preferida de la invención, el aparato se configura para incrementar la longitud de las porciones de la señal con precisión de enteros, con un desplazamiento temporal que se incrementa, cuando se obtienen valores de la similitud para diferentes pares de porciones que tienen diferentes desplazamientos temporales. El incremento de la longitud de las porciones de la señal con precisión de enteros es especialmente beneficioso, debido a la baja complejidad del cálculo involucrado en ello. En otras palabras, por ejemplo, no necesita considerarse el sobremuestreo o los retrasos fraccionarios.

De acuerdo con una realización preferida de la invención, el aparato se configura para incrementar la longitud de las porciones de la señal, entre una longitud mínima predeterminada y una longitud máxima predeterminada, linealmente en dependencia con el desplazamiento temporal. La longitud mínima predeterminada se utiliza para un desplazamiento temporal más corto que corresponde a una frecuencia de la altura del sonido máxima, y la longitud máxima predeterminada se utiliza para un desplazamiento temporal más largo que corresponde a una frecuencia de la altura del sonido mínima. La realización descrita ayuda a mantener la complejidad del cálculo dentro de un intervalo prescrito determinado por la longitud mínima predeterminada y la longitud máxima predeterminada. Además, la longitud mínima predeterminada y la longitud máxima predeterminada pueden elegirse de acuerdo, por ejemplo, con el tracto vocal humano, para captar, por ejemplo, todo un ciclo de un periodo de la altura del sonido considerado.

De acuerdo con una realización preferida de la invención, el aparato se configura para elegir la longitud de las porciones de la señal basándose en

$$Len(d) = m \cdot d + startlen - Pitmin \cdot m,$$

en donde *d* es el desplazamiento temporal dado, *startlen* es una longitud mínima predeterminada para las porciones de la señal, *Pitmin* es un valor predeterminado del desfase de la altura del sonido considerado más pequeño, que representa un valor mínimo para d, y m es un factor mediante el cual se escala el desplazamiento temporal dado, en donde por ejemplo, $m \leq 1$. Además, el aparato se configura para elegir la longitud de las porciones de la señal como un valor entero cercano a *Len(d)*. La elección de un valor entero cercano a *Len(d)* puede basarse en una función de redondeo, una función de suelo, una función de techo o una función de truncamiento. La función de redondeo redondea el valor de *Len(d)* al valor entero más cercano, la función de suelo redondea el valor de *Len(d)* al entero más cercano hacia menos infinito, la función de techo redondea el valor de *Len(d)* hacia el siguiente entero en la dirección de más infinito y la función de truncamiento elimina cualquier valor decimal de *Len(d)*, regresando así a un valor entero.

De acuerdo con una realización preferida de la invención, el aparato se configura para calcular un valor de autocorrelación basándose en dos porciones de la señal desplazadas temporalmente de la señal de audio, desplazadas temporalmente por el desplazamiento temporal dado, con el fin de obtener el valor de la similitud, en el que un valor de la similitud puede ser un valor de autocorrelación o un valor derivado de un valor de autocorrelación. Además, el número de valores de muestra de la señal de audio considerado en el cálculo del valor de autocorrelación se determina por la longitud elegida. El uso de una autocorrelación para la estimación de la altura del sonido es especialmente beneficioso debido a una baja complejidad del cálculo involucrado en el cálculo de una autocorrelación. Variar el número de valores de muestra utilizado para calcular el valor de autocorrelación como se describió permite la estimación de frecuencias de la altura del sonido más exactas, mientras que se evita una longitud de la suma de la autocorrelación innecesariamente larga para los desplazamientos temporales pequeños.

De acuerdo con una realización preferida de la invención, el aparato se configura para obtener los valores de la similitud basándose en

$$R'(d) = \sum_{n=0}^{Len(d)} s(n)s(n-d),$$

en donde *s(n)* es una muestra de la señal de audio al tiempo n, *Len(d)* es una información sobre la longitud de las porciones de la señal para el desplazamiento temporal dado d y d es el desplazamiento temporal dado. El límite superior de la suma puede, por ejemplo, ser también *Len(d)*-1 y el valor d del desplazamiento temporal puede estar en el intervalo [*Pitmin, Pitmax*].

El cálculo de los valores de la similitud de la forma descrita ofrece una manera rápida y flexible de obtener los valores de autocorrelación. Especialmente, el límite superior de la suma (*Len(d)* o *Len(d)*-1), que depende del desplazamiento temporal considerado (*d*), puede proporcionar una porción de la señal suficientemente larga para comprender un periodo completo de la frecuencia de la altura del sonido a ser determinada.

De acuerdo con una realización preferida de la invención, el aparato se configura para obtener una información de la ubicación de un valor máximo de una pluralidad de valores de la similitud. Además, el aparato se configura para obtener una información de la altura del sonido basándose en la información de la ubicación que corresponde a un desplazamiento temporal considerado del valor máximo. La realización descrita es especialmente útil para reducir la complejidad del cálculo, puesto que una búsqueda de un valor máximo puede realizarse con una baja complejidad del cálculo. Esto puede formularse, por ejemplo, como

$$R(T_0) = \max_d R(d),$$

o

$$R'(T_0) = \max_d R'(d),$$

en donde $d \in [Pitmin; Pitmax]$ y T_0 denotan la ubicación de un máximo encontrado.

De acuerdo con una realización preferida de la invención, el aparato se configura para aplicar una normalización al valor de la similitud utilizando al menos dos valores de normalización. Los dos valores de normalización comprenden un primer valor de normalización que representa una característica estadística, por ejemplo, un valor de la energía, de una primera porción del par dado de porciones y un segundo valor de normalización que representa una característica estadística, por ejemplo, un valor de la energía, de una segunda porción del par dado de porciones. La normalización se aplica al valor de la similitud con el fin de derivar un valor normalizado de la similitud. La normalización descrita es útil para compensar las fluctuaciones de energía en la señal de audio, por ejemplo, fluctuaciones de la energía en una señal del habla. Así, se proporcionan los valores de la similitud que son comparables en un intervalo amplio de desplazamientos temporales, haciendo factible un resultado más exacto de la determinación de la altura del sonido.

De acuerdo con una realización preferida de la invención, el aparato se configura para obtener a valor normalizado de la similitud R(d) basándose en

$$R(d) = \frac{R\prime(d)w(d)}{\sqrt{norm(0)norm(d)}},$$

en donde R'(d) es un valor de la similitud y *w(d)* es una función de ventana. La normalización del valor de la similitud de la manera descrita permite una determinación más exacta de una información de la altura del sonido, debido a menos fluctuación de la energía del valor de la similitud. Especialmente, el valor considerado R'(d) puede estar sujeto a variaciones de la energía en las porciones de la señal consideradas para su determinación. El empleo de la normalización descrita libera al valor R(d) de las variaciones de la energía en las porciones de la señal consideradas.

De acuerdo con una realización preferida de la invención, el aparato se configura para derivar de manera recursiva un valor de normalización, por ejemplo, un valor de norma, para un nuevo desplazamiento temporal d de un valor de normalización para un desplazamiento temporal previo, por ejemplo, d-1, *d*-2 y así sucesivamente, sumando uno o más valores de la energía de las muestras de la señal incluidas en una nueva porción de la señal y no incluidas en una porción de la señal anterior, y sustrayendo uno o más valores de la energía de las muestras de la señal incluidas en la porción de la señal anterior y no incluidas en la nueva porción de la señal. El cálculo recursivo

descrito del valor de normalización permite un cálculo rápido y que ahorra memoria de un valor de normalización basado en un valor de normalización previo.

De acuerdo con una realización preferida de la invención, el aparato se configura para obtener un valor de normalización *norm(d)* basado en

$$norm(d) = norm(d-1) + x_d^2 - x_{d+Len(d)}^2,$$

en donde x_d es una muestra de la señal de audio contenida en la porción de la señal de acuerdo con el desplazamiento temporal *d* pero no contenida en la porción de la señal de acuerdo con el desplazamiento temporal *d*-1, $x_{d+Len(d)}$ es una muestra de la señal de audio no contenida en la porción de la señal de acuerdo con el desplazamiento temporal *d* pero contenida en la porción de la señal de acuerdo con el desplazamiento temporal *d*-1 de la señal de audio y *norm(d-1)* es un valor de normalización obtenido para una porción de la señal considerada previamente de acuerdo con el desplazamiento temporal *d*-1 fuera de la nueva porción de la señal del desplazamiento temporal *d*. La manera descrita de obtener un valor de normalización permite una manera rápida y simple de calcular un valor de normalización basado en un valor de normalización previo. Además, la estimación del valor de normalización de la forma descrita es especialmente adecuada para las realizaciones de la invención empleadas en dispositivos portátiles con bajo consumo de energía, puesto que el cálculo exhibe baja complejidad y baja demanda de memoria.

De acuerdo con una realización preferida adicional de la invención, el aparato se configura para determinar una información, por ejemplo, un índice o una información máxima local que es un resultado de una verificación máxima local, sobre una característica de un máximo identificado de una secuencia de valores de la similitud obtenida para diferentes desplazamientos temporales. Además, el aparato se configura para proporcionar una frecuencia de la altura del sonido basándose en el máximo identificado si la información sobre la característica del máximo identificado indica que el máximo identificado es un máximo local. Además, el aparato se configura para proceder a considerar uno o más de otros valores de la similitud que son diferentes del valor máximo previamente identificado para estimar la frecuencia de la altura del sonido si la información sobre la característica del máximo no indica que el máximo es un máximo local, por ejemplo, si indica que la ubicación está en un límite de un intervalo de búsqueda. Una información inexacta de la altura del sonido puede deberse al hecho de que está basada en un máximo identificado que no es un máximo local. Por lo tanto, una verificación del máximo identificado y del tratamiento resultante del máximo identificado en la forma descrita es útil para evitar la determinación de la información inexacta de la altura del sonido.

