(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 531 137**

(51) Int. CI.:

G10L 25/78 (2013.01)
G10L 19/02 (2013.01)
G10L 25/51 (2013.01)
G10L 19/20 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **28.04.2011 E 11717266 (8)**

(97) Fecha y número de publicación de la concesión europea: **31.12.2014 EP 2702585**

(54) Título: **Clasificación de señales de audio basada en marcos**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
11.03.2015

(73) Titular/es:

TELEFONAKTIEBOLAGET LM ERICSSON (PUBL) (100.0%)
164 83 Stockholm, SE

(72) Inventor/es:

GRANCHAROV, VOLODYA y
NÄSLUND, SEBASTIAN

(74) Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

DESCRIPCIÓN

Clasificación de señales de audio basada en marcos

Campo técnico

La presente tecnología se refiere a la clasificación de señales de audio basadas en marcos o cuadros.

Antecedentes

Los métodos de clasificación de señales de audio son diseñados bajo diferentes supuestos: enfoque de tiempo real o fuera de línea, requisitos de diferente memoria y complejidad, etc.

Para un clasificador utilizado en codificación de audio, la decisión tiene que ser tomada normalmente sobre la base de marco a marco, basada totalmente en las estadísticas anteriores de señales. Muchas aplicaciones de codificación de audio, tales como codificación en tiempo real, imponen también fuertes limitaciones sobre la complejidad informática del clasificador.

La referencia [1] describe un discriminador (clasificador) complejo de habla/música basado en un estimador *a posteriori* del máximo Gaussiano multidimensional, una clasificación de modelo de mezcla Gaussiana, un esquema de división espacial basado en ramificaciones k-d o un clasificador de vecino más próximo. Con el fin de obtener una tasa de errores de decisión aceptable es necesario también incluir características de señales de audio que requieran una gran latencia.

La referencia [2] describe un discriminador de habla/música basado parcialmente en Frecuencias Espectrales en Línea (LSFs). Sin embargo, la determinación de LSFs es un procedimiento bastante complejo.

La referencia [5] describe la detección de actividad de voz basándose en la envoltura Modulada en Amplitud (AM) de un segmento de señal.

Compendio

Un objeto de la presente tecnología es la clasificación de señales de audio basada en marcos, de baja complejidad.

Este objeto se consigue de acuerdo con las reivindicaciones adjuntas.

Un primer aspecto de la presente tecnología implica un método de clasificación de señales de audio basada en marcos, que incluye los siguientes pasos:

- Determinar, para cada uno de un número predeterminado de marcos consecutivos, medidas de características que representen al menos las siguientes características: un coeficiente de auto-correlación, energía de señal de marco en un dominio comprimido, variación de energía de señales entre marcos.
- Comparar cada medida de característica determinada con al menos un correspondiente intervalo de características predeterminado.
- Calcular, para cada intervalo de características, una medida de fracciones que represente el número total de medidas correspondiente de características que esté comprendido dentro del intervalo de características.
- Clasificar el último de los marcos consecutivos como habla si cada medida de fracción se sitúa dentro de un intervalo de fracciones correspondiente, y como no-habla, de otro modo.

Un segundo aspecto de la presente tecnología implica un clasificador de audio para clasificación de señales de audio basada en marcos, que incluye:

- Un extractor de características configurado para determinar, para cada uno de un número predeterminado de cuadros consecutivos, medidas de características que representen al menos las siguientes características: un coeficiente de auto-correlación, energía de la señal de marco en un dominio comprimido, variación de energía de señales entre marcos.
- Un comparador de medidas de características configurado para comparar cada medida de característica determinada con al menos un correspondiente intervalo predeterminado de características.
- Un clasificador de marcos configurado para calcular, para cada intervalo de características, una medida de fracción que represente el número total de medidas de características correspondientes que estén comprendidas en el intervalo de características, y para clasificar el último de los marcos consecutivos como habla si cada medida de fracción se sitúa dentro de un correspondiente intervalo de fracciones, y como no-habla, de otro modo.

Un tercer aspecto de la presente tecnología implica una disposición de codificador de audio que incluye un

clasificador de audio de acuerdo con el segundo aspecto para clasificar marcos de audio en habla/no-habla y seleccionar con ello un método de codificación correspondiente.

