(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 454 249**

(51) Int. CI.:

G10L 25/78 (2013.01)
G10L 15/14 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **07.10.2010 E 10768905 (1)**

(97) Fecha y número de publicación de la concesión europea: **11.12.2013 EP 2486562**

(54) Título: **Procedimiento de detección de segmentos de habla**

(30) Prioridad:

08.10.2009 ES 200930819

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
10.04.2014

(73) Titular/es:

TELEFÓNICA, S.A. (100.0%)
Gran Vía, 28
28013 Madrid, ES

(72) Inventor/es:

GARCÍA MARTÍNEZ, CARLOS;
DUXANS BARROBÉS, HELENCA;
SENDRA VICENS, MAURICIO y
CADENAS SÁNCHEZ, DAVID

(74) Agente/Representante:

CARPINTERO LÓPEZ, Mario

DESCRIPCIÓN

Procedimiento de detección de segmentos de habla.

Campo de la invención

La presente invención pertenece al área de la tecnología del habla, particularmente reconocimiento del habla y verificación del hablante, en concreto a la detección de habla y ruido.

Antecedentes de la invención

El reconocimiento automático del habla es una tarea particularmente complicada. Uno de los motivos es la dificultad de detectar los comienzos y finales de los segmentos de habla pronunciados por el usuario, discriminándolos adecuadamente de los periodos de silencio que se producen antes de que comience a hablar, después de que termine, y los que resultan de las pausas que dicho usuario realiza para respirar mientras habla.

La detección y delimitación de los segmentos de habla pronunciados es fundamental por dos motivos. En primer lugar, por motivos de eficiencia computacional: los algoritmos utilizados en reconocimiento del habla son bastante exigentes en lo que a carga computacional se refiere, por lo que aplicarlos a toda la señal acústica, sin eliminar los periodos en los que no está presente la voz del usuario, supondría disparar la carga de procesamiento y, en consecuencia, provocaría retrasos considerables en la respuesta de los sistemas de reconocimiento. En segundo lugar, y no menos importante, por motivos de eficacia: la eliminación de los segmentos de señal que no contienen la voz del usuario, limita considerablemente el espacio de búsqueda del sistema de reconocimiento, reduciendo sustancialmente su tasa de error. Por estos motivos, los sistemas comerciales de reconocimiento automático del habla incluyen un módulo de detección de segmentos de habla y de ruido.

Como consecuencia de la importancia de la detección de segmentos de habla, los esfuerzos para conseguir llevar a cabo esta tarea adecuadamente han sido muy numerosos.

Por ejemplo, la solicitud de patente japonesa JP-A-9050288 divulga un procedimiento de detección de segmentos de habla. En concreto, se determinan los puntos de inicio y finalización del segmento de habla mediante la comparación de la amplitud de la señal de entrada con un umbral. Este procedimiento presenta el inconveniente de que el funcionamiento depende del nivel de la señal de ruido, por lo que sus resultados no son adecuados en presencia de ruidos de gran amplitud.

A su vez, la solicitud de patente japonesa JP-A-1244497 divulga un procedimiento de detección de segmentos de habla basado en el cálculo de la energía de la señal. En concreto, se calcula la energía media de las primeras tramas de habla y utiliza el valor obtenido como estimación de la energía de la señal de ruido superpuesta a la voz. A continuación, se detectan los pulsos de voz mediante la comparación de la energía de cada trama de la señal con un umbral dependiente de la energía de la señal de ruido estimada. De esta forma, se compensa la posible variabilidad de valores de energía de la señal de ruido. Sin embargo, el procedimiento no funciona correctamente cuando aparecen segmentos de ruido de gran amplitud y corta duración.

En la patente estadounidense US-6317711 también se divulga un procedimiento de detección de segmentos de habla. En este caso, para cada trama de señal se obtiene un vector de características mediante una parametrización LPC-cepstra y MEL-cepstra. A continuación, se busca el valor mínimo de dicho vector y se normalizan todos los elementos de dicho vector dividiendo su valor por este valor mínimo. Finalmente se compara el valor de la energía normalizada con un conjunto de umbrales predeterminados para detectar los segmentos de habla. Este procedimiento ofrece mejores resultados que el anterior, aunque sigue presentando dificultades para detectar segmentos de habla en condiciones de ruido desfavorables.

En la patente estadounidense US-6615170 se divulga un procedimiento alternativo de detección de segmentos de habla que, en lugar de basarse en la comparación de un parámetro o un vector de parámetros con un umbral o conjunto de umbrales, se basa en el entrenamiento de modelos acústicos de ruido y de habla y en la comparación de la señal de entrada con dichos modelos, determinando si una determinada trama es habla o ruido mediante la maximización de la máxima verosimilitud. En el documento FR 2 856 506 A1 se utiliza una máquina de estados

Aparte de estas patentes y otras similares, el tratamiento de la tarea de la detección de segmentos de habla y ruido en la literatura científica es muy extenso, existiendo numerosos artículos y ponencias que presentan diferentes procedimientos de llevar a cabo dicha detección. Así, por ejemplo, en "*Voice Activity Detection Based on Conditional MAP Criterion*" (Jong Won Shin, Hyuk Jin Kwon, Suk Ho Jin,Nam Soo Kim; en IEEE Signal Processing Letters, ISSN: 1070-9908, Vo. 15, Feb. 2008) se describe un procedimiento de detección de habla basado en una variante del criterio MAP (*maximum a posteriori*), que clasifica las tramas de señal en habla o ruido basándose en parámetros espectrales y utilizando umbrales diferentes dependiendo de los resultados de clasificación inmediatamente anteriores.

En lo que respecta al ámbito de la normalización, cabe destacar la recomendación de un procedimiento de detección de habla incluida en el estándar de la ETSI de reconocimiento del habla distribuido (ETSI ES 202 050 v1.1.3.

Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms. Technical Report ETSI ES 202 050, ETSI). El procedimiento recomendado en el estándar se basa en el cálculo de tres parámetros de la señal para cada trama de la misma y su comparación con tres umbrales correspondientes, utilizando un conjunto de varias tramas consecutivas para tomar la decisión habla/ruido final.

Sin embargo, a pesar de la gran cantidad de procedimientos propuestos, en la actualidad la tarea de detección de segmentos de habla sigue presentando importantes dificultades. Los procedimientos propuestos hasta el momento, tanto los basados en la comparación de parámetros con umbrales, como los basados en clasificación estadística, son insuficientemente robustos en condiciones desfavorables de ruido, especialmente en presencia de ruido no estacionario, lo que provoca un aumento de los errores de detección de segmentos de habla en tales condiciones. Por este motivo, la utilización de estos procedimientos en entornos particularmente ruidosos, como es el caso del interior de automóviles, presenta importantes problemas.