De acuerdo con una realización preferida de la invención, el aparato se configura para determinar si un máximo identificado se localiza en el límite de la secuencia de valores de la similitud como la información sobre una característica del máximo identificado. Si un máximo se localiza en el límite de la secuencia de valores de la similitud, los valores más allá de este límite pueden ser incluso mayores que el máximo identificado y, por lo tanto, el máximo identificado puede no representar un máximo local verdadero. En otras palabras, es bueno saber si un máximo identificado está en el límite con el fin de reaccionar de manera adecuada. Una reacción, por ejemplo, podría ser elegir un máximo local verdadero dentro de la secuencia de valores de la similitud, puesto que la ubicación del máximo identificado previamente puede no representar un valor del desfase de la altura del sonido válido.

De acuerdo con una realización preferida de la invención, el aparato se configura para considerar de manera selectiva uno o más de otros valores de la similitud más allá del límite de la secuencia de valores de la similitud, por ejemplo, más allá de un intervalo de búsqueda inicial, si la información sobre una característica del máximo identificado indica que el máximo identificado se localiza en el límite de la secuencia de valores de la similitud. El tener la oportunidad de considerar uno o más de otros valores de la similitud más allá del límite de la secuencia de valores de la similitud ayuda a asegurar que se obtenga una información de la altura del sonido exacta y válida.

De acuerdo con una realización preferida de la invención, el aparato se configura para determinar una información de la altura del sonido en una búsqueda de bucle abierto o en una búsqueda de bucle cerrado. La realización descrita es útil para utilizarse en los codificadores de la señal de audio que se configuran para tener una determinación de la información de la altura del sonido de dos etapas, por ejemplo, una búsqueda de bucle abierto y una búsqueda de bucle cerrado.

Una realización de la invención proporciona un método para determinar una información de la altura del sonido basándose en una señal de audio según la reivindicación 18.

Una realización preferida adicional de la invención es un programa informático con un código del programa para

realizar el método cuando el programa informático se ejecuta en un ordenador o un microcontrolador. El programa descrito es especialmente adecuado para el empleo en dispositivos móviles, por ejemplo, teléfonos móviles.

Las realizaciones preferidas adicionales de acuerdo con la invención describen una búsqueda robusta de la altura del sonido con un tamaño de correlación adaptativo.

Breve descripción de las figuras

En lo siguiente, las realizaciones de la presente invención se explicarán con referencia a los dibujos acompañantes, en los cuales:

La figura 1 muestra un diagrama de flujo de un aparato de acuerdo con una realización de la invención;

la figura 2 muestra un diagrama de flujo de un aparato de acuerdo con una realización de la invención;

la figura 3 muestra una gráfica de acuerdo con una realización de la invención;

la figura 4 muestra una gráfica de acuerdo con una realización de la invención;

la figura 5 muestra una gráfica de acuerdo con una realización de la invención;

la figura 6 muestra un esquema de una señal; y

la figura 7 muestra un diagrama de flujo de un método de acuerdo con una realización de la invención.

Descripción detallada de las realizaciones

La figura 1 representa un diagrama de flujo de un aparato 100 de acuerdo con una realización de la invención, para la determinación de una información de la altura del sonido 160. El aparato 100 utiliza como entradas una señal de audio 110, por ejemplo, una señal del habla, y un valor del desplazamiento temporal 120. Basándose en el desplazamiento temporal 120, el aparato 100 elige una longitud de una porción de la señal (por ejemplo, utilizando un bloque 140) y proporciona una información 140a que describe una longitud de las porciones de la señal para la determinación 135 de un par de porciones utilizado para obtener 130 un valor de la similitud 130a (por ejemplo, en el bloque o el conseguidor del valor de la similitud 130). Basándose en el valor de la similitud 130a, la información de la altura del sonido 160 puede obtenerse en una determinación de la altura del sonido opcional (por ejemplo, en el bloque o determinador de la altura del sonido 150). La longitud 140a de la porción de la señal se determina como linealmente dependiente del desplazamiento temporal 120. La longitud 140a proporcionada de las porciones de la señal se utiliza para determinar 135 un par de porciones de la señal de audio 110, en las que la longitud 140a de este par de porciones de la señal se basa de manera flexible en el desplazamiento temporal 120. Así, un valor de la similitud 130a obtenido basándose en el par de porciones proporciona un valor de la similitud 130a confiable para la determinación de una frecuencia de la altura del sonido. Por ejemplo, si se considera un periodo largo de la altura del sonido, que corresponde a un desplazamiento temporal 120 grande, la longitud 140a elegida de las porciones de la señal serán grandes de manera correspondiente, con el fin de poder captar un ciclo completo de la altura del sonido considerada. El aparato descrito ofrece, por lo tanto, una base para una determinación de la altura del sonido confiable, exacta, no compleja y flexible. Además, deberá observarse que el aparato 100 de acuerdo con la figura 1 puede suplementarse con cualquiera de las características y funcionalidades descritas en el presente documento, ya sea de manera individual o en combinación.

La figura 2 muestra un diagrama de flujo de un aparato 200 de acuerdo con una realización de la invención. El aparato 200 toma como una entrada una señal de audio 210 y un valor del desplazamiento temporal 220 y entrega como salida una información de la altura del sonido 260. De acuerdo con el desplazamiento temporal 220, se determina la longitud 240a de las porciones de la señal (en el bloque 240). La longitud 240a determinada de las porciones de la señal se proporciona para la determinación 235 de un par de porciones que, además, se basan en el desplazamiento temporal dado 220 y la señal de audio 210. Basándose en el par determinado de porciones, se obtiene un valor de la similitud 230a (en el bloque 230).

En una etapa opcional adicional (bloque 251), el valor de la similitud 230a se normaliza 251 basándose en los valores de la energía del par determinado de porciones, entregando así un valor normalizado de la similitud 251a. Basándose en el valor de la similitud 230a o en el valor normalizado de la similitud 251a, puede obtenerse una secuencia 252a de valores de la similitud 252 en una etapa opcional (bloque 252). La secuencia 252a obtenida de valores de la similitud se obtiene para un desplazamiento temporal más corto 252b hasta un desplazamiento temporal más largo 252c. Así, el bloque 252 puede proporcionar, por ejemplo, la información del desplazamiento

temporal 220 dentro del intervalo dado (de un desplazamiento temporal más corto 252b hasta un desplazamiento temporal más largo 252c).

En una etapa opcional adicional (bloque 253), la secuencia 252a de valores de la similitud se somete a una función de ventanas 253. De este modo, se obtiene una secuencia con ventanas 253a de los valores de la similitud, en la que la función de ventanas 253 puede mejorar la exactitud de la información de la altura del sonido 260 a ser determinada, enfatizando o desenfatizando ciertos intervalos de la secuencia 252a de valores de la similitud.

Además, la secuencia 252a de valores de la similitud o la secuencia con ventanas 253a de valores de la similitud puede utilizarse en una búsqueda del máximo óptimo 254, para obtener una información de la ubicación del máximo 254a.

Basándose en una información de la ubicación del máximo 254a, en una etapa opcional adicional, se realiza una verificación de una característica de la información de la ubicación del máximo 254a (en el bloque 255). La verificación de la característica de la ubicación del máximo identificado 255 se basa en la información 254a de la ubicación del máximo, el desplazamiento temporal más corto considerado 252b y el desplazamiento temporal más largo considerado 252c. Si la característica del máximo indica que el máximo coincide con el desplazamiento temporal más corto 252b o el desplazamiento temporal más largo 252c, se toma una decisión de que debe considerarse un nuevo valor máximo. El valor máximo a ser considerado puede encontrarse en un intervalo del desplazamiento temporal más corto 252b al desplazamiento temporal más largo 252c, o más allá del desplazamiento temporal más corto 252b o el desplazamiento temporal más largo 252c. Si el nuevo máximo se elige de entre el desplazamiento temporal más corto 252b y el desplazamiento más largo 252c, se elegirá un nuevo máximo local entre los dos valores y se proporcionará como el nuevo máximo local 255a. De manera alternativa, un nuevo valor máximo puede buscarse más allá del desplazamiento temporal más corto 252b o el desplazamiento temporal más largo 252c, y si se encuentra un nuevo valor máximo, la ubicación correspondiente o una información 255a para una ubicación correspondiente se proporcionará. En una etapa opcional final, se realiza una estimación de la frecuencia de la altura del sonido (en el bloque 250).

La señal de audio 210 puede proporcionarse en una versión diezmada, reduciendo así la complejidad del cálculo. Esto se debe al hecho de que una señal diezmada muestra típicamente un velocidad de muestreo reducida y, por lo tanto, exhibe menos muestras por segundo. Esto a su vez, conduce a una complejidad menor del cálculo, puesto que para un intervalo de tiempo equivalente, menos valores de muestra necesitan considerarse que para una señal sobremuestreada o, de manera equivalente, para una señal con una velocidad de muestreo más alta. Por lo tanto, en una primera etapa (no mostrada) la señal de audio 210 puede diezmarse a una frecuencia de muestreo, que varía, por ejemplo, entre 5,3 y 8 kHz, dependiendo de la velocidad de muestreo de entrada.

En lo siguiente, se describirá cómo pueden determinarse la información de la longitud 240a de las porciones de la señal por el bloque 240. La figura 3 muestra una gráfica 300 de acuerdo con un aspecto de la invención. En el eje horizontal 310, se muestra el valor del desplazamiento temporal d. Un desplazamiento temporal más corto 310a y un desplazamiento temporal más largo 310b se indican en el eje horizontal, denominados *Pitmin* y *Pitmax*, respectivamente, que pueden corresponder al desplazamiento temporal más corto 252b y al desplazamiento temporal más largo 252b en la figura 2. En el eje vertical 320, se muestra la longitud de las porciones de la señal consideradas, en el que esta longitud puede representarse por la información de la longitud 140a o 240a. Una longitud mínima 320a y una longitud máxima 320b se indican en el eje vertical, denominadas *startlen* y *stoplen*, respectivamente. La línea 330 ilustra un incremento lineal de la longitud de las porciones de la señal con el desplazamiento temporal que se incrementa. Además, el desplazamiento temporal más corto 310a se denomina como *Pitmin*, que corresponde al valor mínimo de la altura del sonido considerado y el desplazamiento temporal más largo 310b se denomina como *Pitmax*, que corresponde al valor máximo de la altura del sonido considerado. La gráfica 300 ilustra la elección de la longitud de las porciones de la señal utilizada para obtener el valor de la similitud, permitiendo un cálculo eficiente y confiable de la determinación de la altura del sonido.