Un cuarto aspecto de la presente tecnología implica una disposición de codec (codificación-descodificación) de audio que incluye un clasificador de audio de acuerdo con el segundo aspecto para clasificar marcos de audio en habla/no-habla para seleccionar un método de pos-filtración correspondiente.

Un quinto aspecto de la presente tecnología implica un dispositivo de comunicaciones de audio que incluye una disposición de codificador de audio de acuerdo con los aspectos tercero y cuarto.

Son ventajas de la presente tecnología la baja complejidad y lógica de decisión sencilla. Estas características la hacen especialmente apropiada para codificación de audio en tiempo real.

Breve descripción de los dibujos

La tecnología, junto con otros objetos y ventajas de la misma, se comprenderán mejor haciendo referencia a la descripción que sigue tomada junto con los dibujos que se acompañan, en los cuales:

La figura 1 es un diagrama de bloques que ilustra un ejemplo de una disposición de codificador de audio que utiliza un clasificador de audio;

La figura 2 es un diagrama que ilustra el seguimiento de energía máxima;

La figura 3 es un histograma que ilustra la diferencia entre habla y música para una característica concreta;

La figura 4 es un diagrama de flujo que ilustra la presente tecnología;

La figura 5 es un diagrama de bloques que ilustra otro ejemplo de una disposición de codificador de audio que utiliza un clasificador de audio;

La figura 6 es un diagrama de bloques que ilustra un ejemplo de realización de un clasificador de audio;

La figura 7 es un diagrama de bloques que ilustra un ejemplo de realización de un comparador de medidas de características en el clasificador de audio de la figura 6;

La figura 8 es un diagrama de bloques que ilustra un ejemplo de realización de un clasificador de marcos del clasificador de audio de la figura 6;

La figura 9 es un diagrama de bloques que ilustra un ejemplo de realización de un calculador de fracciones del clasificador de marcos de la figura 8;

La figura 10 es un diagrama de bloques que ilustra un ejemplo de realización de un selector de clase del clasificador de marcos de la figura 8;

La figura 11 es un diagrama de bloques de un ejemplo de realización de un clasificador de audio;

La figura 12 es un diagrama de bloques que ilustra otro ejemplo de una disposición de codificador de audio que utiliza un clasificador de audio;

La figura 13 es un diagrama de bloques que ilustra un ejemplo de una disposición de codec de audio que utiliza una decisión de habla/no-habla procedente de un clasificador de audio 12; y

La figura 14 es un diagrama de bloques que ilustra un ejemplo de un dispositivo de comunicación de audio que utiliza una disposición de codificador de audio.

Descripción detallada

En la siguiente descripción m indica el índice de muestra de audio en un marco y n indica el índice de marco. Un marco se define como un corto bloque de la señal de audio, por ejemplo de 20-40 ms, que contiene *M* muestras.

La figura 1 es un diagrama de bloques que ilustra un ejemplo de una disposición de codificador de audio que utiliza un clasificador de audio. Marcos consecutivos, denominados MARCO n, MARCO n+1, MARCO n+2, ….de muestras de audio se hacen avanzar hacia un codificador 10, el cual los codifica convirtiéndolos en una señal codificada. Un clasificador de audio de acuerdo con la presente tecnología ayuda al codificador 10 clasificando los marcos en habla/no-habla. Esto permite que el codificador utilice diferentes esquemas de codificación para diferentes tipos de señales de audio, tales como habla/música o habla/ruido de fondo.

La presente tecnología está basada en un conjunto de medidas de características que pueden ser calculadas directamente a partir de la forma de onda de la señal (o su representación en un dominio de frecuencias, como se describirá más adelante) con una complejidad de cálculo muy baja.

Las siguientes medidas de características son extraídas de la señal de audio sobre una base de cuadro a cuadro:

1. Una medida de característica que representa un coeficiente de auto-correlación entre muestras $x_m(n)$, preferiblemente el coeficiente de auto-correlación normalizado de primer orden. Esta medida de característica puede ser, por ejemplo, representada por:

$$T_n = \frac{\sum_{m=1}^{M} x_m(n) x_{m-1}(n)}{\sum_{m=2}^{M} x_m^2(n)} \quad (1)$$

2. Una medida de característica que representa la energía de la señal de marco en un dominio comprimido. Esta característica puede ser representada, por ejemplo, por:

$$E_n = 10 \cdot \log_{10}\left(\frac{1}{M}\sum_{m=1}^{M} x_m^2(n)\right) \quad (2)$$

donde la compresión es proporcionada por la función logarítmica.