Es decir, los procedimientos de detección de segmentos de habla propuestos hasta el momento, tanto los basados en la comparación de parámetros de la señal con umbrales como los basados en comparación estadística, presentan importantes problemas de robustez en entornos de ruido desfavorables. Su funcionamiento se degrada en particular ante la presencia de ruidos de carácter no estacionario.

Como consecuencia de la falta de robustez en determinadas condiciones, resulta inviable o particularmente difícil la utilización de sistemas de reconocimiento automático del habla en determinados entornos (como por ejemplo, el interior de automóviles). En estos casos, el empleo de procedimientos de detección de segmentos de habla basados en comparación de parámetros de la señal con umbrales, o bien basados en comparaciones estadísticas, no proporciona resultados adecuados. En consecuencia, los reconocedores automáticos del habla obtienen numerosos resultados erróneos, así como frecuentes rechazos de las pronunciaciones del usuario, lo que dificulta enormemente la utilización de este tipo de sistemas.

Descripción de la invención

La invención se refiere a un procedimiento de detección de segmentos de habla de acuerdo con la reivindicación 1. Realizaciones preferidas del procedimiento se definen en las reivindicaciones dependientes.

La presente propuesta trata de hacer frente a tales limitaciones, ofreciendo un procedimiento de detección de segmentos de habla robusto en entornos ruidosos, incluso en presencia de ruidos no estacionarios. Para ello, el procedimiento propuesto se basa en la combinación de tres criterios para tomar la decisión de clasificar los segmentos de la señal de entrada como habla o como ruido. En concreto, se utiliza un primer criterio relacionado con la energía de la señal, basado en la comparación con un umbral. Como segundo criterio se utiliza una comparación estadística de una serie de parámetros espectrales de la señal con unos modelos de habla y de ruido. Y se utiliza un tercer criterio basado en la duración de los distintos pulsos de voz y ruido, basado en la comparación con un conjunto de umbrales.

El procedimiento de detección de segmentos de habla propuesto se realiza en tres etapas. En la primera etapa se descartan las tramas de señal cuya energía no supera un cierto umbral energético, cuyo valor se actualiza automáticamente en tiempo real en función del nivel de ruido existente. En la segunda etapa, las tramas de habla no descartadas se someten a un procedimiento de toma de decisión que combina los tres criterios expuestos para clasificar dichas tramas como habla o ruido. Finalmente, en la tercera etapa se lleva cabo una validación de los segmentos de habla y ruido obtenidos según un criterio de duración, eliminándose los segmentos cuya duración no supere un cierto umbral.

La combinación de los tres criterios, así como la realización del procedimiento en las tres etapas propuestas permite obtener los segmentos de habla y ruido con mayor precisión que la obtenida con otros procedimientos, especialmente en condiciones de ruido desfavorables. Esta detección de segmentos se lleva a cabo en tiempo real y, por tanto, puede aplicarse en sistemas de reconocimiento automático del habla interactivos.

El objeto de la presente invención es un procedimiento de detección de segmentos de habla y de ruido en una señal digital de audio de entrada, estando dividida dicha señal de entrada en una pluralidad de tramas que comprende:

- una primera etapa en la que se realiza una primera clasificación de una trama como ruido si el valor medio de la energía para esta trama y las N tramas anteriores no es superior a un primer umbral de energía, siendo N un número entero mayor que 1;
- una segunda etapa en la que para cada trama que no ha sido clasificada como ruido en la primera etapa se decide si dicha trama se clasifica como ruido o como habla basándose en combinar al menos un primer criterio de similitud espectral de la trama con modelos acústicos de ruido y de habla, un segundo criterio de análisis de energía de la trama respecto a un segundo umbral de energía, y un tercer criterio de duración consistente en utilizar una máquina de estados para detectar el inicio de un segmento como acumulación de un número determinado de tramas consecutivas con similitud espectral superior a un primer umbral acústico y otro número determinado de tramas consecutivas con similitud espectral inferior a dicho primer umbral acústico para detectar el fin de dicho segmento;

- una tercera etapa en la que se revisa la clasificación como habla o como ruido de las tramas de señal llevada a cabo en la segunda etapa utilizando criterios de duración, clasificando como ruido los segmentos de habla de duración inferior a un primer umbral de duración mínima de segmento, así como aquellos que no contienen un determinado número de tramas consecutivas que simultáneamente superan dicho umbral acústico y dicho segundo umbral de energía.

Es decir, el procedimiento de la invención se realiza en tres etapas: una primera basada en umbral de energía, una segunda etapa de toma de decisión multicriterio y una tercera de comprobación de duraciones.

La toma de decisión de la segunda etapa está basada en:

- Por un lado, la utilización simultánea de tres criterios: similitud espectral, valor energético y duración (es necesario un mínimo número de tramas consecutivas similares espectralmente al modelo de ruido al final del segmento para dar éste por terminado).
- Por otro, la utilización de diferentes estados, lo que introduce cierta histéresis tanto para detectar el comienzo del segmento (hace falta acumular varias tramas consimilitud espectral superior al umbral) como para el final del mismo (histéresis).

Esto hace que mejore el funcionamiento eliminando falsos principios y finales de segmento.

En la tercera etapa se utilizan preferiblemente dos umbrales de duración:

- Un primer umbral de duración mínima de segmento.
- Un segundo umbral de duración de tramas consecutivas que cumplen tanto el criterio de similitud espectral como el de energía mínima.

La utilización de este doble umbral mejora frente a ruidos impulsivos y balbuceos del usuario.

La invención puede utilizarse como parte de un sistema de reconocimiento del habla. También puede utilizarse como parte de un sistema de identificación o verificación del locutor, o bien como parte de un sistema de detección acústica del idioma o de indexado acústico de contenidos multimedia.

La utilización de los criterios de duración, tanto en la segunda como en la tercera etapa, hace que el procedimiento clasifique correctamente ruidos no estacionarios y balbuceos del usuario, algo que no consiguen hacer los procedimientos conocidos hasta el momento: los criterios basados en umbrales energéticos no son capaces de discriminar los ruidos no estacionarios con altos valores de energía, mientras que los criterios basados en comparación de características acústicas (sean en el dominio del tiempo, sean en el dominio espectral) no son capaces de discriminar sonidos guturales y balbuceos del usuario, dado su similitud espectral con los segmentos de habla. Sin embargo, la combinación de similitud espectral y energía permite discriminar un mayor número de este tipo de ruidos de los segmentos de habla. Y el empleo de criterios de duración permite evitar que los segmentos de señal con este tipo de ruidos sean clasificados erróneamente como segmentos de habla.