Tomando como referencia a la figura 4, la búsqueda de una información de la ubicación del máximo 254a o 255a se ilustra como realizada, por ejemplo, en el bloque 254 o 255. La figura 4 muestra una gráfica 400 de acuerdo con un aspecto de la invención. En el eje horizontal 410, se muestra el desplazamiento temporal d, que puede ser el desplazamiento temporal 120 o 220. En el eje vertical 420 se muestran los valores del valor de la similitud, por ejemplo, los valores de autocorrelación, que pueden ser el valor de la similitud 130a, 230a o 251a obtenido en el bloque 130 o 230. Una curva 430 muestra una evolución ejemplar de los valores de la similitud, por ejemplo, la secuencia 252a de valores de la similitud, que dependen del desplazamiento temporal d. La curva 430 tiene un máximo local $R(T_0)$ entre las líneas punteadas verticalmente denominadas *Pitmin* y *Pitmax*. El valor a la izquierda del máximo local $R(T_0\text{-}1)$ es más pequeño que $R(T_0)$ y el valor a la derecha de $R(T_0)$, $R(T_0\text{+}1)$ es más pequeño que $R(T_0)$, de este modo, $R(T_0)$ puede caracterizarse como un máximo local verdadero. Además, las líneas punteadas verticalmente denominadas *Pitmin* y *Pitmax* ilustran el intervalo en el cual puede realizarse una búsqueda del máximo (por ejemplo, en el bloque 254) y para el cual se obtienen valores d de los valores de la similitud del

desplazamiento temporal para formar la secuencia 252a. La búsqueda del máximo puede, por ejemplo, ser la búsqueda del máximo como se indica en el bloque 254 en el aparato 200. Además, se identifica un máximo, que corresponde con la línea punteada verticalmente denominada *Pitmin*. Sin embargo, este máximo identificado no es un máximo local verdadero, puesto que un máximo local más alto está disponible fuera del intervalo de búsqueda. Por lo tanto, el máximo que coincide con *Pitmin*, *R*(*Pitmin*) es un máximo falso. Tomando como referencia la figura 2, la curva 430 descrita puede representar la secuencia 252a en la cual se realiza una búsqueda en el bloque 254. La búsqueda 254 puede identificar el valor *R*(*Pitmin*) como el máximo y, por lo tanto, regresa a *Pitmin* como la información de la ubicación del máximo 254a. La información obtenida de la ubicación del máximo 254a puede utilizarse en la verificación 255 de la característica del máximo. La verificación 255 puede identificar la información de la ubicación del máximo 254 para indicar que el máximo se localiza en el límite del intervalo de búsqueda. En respuesta a este hallazgo, en una implementación, la verificación (bloque 255) puede descartar el máximo en *Pitmin* y en su lugar elegir un máximo local verdadero dentro del intervalo de búsqueda que corresponde a $R(T_0)$. Esto da como resultado una información de la ubicación del máximo 255a está caracterizada por T_0 en lugar de *Pitmin*.

En lo siguiente, una implementación alternativa de la verificación (bloque 255) se describirá tomando como referencia la figura 5. La figura 5 muestra una gráfica 500 de acuerdo con un aspecto de la invención. En el eje horizontal 510, se muestra el valor del desplazamiento temporal. Además, en el eje vertical 520, se muestra el valor de la similitud que depende del desplazamiento temporal. Además, una curva 530 se traza en la gráfica 500, que ilustra, por ejemplo, los valores de la similitud, por ejemplo, 130a, 230a o 251a. La curva 530 es similar a la curva 430 en la figura 4 y muestra un procedimiento alternativo si la verificación 255 encuentra que una información de la ubicación del máximo 254a indica que un máximo se localiza en el límite del intervalo de búsqueda. La gráfica 500 muestra un valor máximo de la curva 530 en la intersección con la línea punteada verticalmente denominada *Pitmin* con respecto a los valores a la derecha de la misma, como se ilustra ya en la gráfica 400 de la figura 4 (*R*(*Pitmin*) es un máximo entre *d*=*Pitmin* y *d*=*Pitmax*). De manera alterna, para el procedimiento descrito en la figura 4, el intervalo de búsqueda se extiende más allá de *Pitmin* para la verificación 255 si el máximo encontrado *R*(*Pitmin*) es verdaderamente un máximo local (con valores más pequeños en ambos lados). Aunque la búsqueda más allá de *Pitmin* encuentra un nuevo máximo local *R*(*Pitmin*-2), que a su vez será regresado como una información (nueva, revisada) de la ubicación del máximo 255a. Los valores adicionales de la similitud más allá del valor de la similitud *R*(*Pitmin*) pueden por ejemplo, estar disponibles, debido al hecho de que esta búsqueda adicional se realiza en una versión sobremuestreada de la curva 430 de la figura 4. Por lo tanto, ningún nuevo cálculo puede ser necesario para la recuperación de los valores más allá de *R*(*Pitmin*), excepto para un sobremuestreado de la secuencia de valores de la similitud empleada previamente.

La figura 6 muestra una gráfica ilustrativa de una señal de audio, por ejemplo, de la señal de audio 110 y 210. La señal tiene un seccionamiento por cuadros y se muestran tres cuadros. Dos flechas indican el desplazamiento temporal más corto *Pitmin* y el desplazamiento temporal más largo *Pitmax*, y la flecha marcada como ventana del desfase indica la variabilidad de la ventana del desfase a la escala entre los valores *Pitmin* y *Pitmax*.

La figura 7 ilustra un diagrama de flujo 700 de un método de acuerdo con un aspecto de la invención. En una primera etapa, se determina la longitud de las porciones de la señal 710, en las que la longitud es linealmente dependiente del desplazamiento temporal considerado. Posteriormente, basándose en la longitud determinada, se determinan un par de porciones de la señal 720. Además, basándose en el par determinado de las porciones de la señal, se obtienen los valores de la similitud 730. Opcionalmente, en una etapa final basada en el valor determinado de la similitud, se determina una información de la altura del sonido 740.

El método 700 puede suplementarse con cualquiera de las características y funcionalidades descritas en el presente documento, también con respecto al aparato.

Aspectos adicionales y conclusión

En lo siguiente, se tratan algunos aspectos y opiniones de acuerdo con la presente invención.

Un aspecto de acuerdo con la invención es encontrar la frecuencia fundamental, es decir el valor de la altura del sonido (también llamado el valor del desfase en el dominio del tiempo), en una señal del habla, utilizando el método de autocorrelación. En el codificador del habla códec AMR-WB [1], la búsqueda de la altura del sonido se divide en una búsqueda de la altura del sonido de bucle abierto y de bucle cerrado. La búsqueda de la altura del sonido de bucle abierto es un proceso de estimar el desfase casi óptimo de la entrada del habla ponderada. Dependiendo del modo, se realiza un análisis de la altura del sonido de bucle abierto una vez por trama (cada 20 ms) o dos veces por trama (cada 10 ms) para encontrar dos estimados del desfase de la altura del sonido en cada trama. Esto se hace con el fin de simplificar el análisis de la altura del sonido y confinar la búsqueda de la altura del sonido de bucle cerrado a un pequeño número de desfases alrededor de los desfases estimados de bucle abierto. En algunas realizaciones, tal procedimiento puede utilizarse opcionalmente.

El intervalo de búsqueda se ajusta al tracto vocal humano. Por lo tanto, el algoritmo de búsqueda de la atura, por ejemplo, de AMR-WB, se restringe para buscar solo entre el valor mínimo de la altura del sonido de 55 Hz y el valor máximo de la altura del sonido de 380 Hz. El códec AMR-WB [1] está utilizando un tamaño fijo de la ventana de búsqueda para la autocorrelación. Se ha encontrado que este tamaño fijo de la ventana de búsqueda no es óptimo: algunas veces la ventana de correlación para la estimación del desfase de la altura del sonido puede fallar en contener un ciclo de la altura del sonido completo, haciendo así difícil o no significativa la correlación; si la ventana es demasiado grande, puede causar problemas de complejidad y también incrementar la dificultad para detectar un desfase corto de la altura del sonido. También se ha encontrado que una ventana sobredimensionada costaría mucha complejidad adicional. El códec VMR-WB [2] y EVS [3] están utilizando, respectivamente tres y hasta cuatro diferentes longitudes para la ventana de autocorrelación, divididas en cuatro secciones: [10, 16], [17, 31], [32, 61] y [62, 115], en donde el intervalo de la altura del sonido es de 10 a 115. Se ha encontrado que una desventaja principal es que los valores de la altura del sonido dentro de una sección están utilizando el mismo tamaño de autocorrelación y, por lo tanto, no se tratan igualmente, lo que puede conducir a valores de la altura del sonido erróneos. Por ejemplo, los valores de la altura del sonido de 62 y 115 están utilizando la misma longitud de la autocorrelación de 115. En algunos códecs, los valores de la altura del sonido de las últimas tramas se toman en cuenta. Sin embargo, el conocimiento previo sobre el último valor de la altura del sonido no siempre está disponible, por ejemplo, en los códecs que operan en el dominio de la frecuencia en donde no se necesitan valores de la altura del sonido para el procesamiento normal, como AAC-ELD [4].

En lo siguiente, varios aspectos de la presente invención se discuten adicionalmente.