Otro ejemplo es:

$$E_n = \left(\frac{1}{M}\sum_{m=1}^{M} x_m^2(n)\right)^{\alpha} \quad (3)$$

donde $0 < \alpha < 1$ es un factor de compresión. Una razón para preferir un dominio comprimido es que este emula el sistema de audición humano.

3. Una medida de característica que representa la variación de energía de la señal de marco entre marcos adyacentes. Esta medida de característica puede, por ejemplo, ser representada por:

$$\Delta E_n = \frac{\left\| E_n - E_{n-1} \right\|}{E_n + E_{n-1}} \quad (4)$$

Las medidas de características T_n, E_n, ΔE_n se calculan para cada marco y se usan para deducir ciertas estadísticas de señales. En primer lugar, se comparan T_n, E_n, ΔE_n con respectivos criterios previamente definidos (véanse las dos primeras columnas de la Tabla 1 siguiente), y las decisiones binarias para cierto número de marcos anteriores, por ejemplo N = 40 marcos anteriores se conservan en una memoria temporal. Obsérvese que algunas medidas de características (por ejemplo T_n, E_n en la Tabla 1) pueden ser asociadas a varios criterios. A continuación, las estadísticas (fracciones) de señal se obtienen a partir de valores almacenados temporalmente. Finalmente, un procedimiento de clasificación se basa en las estadísticas de señales.

Tabla 1

Parámetro	Criterio	Intervalo de caracterís.	Ejemplo de interv. de caracterís.	Fracción	Intervalo de Fracción	Ejemplo de intervalo de fracción
T_n	$T_n \leq \Theta_1$	$\{0, \Theta_1\}$	{0,0.98}	Φ_1	$\{T_{11}, T_{21}\}$	{0,0.65}
	$T_n \in \{\Theta_2, \Theta_3\}$	$\{\Theta_2, \Theta_3\}$	{0.8,0.98}	Φ_2	$\{T_{12}, T_{22}\}$	{0,0.375}
E_n	$E_n \geq \Theta_4 E_n^{MAX}$	$\{\Theta_4 E_n^{MAX}, \Omega\}$	$\{0.62E_n^{MAX}, \Omega\}$	Φ_3	$\{T_{13}, T_{23}\}$	{0,0.975}
	$E_n < \Theta_5$	$\{0, \Theta_5\}$	{0,42.4}	Φ_4	$\{T_{14}, T_{24}\}$	{0.025,1}
ΔE_v	$\Delta E_v > \Theta_6$	$\{\Theta_6, 1\}$	{0.065,1}	Φ_5	$\{T_{15}, T_{25}\}$	{0.075,1}

La columna 2 de la Tabla 1 describe ejemplos de los diferentes criterios para cada medida de característica T_n, E_n, ΔE_n. Aunque estos criterios parecen muy diferentes a primera vista, son realmente equivalentes a los intervalos de características ilustrados en la columna 3 de la Tabla 1. De ese modo, en una ejecución práctica los criterios pueden ser realizados probando si las medidas de características caen dentro de sus respectivos intervalos de características. Ejemplos de intervalos de características se dan en la columna 4 de la Tabla 1.

En la Tabla 1 se observa también que, en este ejemplo, el primer intervalo de características para la medida de característica E_n está definido por un parámetro auxiliar E_n^{MAX}. Este parámetro auxiliar representa el máximo de señal y es preferiblemente seguido de acuerdo con:

$$E_n^{MAX} = (1-\mu)E_{n-1}^{MAX} + \mu E_n \quad (5)$$

$$\mu = \begin{cases} 0{,}557 & si \quad E_n \geq E_{n-1}^{MAX} \\ 0{,}038 & si \quad E_n < E_{n-1}^{MAX} \\ 0{,}001 & si \quad E_n < 0{,}62E_{n-1}^{MAX} \end{cases}$$

Como se aprecia en la figura 2, este algoritmo de seguimiento tiene la propiedad de que los aumentos de energía de la señal son seguidos inmediatamente, mientras que las disminuciones de energía de señal son seguidas sólo lentamente.