Por otra parte, el modo en que se combinan los tres criterios en las etapas descritas del procedimiento optimiza la capacidad de clasificar correctamente los segmentos de habla y ruido. En concreto, la aplicación de un primer umbral de energía evita que segmentos con bajo contenido energético se tengan en cuenta en la comparación acústica. De esta forma, se evitan resultados impredecibles, algo habitual en procedimientos de detección basados en comparación acústica que no filtran este tipo de segmentos, así como los que comparan un vector de características mixto, con características espectrales y energéticas. La utilización de un segundo umbral de energía, evita que en la primera etapa se eliminen segmentos de habla con niveles bajos de energía, ya que permite utilizar un primer umbral energético poco restrictivo, que elimine sólo los segmentos de ruido con muy bajo nivel de energía, dejándose la eliminación de segmentos de ruido de mayor potencia para la segunda etapa, en la que interviene el segundo umbral energético, más restrictivo. La utilización combinada de los umbrales acústicos y energético en la segunda etapa permite discriminar los segmentos de ruido de los de habla: por un lado, la exigencia de superar ambos umbrales evita clasificar como habla los segmentos de ruido de alta energía pero con características espectrales diferentes del habla (ruidos no estacionarios, como golpes o chasquidos) y los segmentos de ruido similares acústicamente al habla pero con baja energía (balbuceos y sonidos guturales); por otro lado, la utilización de dos comparaciones independientes en lugar de un vector de características mixto (acústico y energético) permite ajustar el procedimiento de detección. El empleo de criterios de duración en esta segunda etapa (necesidad de superar un umbral de puntuaciones acústicas acumuladas al inicio del segmento de habla, y de concatenar un número mínimo de tramas de señal de ruido al final del mismo), permite detectar como ruido los segmentos de señal con ruidos no estacionarios de corta duración, así como clasificar como habla los segmentos correspondientes a sonidos que, aun siendo habla, tienen menor tono, como es el caso de los fonemas correspondientes a consonantes oclusivas y fricativas (k, t, s,…). Finalmente, el empleo de la tercera etapa permite hacer un filtrado final, eliminando los segmentos de ruido que han sido clasificados como habla pero no alcanzan la duración mínima, corrigiendo los errores de las dos primeras etapas del procedimiento con un procedimiento diferente respecto a todos los utilizados en otros procedimientos.

La correcta clasificación de los tramos de señal con ruidos de energía alta y con balbuceos, hace que el

procedimiento se puede emplear en sistemas de reconocimiento en diferentes entornos: oficina, hogar, interior de automóviles, etc., y con diferentes canales de utilización (microfónico o telefónico). Asimismo, es aplicable en diferentes tipos de aplicaciones vocales: servicios vocales de información, control vocal de equipos, etc.

Breve descripción de los dibujos

Para complementar la descripción que se está realizando y con objeto de ayudar a una mejor comprensión de las características de la invención, a continuación se pasa a describir de manera breve un modo de realización de la invención, como ejemplo ilustrativo y no limitativo de ésta.

La Figura 1 representa un diagrama de bloques del procedimiento de detección de segmentos de habla.
La Figura 2 muestra un diagrama de estados del proceso de clasificación de tramas de habla y ruido.
La Figura 3 muestra el procedimiento de comprobación de tramas que cumplen simultáneamente umbrales acústico y energético.
La Figura 4 representa el Diagrama de flujo de la validación de umbrales de duración.

Descripción de una realización preferida de la invención

De acuerdo con la realización preferida de la invención, el procedimiento de detección de segmentos de habla y ruido se lleva a cabo en tres etapas.

Como paso previo al procedimiento se divide la señal de entrada en tramas de muy corta duración (entre 5 y 50 milisegundos), que son procesadas una tras otra.

Como se muestra en la figura 1, en una primera etapa 10, para cada trama 1 se calcula su energía. Se calcula (bloque 11: cálculo energía media N últimas tramas) el promedio del valor de la energía para esta trama y las N tramas anteriores, siendo N un número entero cuyos valores varían dependiendo del entorno; típicamente N=10 en entornos poco ruidosos y N>10 para entornos ruidosos. Tras ello, se compara (bloque 12: validación umbral de energía media) este valor medio con un primer umbral de energía Umbral_energ1, cuyo valor es modificado en la segunda etapa en función del nivel de ruido, y siendo configurable el valor inicial del mismo; típicamente, para tramas de 10 ms, Umbral_energ1=15, valor que puede ajustarse según la aplicación. Si el valor medio de energía de las últimas tramas no supera dicho primer umbral de energía Umbral_energ1, la trama es clasificada definitivamente como ruido y se finaliza el procesado de la misma, comenzando el proceso de la siguiente trama de la señal. Si, por el contrario, el valor medio sí supera dicho primer umbral de energía, la trama continúa procesándose, pasando a la segunda etapa 20 del procedimiento.

En la segunda etapa 20 se realizan dos procesos:

- una comparación estadística de la trama que se está procesando con unos modelos acústicos de habla y de ruido (bloque 21: comparación estadística con modelos acústicos (algoritmo Viterbi)), y
- un proceso de clasificación de la trama (bloque 22: clasificación de tramas) como habla o ruido (véase figura 2).

Para llevar a cabo la comparación estadística, se obtiene en primer lugar un vector de características consistente en un conjunto de parámetros espectrales obtenidos a partir de la señal. En concreto, se selecciona un subconjunto de los parámetros que componen el vector de características propuesto en el estándar ETSI ES 202 050.

A continuación se describe cómo se realiza la selección del subconjunto de parámetros:

- Se estiman en primer lugar las funciones densidad de probabilidad del valor de cada uno de los parámetros para las tramas de habla y las de ruido, a partir de los valores del parámetro obtenidos con un conjunto de señales acústicas de habla y ruido distintas de las que se van a analizar.
- Haciendo uso de las funciones densidad de probabilidad estimadas, se calcula la probabilidad de error de clasificación de cada parámetro.
- Se crea una lista de los parámetros ordenados de menor a mayor valor de esta probabilidad de error.
- Se elige un subconjunto formado por los N primeros parámetros de la lista, estando el valor de N comprendido entre 0 y 39. Típicamente N=5, pero puede variar en función de la aplicación.

La comparación estadística requiere la existencia de unos modelos acústicos de habla y ruido. En concreto, se emplean modelos ocultos de Márkov (HMM, Hidden Markov Model) para modelar estadísticamente dos unidades acústicas: una representa las tramas de habla y otra representa las tramas de ruido. Estos modelos se obtienen antes de utilizar el procedimiento de detección de segmentos de habla y ruido de la presente invención. Para ello, con carácter previo, se entrenan estas unidades acústicas, utilizando para ello grabaciones que contienen segmentos de habla y ruido etiquetados como tales.

La comparación se lleva a cabo utilizando el algoritmo de Viterbi. De esta forma, a partir del vector de características obtenido en la trama que se está procesando, de los modelos estadísticos de habla y ruido, y de los datos de comparación de las tramas procesadas anteriormente, se determina la probabilidad de que la trama actual sea habla

y la probabilidad de que sea ruido. Asimismo se calcula un parámetro de puntuación acústica calculado al dividir la probabilidad de que la trama sea habla entre la probabilidad de que la trama sea ruido.