Un aspecto de la invención presenta un enfoque con una baja complejidad y una búsqueda robusta de la altura del sonido utilizando un tamaño de autocorrelación adaptativo de la altura del sonido con precisión de enteros. No necesita ningún conocimiento previo de la señal, como los valores previos de la altura del sonido. Tal enfoque puede, por ejemplo, implementarse utilizando la selección de la longitud de las porciones de la señal como se realiza por los bloques 140,240. Por razones de complejidad, la búsqueda de la altura del sonido puede separarse en dos etapas similares a la búsqueda de la altura del sonido en el códec AMR-WB [1].

En el códec AMR-WB [1], el intervalo de búsqueda para la búsqueda de la altura del sonido se adapta en el tracto vocal humano. Por lo tanto, se observan los valores de la altura del sonido de 55 Hz a 376 Hz a la velocidad de muestreo de 12,8 kHz. Basándose en esto, los límites de *Pitmax* = 872 muestras y *Pitmin* = 126 muestras para una velocidad de muestreo de 48 kHz se utilizarán en un enfoque de acuerdo con un aspecto de la invención. Esto corresponde a los valores de la altura del sonido de 55 Hz a 380 Hz.

De acuerdo con un aspecto adicional de la invención, en una primera etapa, la señal, por ejemplo, la señal 110 o 210, se submuestrea como en el códec AMR-WB [1], por ejemplo, en una etapa no mostrada de los aparatos 100 y 200. Pero en lugar de diezmar la señal a una frecuencia de muestreo fija de 6,4 kHz, la señal (por ejemplo, la señal 110 o 210) se diezma a una frecuencia de muestreo que varía entre 5,3 y 8 kHz, dependiendo de la velocidad de muestreo de entrada. El factor de diezmado *decim* se elige como:

$$decim = \begin{cases} 2, & fs \leq 16\,kHz \\ 3, & fs \leq 24\,kHz \\ 4, & fs \leq 32\,kHz \\ 6, & fs > 32\,kHz \end{cases}$$

en donde *fs* es la velocidad de muestreo de entrada. Se hace un submuestreo vía un filtro FIR con las derivaciones siendo

[0,0101, 0,2203, 0,5391, 0,2203, 0,0101] para *decim*=2,

[0,0068, 0,0664, 0,2465, 0,3608, 0,2465, 0,0664, 0,0068] para *decim*=3,

[0,0051, 0,0294, 0,1107, 0,2193, 0,2710, 0,2193, 0,1107, 0,0294, 0,0051] para *decim*=4

y

[0,0034, 0,0106, 0,0333, 0,0739, 0,1236, 0,1648, 0,1809, 0,1648, 0,1236, 0,0739, 0,0333, 0,0106, 0,0034] para *decim*=6 (por ejemplo, con el fin de evitar el solapamiento).

De acuerdo con un aspecto de la invención, una búsqueda de la altura del sonido puede hacerse en la versión

submuestreada (por ejemplo, en la señal 110, 210) vía el método de autocorrelación en un bucle iterativo (por ejemplo, controlado por el bloque 252) del desfase mínimo $pitmin = \frac{Pitmin}{decim}$ al valor del desfase máximo $pitmax = \frac{Pitmax}{decim}$ con el tamaño de autocorrelación (representado, por ejemplo, por la información de la longitud 240a) que va de 5 ms a 10 ms con precisión de enteros.

En algunos algoritmos, existe la posibilidad de que el máximo de la función de autocorrelación corresponda a un múltiplo o submúltiplo del desfase de la altura del sonido d y que el desfase de la altura del sonido estimado, por lo tanto, no sea correcto. El documento EP0628947 [5] trata este problema aplicando una función de ponderación *w(d)* a la función de autocorrelación R:

$$R(d) = R(d) \cdot w(d), \qquad \mathrm{d} = \mathrm{pitmin} \ldots pitmax$$

en donde la función de ponderación tiene la siguiente forma: $w(d) = i^{\log_2 K}$. *K* es un parámetro de afinación que se establece a un valor suficientemente bajo para reducir la probabilidad de obtener un máximo para R(d) a un múltiplo del desfase de la altura del sonido pero, al mismo tiempo, suficientemente alto para excluir los submúltiplos del desfase de la altura del sonido. De manera similar al códec AMR-WB [1], este enfoque utiliza la función de ponderación utilizada con *K* = 0,7. La ponderación descrita puede ser la función de ventana como la realizada en el bloque 253.

En algunos algoritmos, como en el códec AMR-WB [1], el máximo valor de autocorrelación se normaliza finalmente, esto permite comparar este máximo a través de las señales o contra un valor umbral. Sin embargo, de acuerdo con un aspecto de la invención, para incrementar la robustez de la búsqueda de la altura del sonido, al hacer la autocorrelación libre de fluctuaciones de energía en la señal, los valores de autocorrelación se normalizan, por ejemplo, en el bloque 251, antes de que se realice la maximización (o búsqueda del máximo), como sigue:

$$R(d) = \frac{R'(d) \cdot w(d)}{\sqrt{norm(0) \cdot norm(d)}}$$

en donde R(d) es el valor de autocorrelación normalizado entre la señal sin desplazamiento y la señal desplazada a la izquierda por *d* muestras, R'(d) es el valor de autocorrelación entre la señal sin desplazamiento y la señal desplazada a la izquierda por *d* muestras, *w*(*d*) es el factor de ponderación de *d*, *norm*(0) es el producto escalar de la parte de la señal sin desplazamiento (por ejemplo, de la primera porción del par de porciones) y *norm*(*d*) es el producto escalar de la parte de la señal desplazada a la izquierda por *d* muestras (por ejemplo, de la segunda porción del par de porciones). (Por ejemplo, R(d) puede corresponder al valor normalizado de la similitud 251a y *R'*(*d*) puede corresponder al valor de la similitud 230a o 130a)

De acuerdo con un aspecto adicional de la invención, para ahorrar complejidad, los valores de normalización *norm*(0) y *norm*(*d*), que pueden utilizarse para la normalización y estimarse en el bloque 251, se calculan con un mecanismo de actualización. Así, *norm*(*d*) puede calcularse como:

$$norm(d) = norm(d-1) + x_d^2 - x_{d+len(d)}^2$$

en donde x_d es la muestra de la señal desplazada a la izquierda por *d* muestras con la ventana de búsqueda de longitud *len(d)*. Solo para los valores iniciales de *norm*(0) y *norm*(*pitmin*), los productos escalares completos tienen que calcularse con *len*(*pitmin*). Si la longitud de la ventana de búsqueda cambia de *d*-1 a *d*, el valor de normalización necesita una actualización adicional de *len*(*d*-1)-*len*(*d*) valores.

De acuerdo con otro aspecto de la invención, otra diferencia mayor para algunos algoritmos de búsqueda de la altura del sonido basados en el método de autocorrelación es que este enfoque solo elige valores de la altura del sonido, que representan un máximo local real, por ejemplo, realizado en el bloque 255. Así, pueden evitarse los resultados falsos de la altura del sonido, que ocurren si un máximo de la autocorrelación está fuera del intervalo de búsqueda (por ejemplo, consultar el ejemplo descrito con respecto a las figuras 4 y 5). Esto significa que el valor del desfase de d se utiliza solo si:

$$R(d-1) \leq R(d) \geq R(d+1).$$

Como se hizo en el códec AMR-WB [1], una segunda etapa de la búsqueda de la altura del sonido (por ejemplo, bucle cerrado) está operando en el dominio de la señal muestreada original y solo utiliza un número pequeño de desfases alrededor del desfase estimado de bucle abierto sobremuestreado T_0. La búsqueda de la altura del sonido, por ejemplo, la búsqueda del máximo en 254, también utiliza una longitud de la ventana de búsqueda *Len* (que puede ser una longitud constante de la ventana de búsqueda en algunas realizaciones), pero es dependiente ahora de T_0, como sigue:

$$Len = m \cdot T_0 + startlen - Pitmin \cdot m$$

en donde

$$m = \frac{(stoplen - startlen)}{Pitmax - Pitmin}$$

y *startlen* = 5 ms y *stoplen* = 10 ms.

De acuerdo con un aspecto adicional de la invención, el intervalo de búsqueda, por ejemplo, en la búsqueda del máximo 254, está limitado por

$$\left[\max\left(Pitmin, T_0 - \frac{\delta}{2}\right), \min(Pitmax, T_0 + \frac{\delta}{2})\right]$$

en donde $\delta = 4 \cdot decim$.

De acuerdo con un aspecto de la invención, el algoritmo elige el valor del desfase T que pertenece al valor máximo de autocorrelación normalizado.

De acuerdo con otro aspecto de la invención, una mejora del método propuesto es que la búsqueda de la altura del sonido en el límite de la búsqueda se maneje con cuidado, como se describió con respecto al bloque 255 y con respecto a las figuras 4 y 5. Si el valor del desfase de *Pitmin* o *Pitmax* se elige en algún método, el algoritmo corre el riesgo de utilizar un valor del desfase falso cuando el máximo real está fuera del intervalo de búsqueda. Esto puede ocurrir incluso con una búsqueda de la altura del sonido como se describió anteriormente, debido a que la búsqueda de la altura del sonido de bucle abierto y de bucle cerrado están trabajando en diferentes resoluciones de la señal debido al submuestreo de la búsqueda de la altura del sonido de bucle abierto. Por lo tanto, este enfoque extiende la búsqueda por un máximo de, por ejemplo, cuatro muestras por encima del límite correspondiente (en el bloque 255). La búsqueda de la altura del sonido se detiene y utiliza el valor del desfase correspondiente, si un primer máximo real de la autocorrelación normalizada se encuentra fuera del intervalo de búsqueda de [*Pitmin Pitmax*]. De otra manera, se selecciona *Pitmin* -4 o *Pitmax* +4.

Aunque algunos aspectos se han descrito en el contexto de un aparato, está claro que estos aspectos también representan una descripción del método correspondiente, en donde un bloque o dispositivo corresponde a una etapa del método o a una característica de una etapa del método. De manera análoga, los aspectos descritos en el contexto de una etapa del método también representan una descripción de un bloque o punto o característica correspondiente de un aparato correspondiente. Algunas o todas las etapas del método pueden ejecutarse mediante (o utilizando) un equipo, como por ejemplo, un microprocesador, un ordenador programable o un circuito electrónico. En algunas realizaciones, una o más de las etapas del método más importantes pueden ejecutarse por tal aparato.

Dependiendo de ciertos requisitos de implementación, las realizaciones de la invención pueden implementarse en hardware o en software. La implementación puede realizarse utilizando un medio de almacenamiento digital, por ejemplo, un disco flexible, un DVD, un Blu-Ray, un CD, una ROM, una PROM, una EPROM, una EEPROM o una memoria FLASH, que tenga señales de control legibles electrónicamente almacenadas en el mismo, que coopera (o que es capaz de cooperar) con un sistema informático programable, de manera que se ejecute el método respectivo. Por lo tanto, el medio de almacenamiento digital puede ser legible por ordenador.

Algunas realizaciones de acuerdo con la invención comprenden un portador de datos que tiene señales de control legibles electrónicamente, que es capaz de cooperar con un sistema informático programable, de manera que se

ejecute uno de los métodos descritos en el presente documento.

Generalmente, las realizaciones de la presente invención pueden implementarse como un producto de un programa informático con un código del programa, el código del programa es operativo para ejecutar uno de los métodos cuando el producto del programa informático se ejecuta en un ordenador. El código del programa puede, por ejemplo, almacenarse en un portador legible por la máquina.

Otras realizaciones comprenden el programa informático para ejecutar uno de los métodos descritos en el presente documento, almacenado en un portador legible por la máquina.

En otras palabras, una realización del método inventivo es, por lo tanto, un programa informático que tiene un código del programa para ejecutar uno de los métodos descritos en el presente documento, cuando el programa informático se ejecuta en un ordenador.

Una realización adicional de los métodos inventivos es, por lo tanto, un portador de datos (o un medio de almacenamiento digital o un medio legible por ordenador) que comprende, grabado en el mismo, el programa informático para ejecutar uno de los métodos descritos en el presente documento. El portador de datos, el medio de almacenamiento digital o el medio grabado son de manera típica tangibles y/o no transitorios.

Una realización adicional del método inventivo es, por lo tanto, un flujo de datos o una secuencia de señales que representa el programa informático para ejecutar uno de los métodos descritos en el presente documento. El flujo de datos o la secuencia de señales puede por ejemplo, configurarse para transferirse mediante una conexión de comunicación de datos, por ejemplo, mediante Internet.

Una realización adicional comprende un medio de procesamiento, por ejemplo, un ordenador, o un dispositivo lógico programable, configurado o adaptado para ejecutar uno de los métodos descritos en el presente documento.

Una realización adicional comprende un ordenador que tiene instalado en el mismo el programa informático para ejecutar uno de los métodos descritos en el presente documento.

Una realización adicional de acuerdo con la invención comprende un aparato o un sistema configurado para transferir (por ejemplo, de manera electrónica u óptica) un programa informático para ejecutar uno de los métodos descritos en el presente documento a un receptor. El receptor puede, por ejemplo, ser un ordenador, un dispositivo móvil, un dispositivo de memoria o lo similar. El aparato o sistema, por ejemplo, puede comprender un servidor de archivos para transferir el programa informático al receptor.

En algunas realizaciones, un dispositivo lógico programable (por ejemplo, una matriz de puertas programable por campos) puede utilizarse para ejecutar algunas o todas las funcionalidades de los métodos descritos en el presente documento. En algunas realizaciones, una matriz de puertas programable por campos puede cooperar con un microprocesador con el fin de ejecutar uno de los métodos descritos en el presente documento. Generalmente, los métodos se ejecutan de manera preferida, por algún aparato de hardware.

El aparato descrito en el presente documento puede implementarse utilizando un aparato, o utilizando un ordenador, o utilizando una combinación de un aparato y un ordenador.

El aparato descrito en el presente documento, o cualquier componente del aparato descrito en el presente documento, puede implementarse al menos parcialmente en hardware y/o en software.

Los métodos descritos en el presente documento pueden ejecutarse utilizando un aparato de hardware, o utilizando un ordenador, o utilizando una combinación de un aparato y un ordenador.

Los métodos descritos en el presente documento, o cualquier componente del aparato descrito en el presente documento, pueden ejecutarse al menos parcialmente por hardware y/o software.

Las realizaciones descritas anteriormente son simplemente ilustrativas para los principios de la presente invención. Se entiende que las modificaciones y variaciones de las disposiciones y los detalles descritos en el presente documento serán evidentes para otros expertos en la técnica. Es la intención, por lo tanto, de estar limitado solo por el alcance de las reivindicaciones inminentes de la patente y no por los detalles específicos presentados por medio de la descripción y la explicación de las realizaciones en el presente documento.

Referencias:

[1] 3GPP, TS 26.190, “Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB)

speech codec; Transcoding functions (Release 12)", 2014.

[2] 3GPP2, C.S0052-A, "Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB), Service Options 62 and 63 for Spread Spectrum Systems", Versión 1.0, abril de 2005

[3] 3GPP, TS 26.445, "Universal Mobile Telecommunitations System (UMTS); LTE; Codec for enhanced Voice Services (EVS); Detailed algorithmic description", versión 12.3.0, lanzamiento 12

[4] AAC-ELD Standard: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=46457

[5] EP0628947 "Method and device for speech signal pitch period estimation and classification in digital speech coders"

REIVINDICACIONES

1. Aparato para determinar una información de la altura del sonido (160; 260), basándose en una señal de audio (110; 210),

en el que el aparato se configura para obtener a valor de la similitud (130a; 230a, 251a) ($R(d)$; $R'(d)$) que está asociado con un par dado de porciones de la señal de audio que tiene un desplazamiento temporal dado (120; 220) (d);

en el que el aparato se configura para elegir una longitud (140a; 240a) (*Len(d)*) de las porciones de la señal de la señal de audio utilizadas para obtener el valor de la similitud ($R(d)$; $R'(d)$) para el desplazamiento temporal dado (d) que depende del desplazamiento temporal dado (d);

en el que el aparato se configura para elegir la longitud (*Len(d)*) de las porciones de la señal para ser linealmente dependiente del desplazamiento temporal dado (d), dentro de una tolerancia de ± 1 muestra; caracterizado por que el aparato se configura para elegir la longitud de las porciones de la señal basándose en

$$Len(d) = m \cdot d + startlen - Pitmin \cdot m,$$

en donde d es el desplazamiento temporal dado, *startlen* es una longitud mínima predeterminada para las porciones de la señal, *Pitmin* es un valor predeterminado del desfase de la altura del sonido considerado más pequeño y m es un factor mediante el cual el desplazamiento temporal dado se escala, y

en el que el aparato se configura para elegir la longitud de las porciones de la señal como un valor entero cercano a *Len(d)* basándose en una función de redondeo, una función de suelo, una función de techo o una función de truncamiento.

2. Aparato según la reivindicación 1, en el que el aparato se configura para obtener una información de la altura del sonido basándose en una secuencia de valores de la similitud (252a).

3. Aparato según la reivindicación 2, en el que el aparato se configura para obtener la secuencia de valores de la similitud basándose en los valores de la similitud para los desplazamientos temporales d en un intervalo que empieza entre 1 ms y 4 ms y que se extiende hasta los desplazamientos temporales entre 15 ms a 25 ms.

4. Aparato según una de las reivindicaciones 1 a 3, en el que el aparato se configura para incrementar paso a paso la longitud de las porciones de la señal en etapas de una muestra con desplazamiento temporal que se incrementa.

5. Aparato según una de las reivindicaciones 1 a 4, en el que el aparato se configura para incrementar la longitud de las porciones de la señal con precisión de enteros con desplazamiento temporal que se incrementa.

6. Aparato según una de las reivindicaciones 1 a 5, en el que el aparato se configura para incrementar la longitud de las porciones de la señal, entre una longitud mínima predeterminada (320a) y una longitud máxima predeterminada (320b), linealmente en dependencia con el desplazamiento temporal dado,

en el que la longitud mínima predeterminada se utiliza para un desplazamiento temporal más corto (252b) que corresponde a una frecuencia de la altura del sonido máxima, y

en el que la longitud máxima predeterminada se utiliza para un desplazamiento temporal más largo (252c) que corresponde a una frecuencia de la altura del sonido mínima.

7. Aparato según una de las reivindicaciones 1 a 6, en el que el aparato se configura para calcular un valor de autocorrelación (230a) ($R'(d)$) basándose en dos porciones de la señal desplazadas temporalmente de la señal de audio, desplazadas temporalmente por el desplazamiento temporal dado (d), con el fin de obtener el valor de la similitud,

en el que un número de valores de muestra de la señal de audio considerado en el cálculo del valor de autocorrelación se determina por la longitud elegida.

8. Aparato según la reivindicación 7, en el que el aparato se configura para obtener los valores de la similitud basándose en

$$R'(d) = \sum_{n=0}^{Len(d)} s(n)s(n-d),$$

en donde *s*(*n*) es una muestra de la señal de audio en el tiempo n, *Len(d)* es una información sobre la longitud de las porciones de la señal para el desplazamiento temporal dado d y d es el desplazamiento temporal dado.

9. Aparato de conformidad con una de las reivindicaciones 1 a 8, en el que el aparato se configura para obtener una información de la ubicación (254a) de un valor máximo de una pluralidad de valores de la similitud; y

en el que el aparato se configura para obtener una información de la altura del sonido basándose en la información de la ubicación del valor máximo.

10. Aparato según una de las reivindicaciones 1 a 9, en el que el aparato se configura para aplicar una normalización (251) al valor de la similitud (*R'*(*d*)) utilizando al menos dos valores de normalización (*norm*(0), *norm*(*d*));

un primer valor de normalización (*norm*(0)) que representa una característica estadística de una primera porción del par de porciones dadas, y

un segundo valor de normalización (*norm*(*d*)) que representa una característica estadística de una segunda porción del par de porciones dadas,

con el fin de derivar un valor normalizado de la similitud (251a) (*R*(*d*)).