El proceso de clasificación de tramas (bloque 22) se lleva a cabo mediante un proceso de toma de decisión (véase figura 2) que tiene en cuenta el parámetro de puntuación acústica obtenido en el proceso de comparación estadística 21 y otros criterios, entre ellos, las decisiones de clasificación como habla o ruido de las tramas anteriores.

Esta figura 2 representa un diagrama de estados, en el que cuando se produce una transición (por ejemplo si la puntuación acústica es menor a “umbral_ac_1”), se pasa al estado indicado por la flecha, y se llevan a cabo los procesos incluidos en dicho estado. Por este motivo los procesos aparecen en el siguiente estado, una vez realizada la transición.

Tal y como se muestra en la figura 2, los pasos del proceso de toma de decisión son los siguientes:

* Estado inicial 210: Se pone a cero un acumulador de puntuaciones acústicas, Acumulador punt. Acústicas (2101). Se clasifican como ruido las posibles tramas previas que estuviesen clasificadas de forma provisional como habla o como ruido (2102).

A continuación se compara el parámetro de puntuación acústica obtenido en la comparación estadística con un primer umbral acústico, Umbral_ac_1.

A) Si no supera dicho primer umbral acústico Umbral_ac_1 se realizan las siguientes acciones:

I. Se clasifica definitivamente la trama actual como ruido (2102).
II. Se actualiza el primer umbral de energía utilizado en la primera etapa, Umbral_energ1 (2103), obteniendo una media (ponderada por un factor de memoria) entre su valor actual y el valor de la energía de la trama actual. El factor de memoria es un valor entre 0 y 1; típicamente tiene un valor de 0.9, ajustable en función de la aplicación.
III. Se pasa a procesar desde la primera etapa 10 del procedimiento la siguiente trama de señal.

B) En caso de que el parámetro de puntuación acústica obtenido en la comparación estadística supere dicho primer umbral acústico Umbral_ac_1, se realizan las siguientes acciones:

I. Se clasifica provisionalmente la trama actual como habla (2201).
II. Se actualiza el valor del acumulador de puntuaciones acústicas con el valor del parámetro de puntuación acústica obtenido en la comparación estadística (2202).
III. Se comprueba (2203) si la energía de la señal supera un segundo umbral de energía, Umbral_energ2 (ver figura 3), calculado a partir del valor actual del primer umbral de energía Umbral_energ1 (utilizado en la primera etapa 10 del procedimiento), cuyo valor se obtiene multiplicando dicho primer umbral de energía Umbral_energ1 por un factor y sumándole un desfase adicional. Este factor tiene un valor configurable entre 0 y 1, y el desfase, también con valor configurable, puede adquirir valores tanto positivos como negativos, oscilando su valor absoluto entre 0 y 10 veces el valor del primer umbral de energía, Umbral_energ1. Si supera dicho segundo umbral de energía, Umbral_energ2, se inicia con valor 1 un primer contador de tramas consecutivas que superan tanto el primer umbral acústico Umbral_ac_1 (de la comparación estadística) como este segundo umbral de energía, Umbral_energ2.
IV. Se pasa al siguiente estado: estado de comprobación de inicio de segmento de habla 220.
V. Se pasa a procesar desde la primera etapa 10 del procedimiento la siguiente trama de señal.

* Estado de comprobación de inicio de segmento de habla 220: se compara el parámetro de puntuación acústica obtenido en la comparación estadística con el primer umbral acústico, Umbral_ac_1.

A) Si no supera dicho primer umbral acústico Umbral_ac_1 se realizan las siguientes acciones:

I. Se clasifican como ruido (2102) tanto la trama en curso como todas las tramas anteriores clasificadas provisionalmente como habla.
II. Se ponen a cero el acumulador de puntuaciones acústicas (2101) y el primer contador de tramas consecutivas que superan tanto el segundo umbral de energía Umbral_energ_2 como el primer umbral de puntuación acústica Umbral_ac_1.
III. Se vuelve (2204) al estado inicial 210.
IV. Se pasa a procesar desde la primera etapa 10 del procedimiento la siguiente trama de señal.

B) En caso de que el parámetro de puntuación acústica obtenido en la comparación estadística supere dicho primer umbral acústico Umbral_ac_1, se realizan las siguientes acciones:

I. Se clasifica provisionalmente la trama actual como habla (2301 ó 2201).
II. Se comprueba (2303 ó 2203) si la energía de la señal supera el segundo umbral de energía, Umbral_energ2 (véase figura 3).

- Si lo supera se incrementa (2203A en fig. 3) el primer contador de tramas consecutivas que superan tanto el primer umbral acústico Umbral_ac_1 de la comparación estadística como el segundo umbral de energía Umbral_energ2.

- Si no lo supera se pone a cero (2203B en fig. 3) dicho primer contador de tramas consecutivas.

III. Se incrementa el valor del acumulador de puntuaciones acústicas (2202) sumándole el valor del parámetro de puntuación acústica obtenido en la comparación estadística.

IV. Se comprueba si el valor del acumulador de puntuaciones acústicas supera un segundo umbral de puntuaciones acústicas acumuladas, Umbral_ac_2.

- Si no supera dicho segundo umbral acústico Umbral_ac_2 se pasa a procesar desde la primera etapa 10 del procedimiento la siguiente trama de señal.
- Si supera dicho segundo umbral acústico Umbral_ac_2:

1º) Se pasa al estado de segmento de habla encontrado 230.
2º) Se pasa a procesar desde la primera etapa 10 del procedimiento la siguiente trama de señal.

* Estado de segmento de habla encontrado 230: se compara el parámetro de puntuación acústica obtenido en la comparación estadística con el primer umbral acústico, Umbral_ac_1.

A) Si el parámetro de puntuación acústica supera dicho primer umbral acústico Umbral_ac_1 se realizan las siguientes acciones:

I. Se clasifica provisionalmente la trama actual como habla (2301).

II. Se comprueba (2303) si la energía de la señal supera el segundo umbral de energía Umbral_energ2 (ver fig. 3).

- Si lo supera se incrementa (2203A en fig. 3) el primer contador de tramas consecutivas que superan tanto el primer umbral acústico Umbral_ac_1 de la comparación estadística como el segundo umbral de energía Umbral_energ2.
- Si no lo supera se pone a cero (2203B en fig. 3) dicho primer contador de tramas consecutivas.

III. Se pasa a procesar desde la primera etapa del procedimiento 10 la siguiente trama de señal.

B) En caso de que el parámetro de puntuación acústica obtenido en la comparación estadística no supere el primer umbral acústico, Umbral_ac_1, se realizan las siguientes acciones:

I. Se clasifica provisionalmente la trama actual como ruido (2401).Se pasa al estado de comprobación de fin de segmento de habla 240.

II. Se inicia a 1 (2302) un segundo contador de número de tramas consecutivas que no superan el umbral acústico modificado (la primera vez debe quedar por debajo de umbral_ac_1 para iniciar el contador; posteriormente los incrementos del contador se hacen cuando no se supere el umbral modificado (dividido por factor de histéresis)). .iv) Se pasa a procesar desde la primera etapa 10 del procedimiento la siguiente trama de señal.