11. Aparato según la reivindicación 10, en el que el aparato se configura para obtener un valor normalizado de la similitud *R*(*d*) basándose en

$$R(d) = \frac{R'(d)w(d)}{\sqrt{norm(0)norm(d)}},$$

en donde R'(d) es un valor de la similitud y *w*(*d*) es una función de ventana.

12. Aparato según una de las reivindicaciones 10 a 11, en el que el aparato se configura para derivar de manera recursiva un valor de normalización para un nuevo desplazamiento temporal d, de un valor de normalización para un desplazamiento temporal previo *d*-1, sumando uno o más valores de la energía de las muestras de la señal incluidas en una nueva porción de la señal y no incluidas en una porción anterior de la señal y sustrayendo uno o más valores de la energía de las muestras de la señal incluidas en la porción anterior de la señal y no incluidas en la nueva porción de la señal.

13. Aparato según una de las reivindicaciones 10 a 12, en el que el aparato se configura para obtener un valor de normalización *norm*(*d*) basándose en

$$norm(d) = norm(d-1) + x_d^2 - x_{d+Len(d)}^2,$$

en donde x_d es una muestra de la señal de audio contenida en la porción de la señal de acuerdo con el desplazamiento temporal *d*, pero no contenida en la porción de la señal de acuerdo con el desplazamiento temporal *d*-1, $x_{d+Len(d)}$ es una muestra de la señal de audio no contenida en la porción de la señal de acuerdo con el desplazamiento temporal d, pero contenida en la porción de la señal de acuerdo con el desplazamiento temporal *d*-1 de la señal de audio y *norm*(*d*-1) es un valor de normalización obtenido para una porción de la señal considerada previamente, de acuerdo con el desplazamiento temporal *d*-1.

14. Aparato según una de las reivindicaciones 1 a 13, en el que el aparato se configura para determinar una información sobre una característica (255a) de un máximo identificado de una secuencia de valores de la similitud (*R*(*d*); *R'*(*d*)) obtenida para diferentes desplazamientos temporales (*d*); y

en el que el aparato se configura para proporcionar una frecuencia de la altura del sonido (250) basándose en el máximo identificado si la información sobre la característica del máximo identificado indica que el máximo identificado es un máximo local; y

en el que el aparato se configura para proceder a considerar uno o más de otros valores de la similitud para estimar la frecuencia de la altura del sonido si la información sobre la característica del máximo no indica que el máximo es un máximo local.

15. Aparato según la reivindicación 14, en el que el aparato se configura para determinar si un máximo identificado se localiza en el límite de la secuencia de valores de la similitud como la información sobre una característica del máximo identificado.

16. Aparato según una de las reivindicaciones 14 a 15, en el que el aparato se configura para considerar de manera selectiva uno o más de otros valores de la similitud más allá del límite de la secuencia de valores de la similitud, si la información sobre una característica del máximo identificado indica que el máximo identificado se localiza en el límite de la secuencia de valores de la similitud.

17. Aparato según una de las reivindicaciones 1 a 16, en el que el aparato se configura para determinar una información de la altura del sonido en una búsqueda de bucle abierto o en una búsqueda de bucle cerrado.

18. Método para determinar una información de la altura del sonido, basándose en una señal de audio, que comprende:

obtener un valor de la similitud (*R*(*d*); *R'*(*d*)) que está asociado con un par de porciones dadas de la señal de audio que tiene un desplazamiento temporal dado (d);

elegir una longitud (*Len(d)*) de las porciones de la señal de la señal de audio utilizadas para obtener el valor de la similitud (*R*(*d*); *R'*(*d*)) para el desplazamiento temporal dado (d) que depende del desplazamiento temporal dado (d); y

en el que la longitud (*Len(d)*) de las porciones de la señal se elige para ser linealmente dependiente del desplazamiento temporal dado (d), dentro de una tolerancia de ± 1 muestra; caracterizado porque el método comprende elegir la longitud de las porciones de la señal basándose en

$$Len(d) = m \cdot d + startlen - Pitmin \cdot m,$$

en donde *d* es el desplazamiento temporal dado, *startlen* es una longitud mínima predeterminada para las porciones de la señal, *Pitmin* es un valor predeterminado del desfase de la altura del sonido considerado más pequeño y m es un factor mediante el cual el desplazamiento temporal dado se escala, y

en el que el método comprende elegir la longitud de las porciones de la señal como un valor entero cercano a *Len(d)* basándose en una función de redondeo, una función de suelo, una función de techo o una función de truncamiento

19. Un programa informático con un código del programa para ejecutar el método según la reivindicación 18, cuando el programa informático se ejecuta en un ordenador o un microcontrolador.

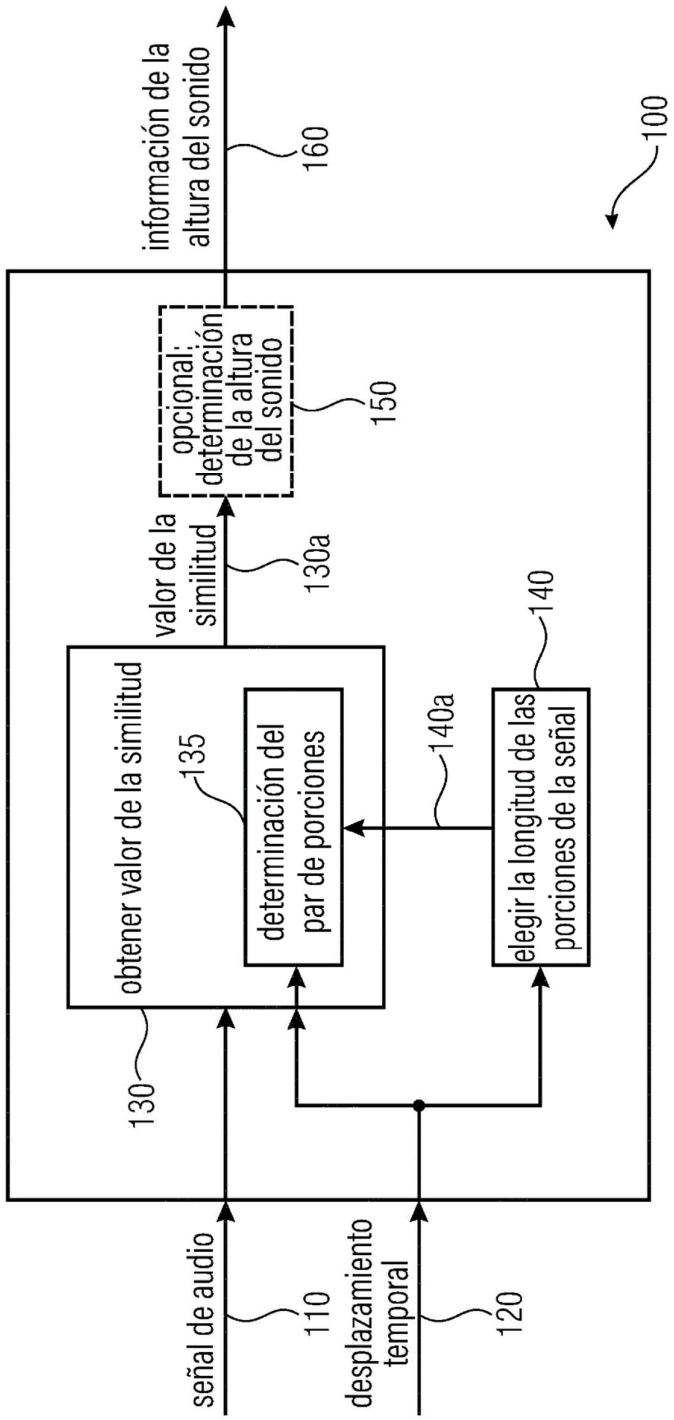
señal de audio
110
desplazamiento temporal
120
obtener valor de la similitud
130
determinación del par de porciones
135
elegir la longitud de las porciones de la señal
140
140a
valor de la similitud
130a
opcional: determinación de la altura del sonido
150
información de la altura del sonido
160
100

Fig. 1

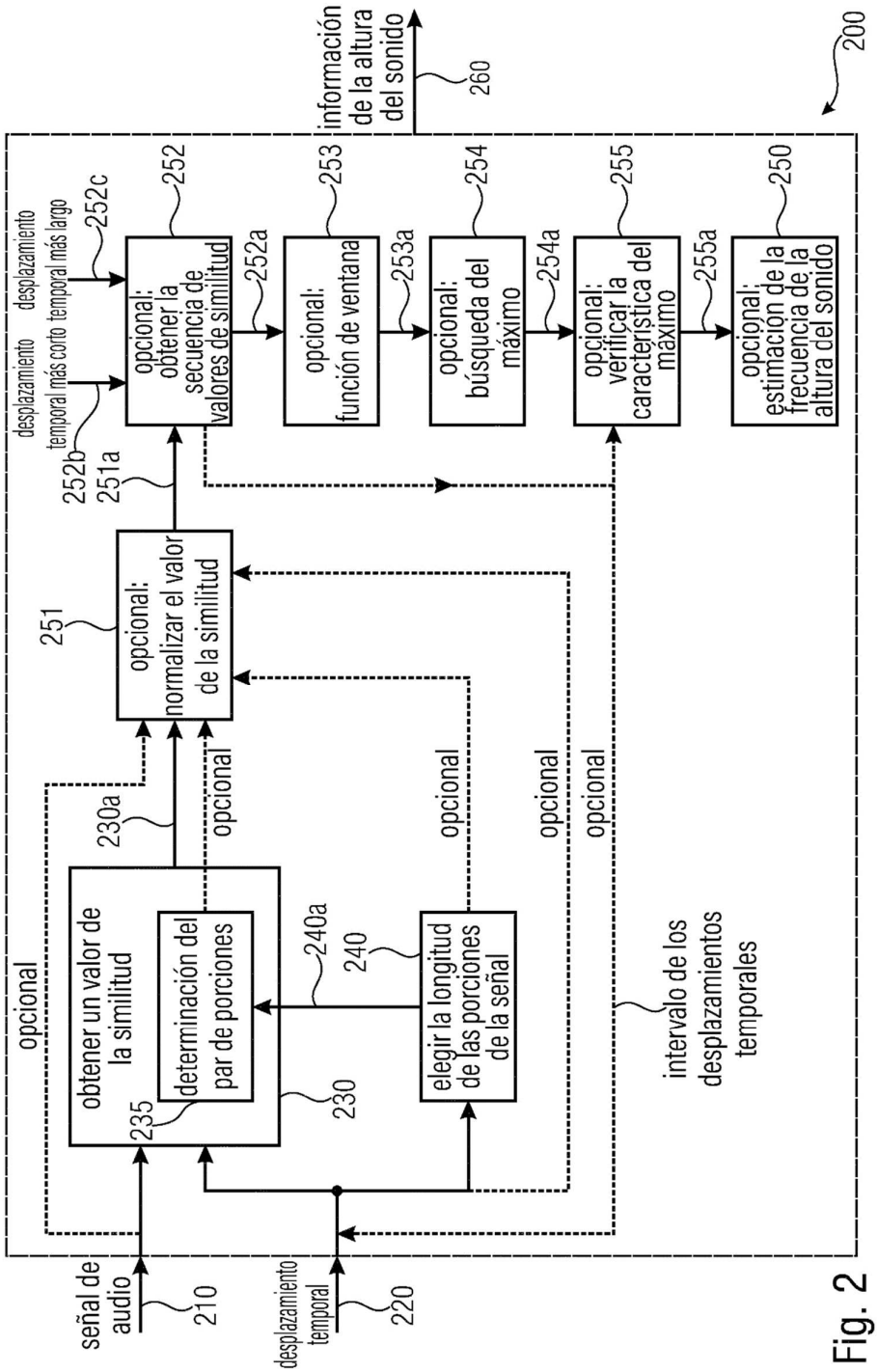
señal de audio
210
desplazamiento temporal
220
opcional
obtener un valor de la similitud
determinación del par de porciones
235
230
240a
240
elegir la longitud de las porciones de la señal
opcional
230a
opcional
opcional: normalizar el valor de la similitud
251
252b
251a
desplazamiento temporal más corto
desplazamiento temporal más largo
252c
opcional: obtener la secuencia de valores de similitud
252
252a
opcional: función de ventana
253
253a
opcional: búsqueda del máximo
254
254a
opcional: verificar la característica del máximo
255
255a
opcional: estimación de la frecuencia de la altura del sonido
250
información de la altura del sonido
260
200
opcional
opcional
intervalo de los desplazamientos temporales

Fig. 2

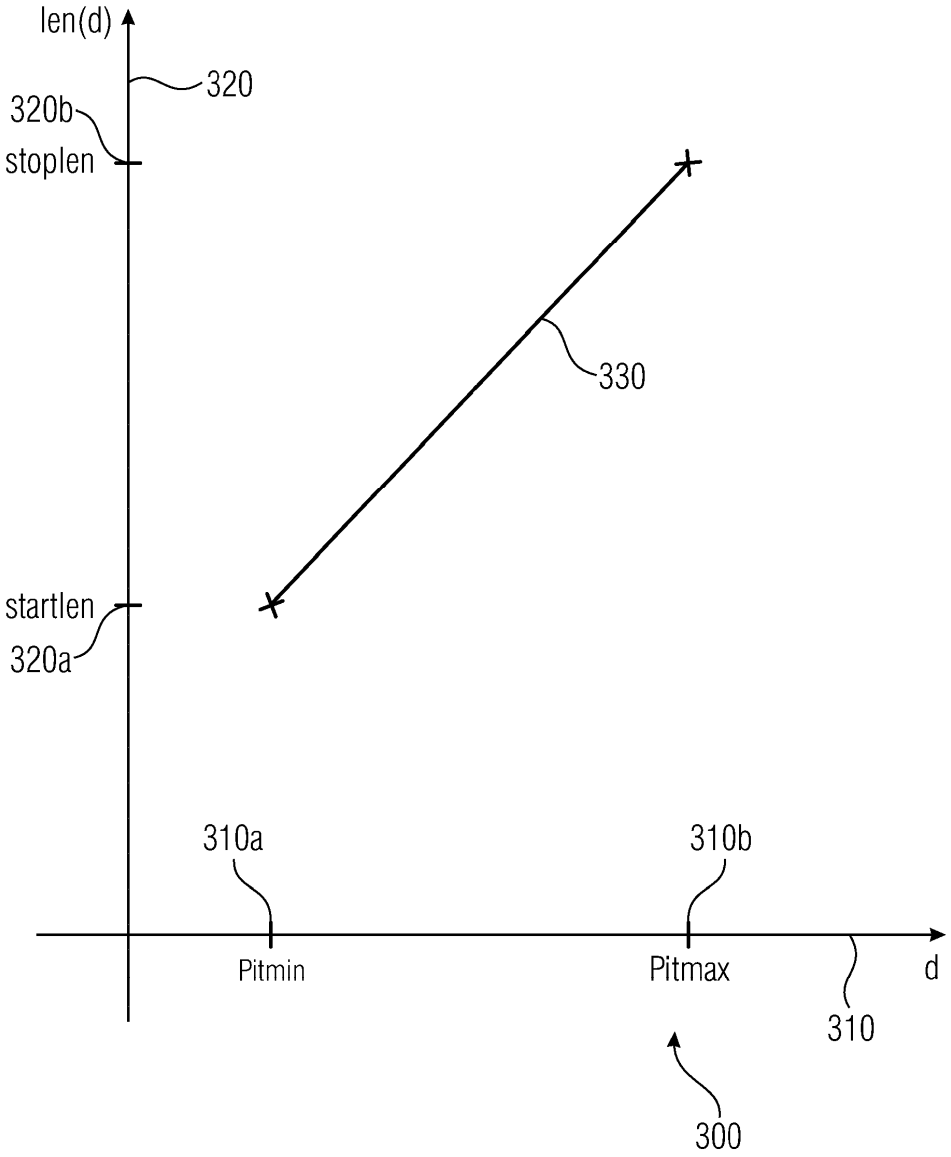
len(d)
320
320b
stoplen
330
startlen
320a
310a
310b
Pitmin
Pitmax
d
310
300

Fig. 3

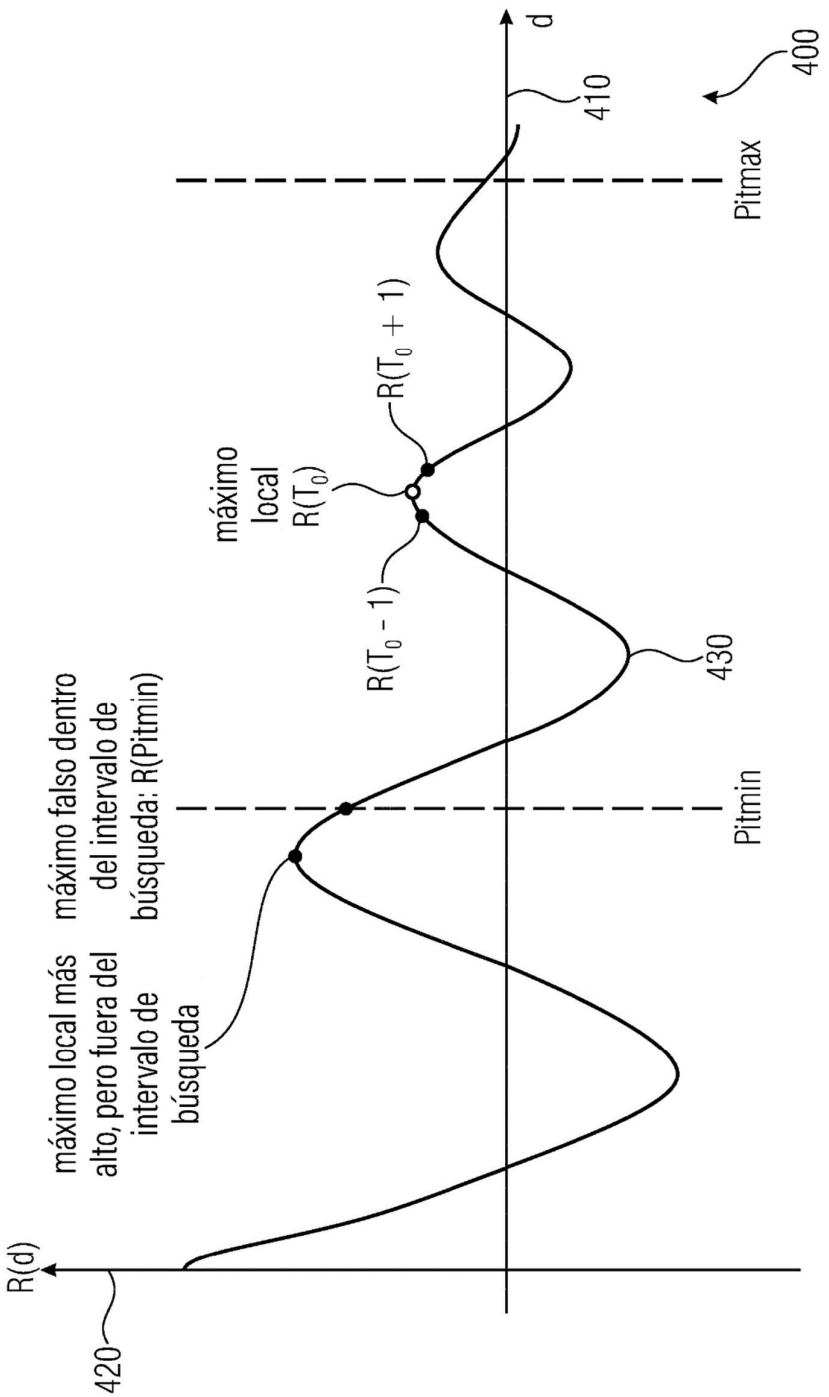
R(d)
d
420
410
400
máximo local más alto, pero fuera del intervalo de búsqueda
máximo falso dentro del intervalo de búsqueda: R(Pitmin)
máximo local R(T0)
R(T0 + 1)
R(T0 - 1)
Pitmax
Pitmin
430