* Estado de comprobación de fin de segmento de habla 240: Se compara el parámetro de puntuación acústica obtenido en la comparación estadística con un umbral modificado resultante de dividir el primer umbral acústico Umbral_ac_1 por un factor de histéresis, Histéresis.

A) Si el parámetro de puntuación acústica supera dicho umbral modificado, Umbral_ac_1/Histéresis se realizan las siguientes acciones:

I. Se clasifica provisionalmente la trama actual como habla. Asimismo, se clasifican provisionalmente como habla las tramas anteriores que se encontraban clasificadas provisionalmente como ruido (2301).

II. Se comprueba (2203 ó 2303) si la energía de la señal supera el segundo umbral de energía, Umbral_energ_2.

- Si lo supera se incrementa (2203A en fig. 3) el primer contador de tramas consecutivas que superan tanto el umbral modificado Umbral_ac_1/Histéresis de la comparación estadística como el segundo umbral de energía Umbral_energ2.
- Si no lo supera se pone a cero (2203B en fig. 3) dicho primer contador de tramas consecutivas.

III. Se pasa al estado de segmento de habla encontrado 230.

IV. Se pasa a procesar desde la primera etapa 10 del procedimiento la siguiente trama de señal.

B) En caso de que el parámetro de puntuación acústica obtenido en la comparación estadística no supere el

umbral modificado Umbral_ac_1/Histéresis, se realizan las siguientes acciones:

I. Se clasifica provisionalmente la trama actual como ruido (2401).

II. Se incrementa (2402) el segundo contador de número de tramas consecutivas que no superan el umbral acústico modificado.

III. Se comprueba si dicho segundo contador de número de tramas consecutivas que no superan el umbral acústico modificado, Umbral_ac_1/Histéresis es mayor que un umbral de duración de búsqueda de fin de pulso de voz, Umbral_dur_fin. Si es mayor, se pasa a la tercera etapa 30 del procedimiento de detección.

En caso contrario, se pasa a procesar desde la primera etapa 10 del procedimiento la siguiente trama de señal.

En la tercera etapa 30 del procedimiento de la presente invención se revisa la clasificación habla/ruido de las tramas de señal llevada a cabo en la segunda etapa utilizando criterios de duración para así finalmente detectar los segmentos de habla 2. Se hacen las siguientes comprobaciones (véase figura 4):

- Si el máximo valor alcanzado durante la segunda etapa 20 por el primer contador de tramas consecutivas que superan tanto el primer umbral acústico Umbral_ac_1 como el segundo umbral de energía Umbral_ energ_2 es menor (300A) que un primer umbral de duración, Umbral_dur1, se considera que el segmento de habla detectado es espurio (310), y se descarta. En consecuencia, todas las tramas de señal clasificadas provisionalmente como habla y como ruido, que cumplan este criterio, se clasifican definitivamente como ruido.

- Si el máximo valor alcanzado durante la segunda etapa 20 de dicho primer contador es mayor o igual (300B) que dicho primer umbral de duración, Umbral_dur_1, se comprueba (301) si el número total de todas las tramas clasificadas provisionalmente como habla supera un segundo umbral de duración Umbral_dur2.

  - En caso de no superarlo (301A), se considera que el segmento de habla detectado es espurio (320) y, en consecuencia, todas las tramas de señal clasificadas provisionalmente como habla o como ruido que cumplan este criterio, se clasifican definitivamente como ruido.
  - Si se supera (301B) este segundo umbral de duración, Umbral_dur2, las tramas clasificadas provisionalmente como habla se clasifican de forma definitiva como habla (330), y las tramas clasificadas provisionalmente como ruido se clasifican definitivamente como ruido.

En la tercera etapa se llevan a cabo, además, las siguientes acciones:

- Se actualiza el primer umbral de energía Umbral_energ1 utilizado en la primera etapa 10 del procedimiento, obteniendo una media (ponderada por un factor de memoria) entre su valor actual y el valor de la energía de la trama actual.
- Se pasa a procesar desde la primera etapa 10 del procedimiento la siguiente trama de señal. En caso de que dicha trama pase a la segunda etapa 20 del procedimiento, el proceso de toma de decisión comenzará desde el estado inicial 210.

La invención ha sido descrita según una realización preferente de la misma, pero para el experto en la materia resultará evidente que múltiples variaciones pueden ser introducidas en dicha realización preferente sin exceder el objeto de la invención reivindicada.

REIVINDICACIONES

1. Procedimiento de detección de segmentos de habla (2) y de ruido en una señal digital de audio de entrada, estando dividida dicha señal de entrada en una pluralidad de tramas (1) que comprende:

- una primera etapa (10) en la que se realiza una primera clasificación de una trama como ruido si el valor medio de la energía para esta trama y las N tramas anteriores no es superior a un primer umbral de energía (umbral_energ1), siendo N un número entero mayor que 1;
- una segunda etapa (20) en la que para cada trama que no ha sido clasificada como ruido en la primera etapa se decide si dicha trama se clasifica como ruido o como habla en base a la combinación de al menos un primer criterio de similitud espectral de la trama con modelos acústicos de ruido y de habla, un segundo criterio de análisis de energía de la trama respecto a un segundo umbral de energía (umbral_energ2) y un tercer criterio de duración consistente en utilizar una máquina de estados para detectar el inicio de un segmento como acumulación de un número determinado de tramas consecutivas con similitud espectral superior a un primer umbral acústico (umbral_ac1) y otro número determinado de tramas consecutivas con similitud espectral inferior a dicho primer umbral acústico para detectar el fin de dicho segmento, en el que en la segunda etapa, para cada trama que no ha sido clasificada como ruido en la primera etapa:

  - se calcula una probabilidad de que la trama sea una trama de ruido comparando unas características espectrales de dicha trama con esas mismas características espectrales de un grupo de tramas clasificadas como ruido que no pertenecen a la señal que se está analizando;
  - se calcula una probabilidad de que la trama sea una trama de habla comparando unas características espectrales de dicha trama con esas mismas características espectrales de un grupo de tramas clasificadas como habla que no pertenecen a la señal que se está analizando;
  - se calcula un estado siguiente de la máquina de estados en función de al menos, una relación entre la probabilidad de que la trama sea una trama de habla y la probabilidad de que la trama sea una trama de ruido, y de un estado actual de dicha máquina de estados, y

- una tercera etapa (30) en la que se revisa la clasificación como habla o como ruido de las tramas de señal llevada a cabo en la segunda etapa utilizando criterios de duración, clasificando como ruido los segmentos de habla de duración inferior a un primer umbral de duración mínima de segmento, así como aquellos que no contienen un determinado número de tramas consecutivas que simultáneamente superan dicho umbral acústico y dicho segundo umbral de energía;

en el que la máquina de estados comprende, al menos, un estado inicial (210), un estado en el que se comprueba que se ha iniciado un segmento de habla (220), un estado en el que se comprueba que continúa el segmento de habla (230), y un estado en el que se comprueba que ha finalizado el segmento de habla (240);

y en el que para producirse una transición entre el estado en el que se comprueba que se ha iniciado un segmento de habla (220) y el estado en el que se comprueba que continúa un segmento de habla (230), se requieren, al menos, dos tramas consecutivas en las que la relación entre la probabilidad de que la trama sea una trama de habla y la probabilidad de que la trama sea una trama de ruido sea superior a dicho primer umbral acústico.