Fig. 4

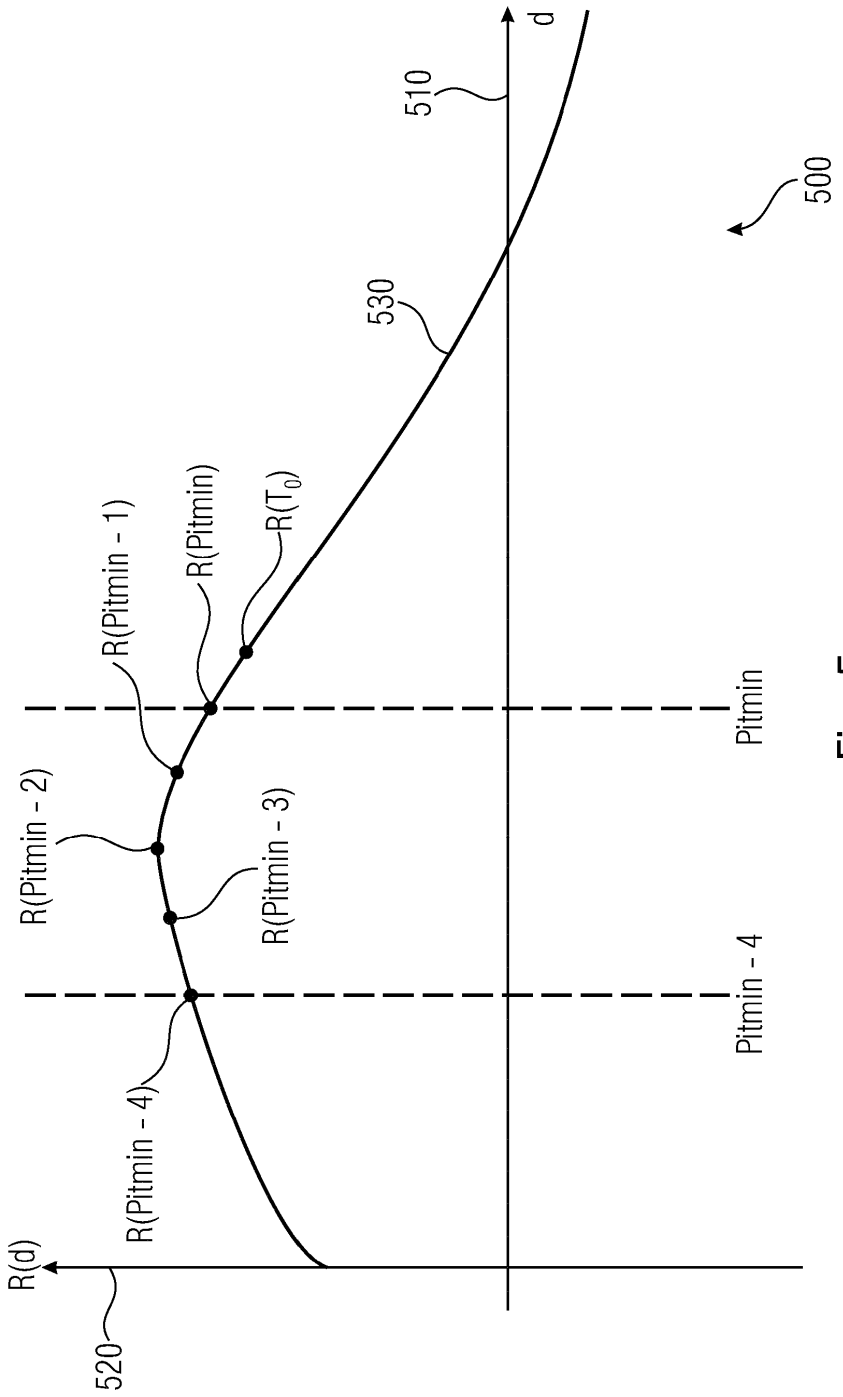
R(d)
d
520
510
500
530
R(Pitmin - 4)
R(Pitmin - 3)
R(Pitmin - 2)
R(Pitmin - 1)
R(Pitmin)
R(T0)
Pitmin - 4
Pitmin

Fig. 5

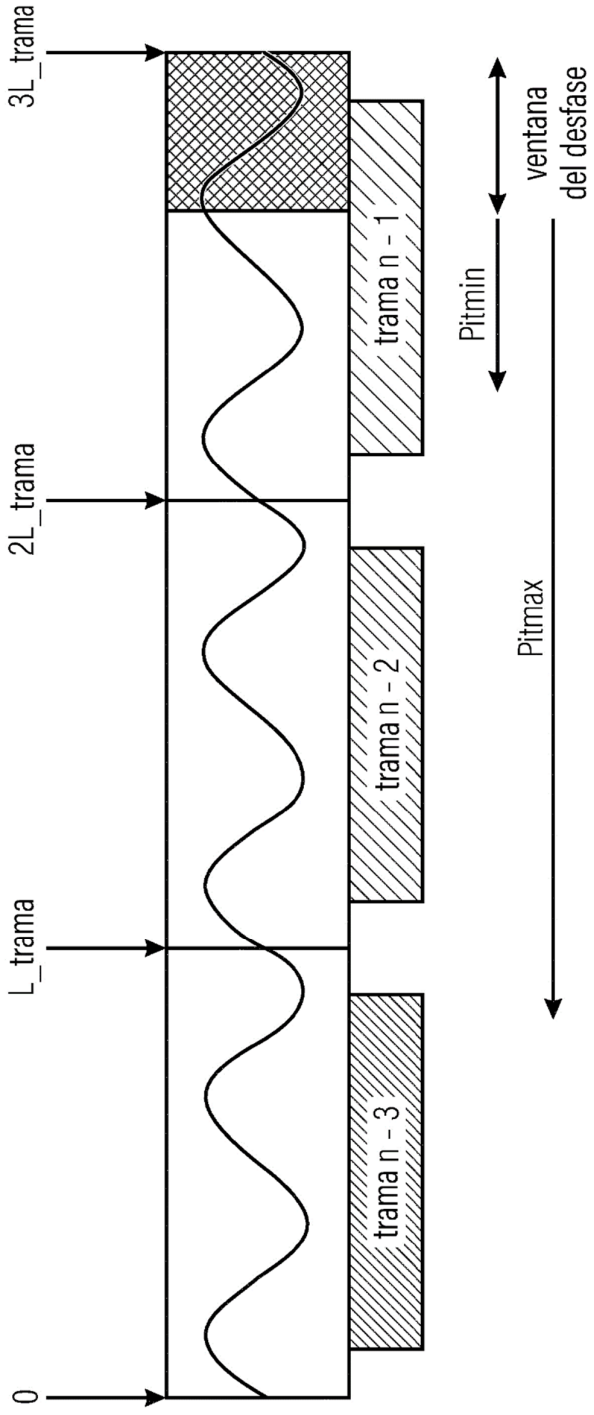
3L_trama
2L_trama
L_trama
0
trama n - 1
trama n - 2
trama n - 3
Pitmin
Pitmax
ventana
del desfase

Fig. 6

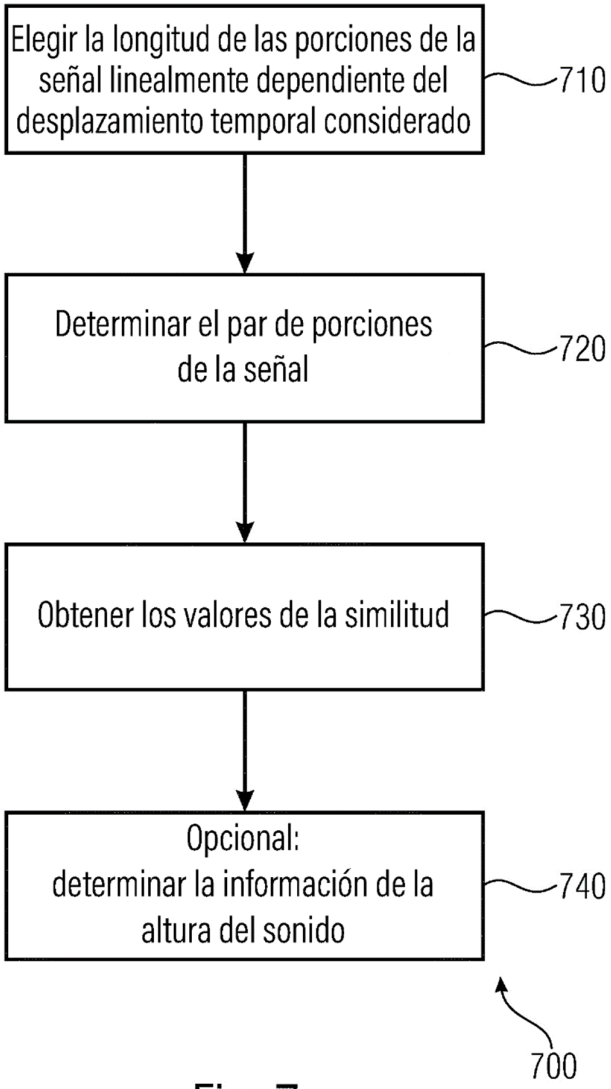
Elegir la longitud de las porciones de la señal linealmente dependiente del desplazamiento temporal considerado
710
Determinar el par de porciones de la señal
720
Obtener los valores de la similitud
730
Opcional: determinar la información de la altura del sonido
740
700

Fig. 7