2. Procedimiento según la reivindicación 1, en el que en dicha tercera etapa se utilizan dos umbrales de duración:

- un primer umbral (umbral_dur1) de duración mínima de segmento o número mínimo de tramas consecutivas clasificadas como habla o como ruido;
- un segundo umbral de duración (umbral_dur2) de tramas consecutivas que en la segunda etapa cumplen tanto el criterio de similitud espectral como el criterio de análisis de energía de la trama.

3. Procedimiento según cualquiera de las reivindicaciones 1-2, en el que dicho criterio de similitud espectral usado en la segunda etapa consiste en un análisis comparativo de características espectrales de dicha trama con características espectrales de dichos modelos acústicos de ruido y de habla previamente establecidos.

4. Procedimiento según la reivindicación 3, en el que dicho análisis comparativo de características espectrales se realiza utilizando el algoritmo de Viterbi.

5. Procedimiento según cualquiera de las reivindicaciones 1-4, en el que dichos modelos acústicos de ruido y de habla previamente establecidos se obtienen modelando estadísticamente dos unidades acústicas, de ruido y habla respectivamente, mediante modelos ocultos de Márkov.

6. Procedimiento según cualquiera de las reivindicaciones 1-5, en el que para producirse una transición entre el estado que comprueba que ha finalizado un segmento de habla (240) y el estado inicial (210) se requieren, al menos, dos tramas consecutivas en las que la relación entre la probabilidad de que la trama sea una trama de habla y la probabilidad de que la trama sea una trama de ruido sea inferior a un primer umbral acústico dividido por un factor.

7. Procedimiento según cualquiera de las reivindicaciones 1-6, en el que el primer umbral de energía utilizado en la primera etapa se actualiza dinámicamente ponderando su valor actual y el valor de energía de las tramas clasificadas como ruido en la segunda y la tercera etapas.

8. Procedimiento según la reivindicación 1-2, en el que el criterio de análisis de la energía de la trama (2203, 2303) consiste en superar un segundo umbral de energía, calculado al multiplicar el primer umbral de energía por un factor y sumarle un desfase.

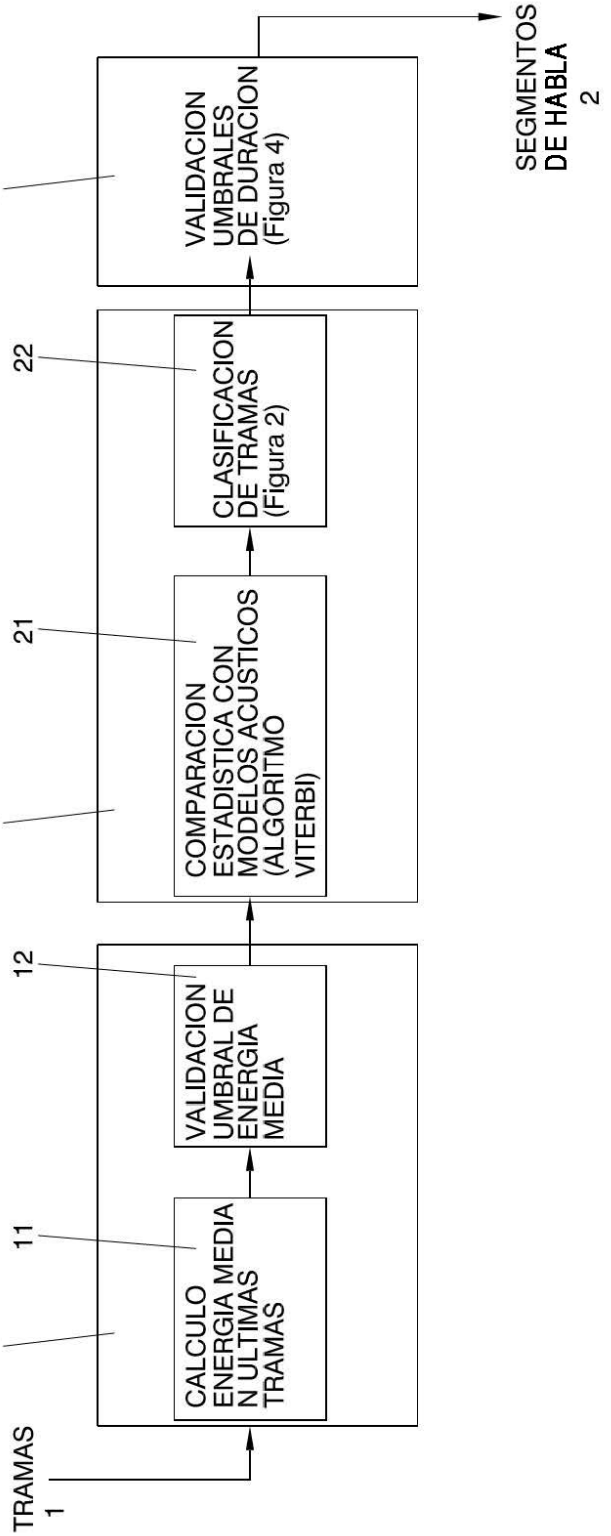
TRAMAS
1
10
CALCULO ENERGIA MEDIA N ULTIMAS TRAMAS
11
12
VALIDACION UMBRAL DE ENERGIA MEDIA
20
21
COMPARACION ESTADISTICA CON MODELOS ACUSTICOS (ALGORITMO VITERBI)
22
CLASIFICACION DE TRAMAS (Figura 2)
30
VALIDACION UMBRALES DE DURACION (Figura 4)
SEGMENTOS DE HABLA
2

FIG. 1

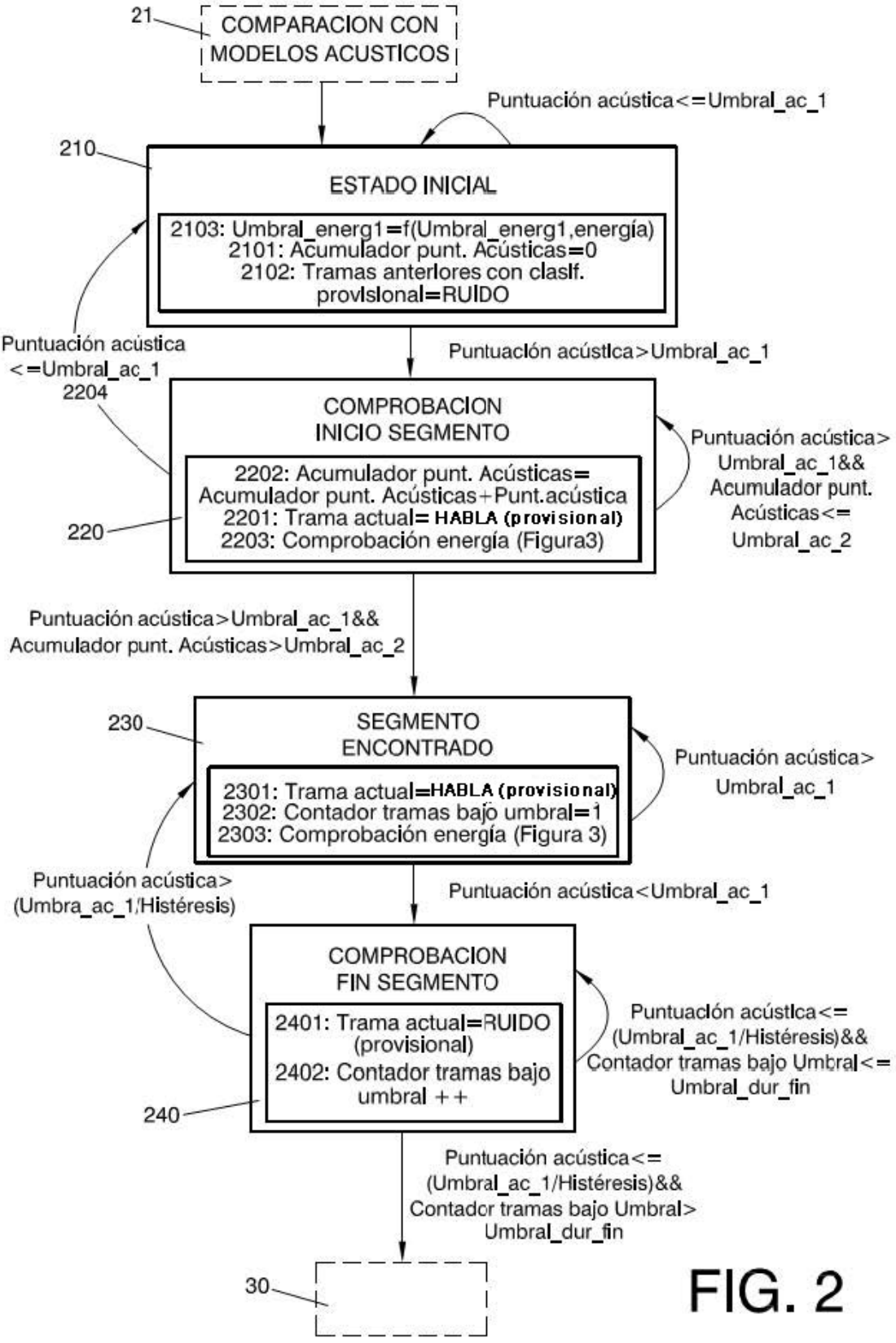
21
COMPARACION CON MODELOS ACUSTICOS
Puntuación acústica<=Umbral_ac_1
210
ESTADO INICIAL
2103: Umbral_energ1=f(Umbral_energ1,energía)
2101: Acumulador punt. Acústicas=0
2102: Tramas anteriores con clasif. provisional=RUIDO
Puntuación acústica <=Umbral_ac_1
Puntuación acústica>Umbral_ac_1
2204
COMPROBACION INICIO SEGMENTO
2202: Acumulador punt. Acústicas= Acumulador punt. Acústicas+Punt.acústica
2201: Trama actual= HABLA (provisional)
2203: Comprobación energía (Figura3)
220
Puntuación acústica> Umbral_ac_1&& Acumulador punt. Acústicas<= Umbral_ac_2
Puntuación acústica>Umbral_ac_1&& Acumulador punt. Acústicas>Umbral_ac_2
230
SEGMENTO ENCONTRADO
2301: Trama actual=HABLA (provisional)
2302: Contador tramas bajo umbral=1
2303: Comprobación energía (Figura 3)
Puntuación acústica> Umbral_ac_1
Puntuación acústica> (Umbra_ac_1/Histéresis)
Puntuación acústica<Umbral_ac_1
COMPROBACION FIN SEGMENTO
2401: Trama actual=RUIDO (provisional)
2402: Contador tramas bajo umbral ++
240
Puntuación acústica<= (Umbral_ac_1/Histéresis)&& Contador tramas bajo Umbral<= Umbral_dur_fin
Puntuación acústica<= (Umbral_ac_1/Histéresis)&& Contador tramas bajo Umbral> Umbral_dur_fin
30

FIG. 2

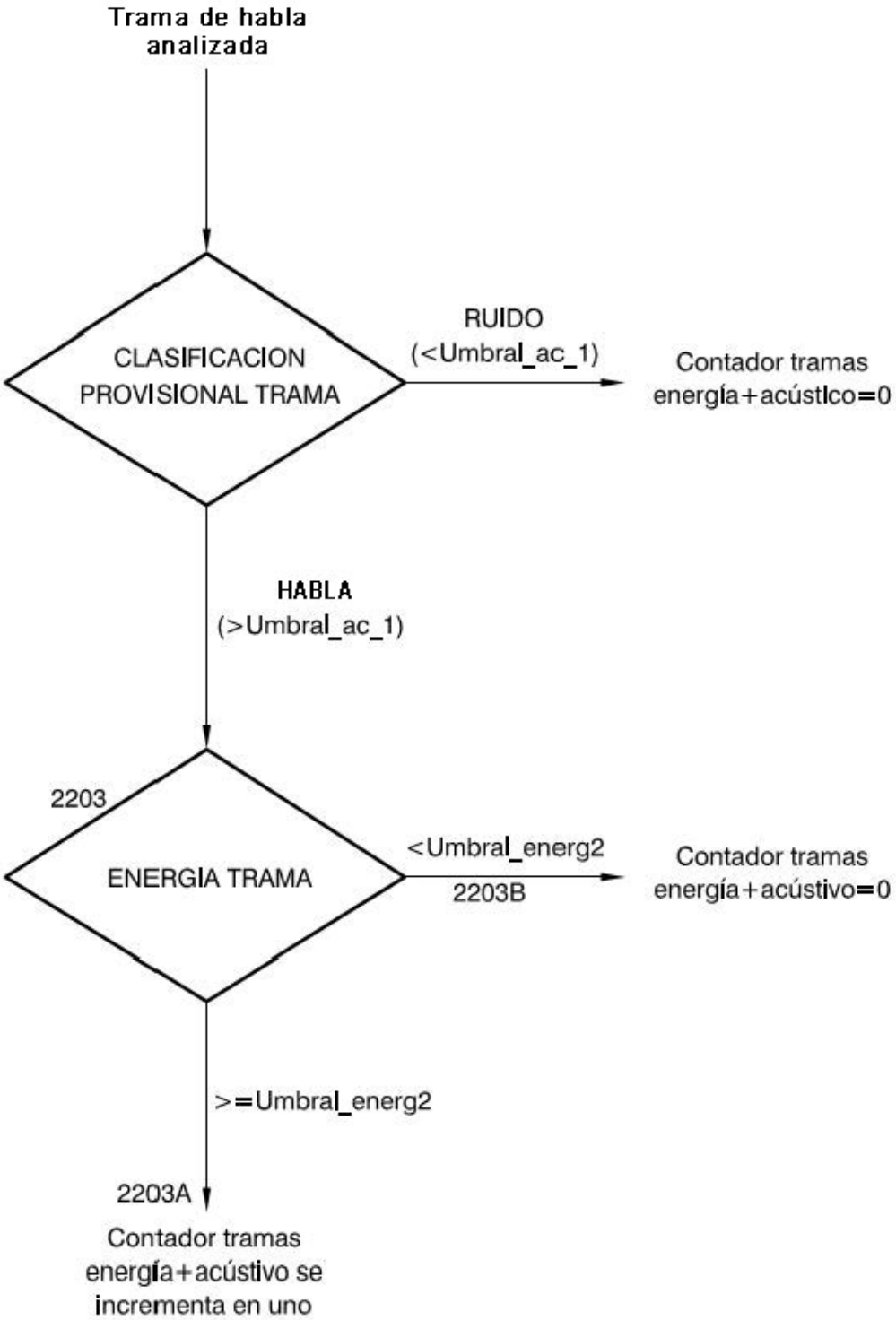
Trama de habla analizada
CLASIFICACION PROVISIONAL TRAMA
RUIDO
(<Umbral_ac_1)
Contador tramas energía+acústico=0
HABLA
(>Umbral_ac_1)
2203
ENERGIA TRAMA
<Umbral_energ2
2203B
Contador tramas energía+acústivo=0
>=Umbral_energ2
2203A
Contador tramas energía+acústivo se incrementa en uno

FIG. 3

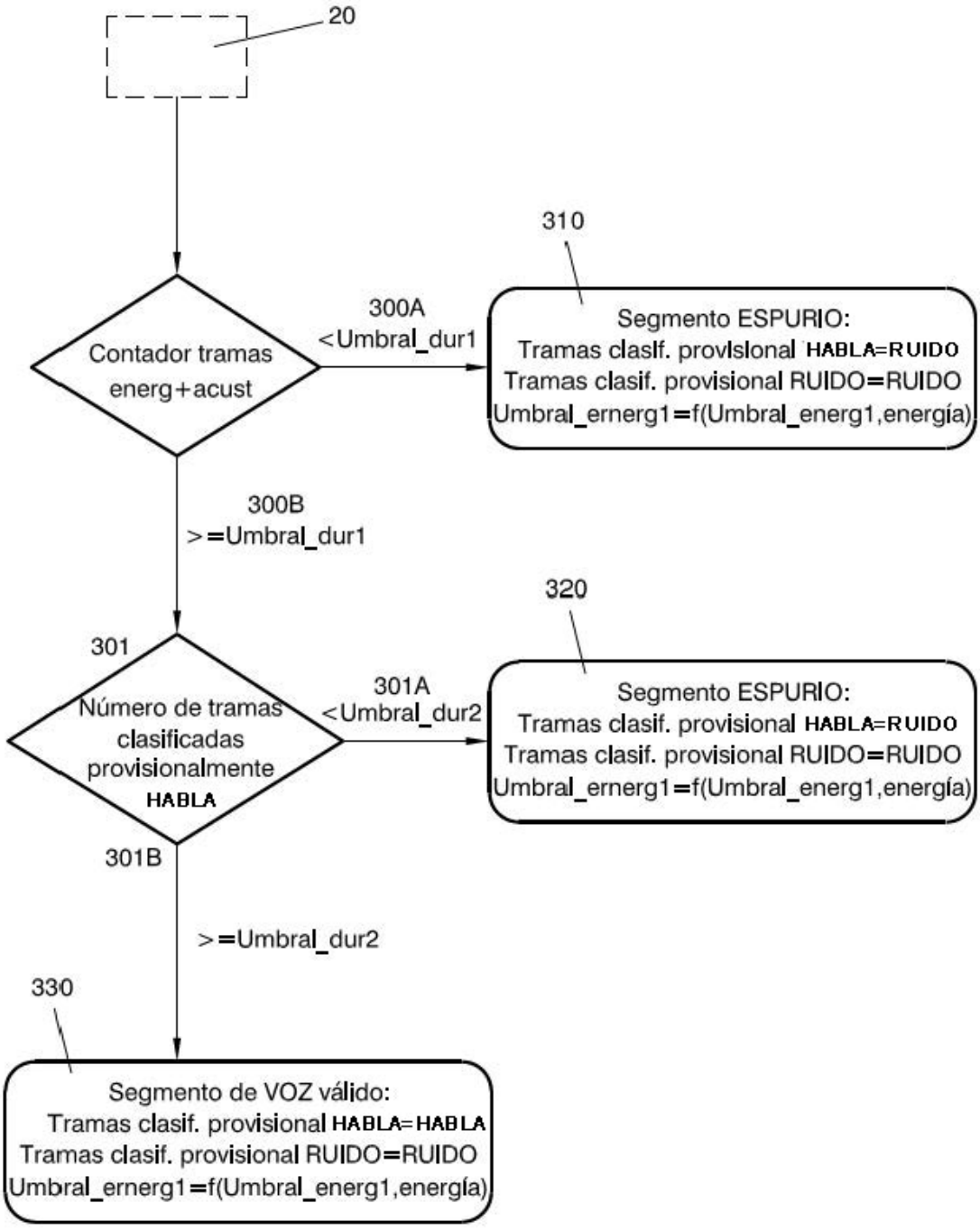
20
310
Contador tramas energ+acust
300A
<Umbral_dur1
Segmento ESPURIO:
Tramas clasif. provisional HABLA=RUIDO
Tramas clasif. provisional RUIDO=RUIDO
Umbral_ernerg1=f(Umbral_energ1,energía)
300B
>=Umbral_dur1
320
301
Número de tramas clasificadas provisionalmente HABLA
301A
<Umbral_dur2
Segmento ESPURIO:
Tramas clasif. provisional HABLA=RUIDO
Tramas clasif. provisional RUIDO=RUIDO
Umbral_ernerg1=f(Umbral_energ1,energía)
301B
>=Umbral_dur2
330
Segmento de VOZ válido:
Tramas clasif. provisional HABLA=HABLA
Tramas clasif. provisional RUIDO=RUIDO
Umbral_ernerg1=f(Umbral_energ1,energía)

FIG. 4