Una alternativa al método de seguimiento descrito es utilizar una memoria temporal grande para almacenar valores anteriores de energía de marco. La longitud de la memoria temporal debe ser suficiente para almacenar valores de energía de marcos durante un periodo de tiempo que sea mayor que la mayor pausa esperada, por ejemplo 400 ms. Para cada nuevo marco, es suprimido el valor de energía de marco más antiguo y se añade el último valor de energía de marco. A continuación se determina el valor máximo de la memoria temporal.

La señal es clasificada como habla si todas las estadísticas de señales (la fracción Φ_1 en la columna 5 de la Tabla 1) pertenece a un intervalo de fracciones previamente definido (columna 6 de la Tabla 1), es decir, $\forall \Phi_1 = \{T_{1j}, T_{2j}\}$. Un ejemplo de intervalos de fracciones se da en la columna 7 de la Tabla 1. Si una o más de las fracciones Φ_l está fuera del intervalo de fracciones correspondiente $\{T_{1j}, T_{2j}\}$, la señal es clasificada como no-habla.

Las estadísticas o fracciones Φ_j de señales seleccionadas son motivadas por observaciones que indican que una señal de habla consiste en una cierta cantidad de segmentos alternativos con voz y sin voz. Una señal de habla puede ser normalmente también activa sólo durante un periodo limitado de tiempo y después ir seguida por un segmento de silencio. Las dinámicas o variaciones de energía son generalmente mayores en una señal de habla que en no-habla, tal como música, véase la figura 3, que ilustra un histograma de Φ_5 sobre bases de datos de habla y música. En la Tabla 2 siguiente se presenta una breve descripción de estadísticas o fracciones Φ_j de señales seleccionadas.

Tabla 2

Φ_1	Mide la cantidad de marcos sin voz en la memoria temporal (una decisión "sin voz" está basada en el espectro distorsionado, que puede estar basado, a su vez, en un coeficiente de auto-correlación)
Φ_2	Mide la cantidad de marcos con voz que no tienen espectro típico de habla distorsionado
Φ_3	Mide la cantidad de marcos de señal activos
Φ_4	Mide la cantidad de marcos que pertenecen a una región de señal de pausa o no activa
Φ_5	Mide la cantidad de marcos con gran dinámica o variación de energía

La figura 4 es un diagrama de flujo que ilustra la presente tecnología. El paso S1 determina, para cada número predeterminado de marcos consecutivos, medidas de características, por ejemplo T_n, E_n, ΔE_n, que representan al menos las características: auto-correlación (T_n), la energía de la señal de marco (E_n) en un dominio comprimido, variación de la energía de señal entre marcos. El paso S2 compara cada medida de características determinada con al menos un correspondiente intervalo de características predeterminado. El paso S3 calcula, para cada intervalo de características, una medida de fracción, por ejemplo Φ_j, que representa el número total de medidas correspondientes de características que caen dentro del intervalo de características. El paso S4 clasifica el último de los marcos consecutivos como habla si cada medida de fracción se sitúa dentro de un correspondiente intervalo de fracciones, y como no-habla en caso contrario.

En los ejemplos dados anteriormente, las medidas dadas de características en (1)-(4) son determinadas en el dominio de tiempo. Sin embargo, también es posible determinarlas en el dominio de frecuencia, como se ilustra en el diagrama de bloques de la figura 5. En este ejemplo de disposición de codificador de audio, el codificador 10 comprende un transformador de frecuencia 10A conectado al codificador de transformación 10B. El codificador 10 puede, por ejemplo, estar basado en la transformación de Coseno Discreta Modificada (MDCT). En este caso, las medidas de características T_n, E_n, ΔE_n se pueden determinar en el dominio de frecuencia a partir de K secciones o intervalos (bins) de frecuencia $X_k(n)$ obtenidas del transformador de frecuencia 10A. Esto no da lugar a ninguna complejidad o retardo adicional de cálculo, ya que la transformación de frecuencia es requerida de cualquier modo por el codificador de transformación 10B. En esta ejecución en dominio de frecuencia, la ecuación (1) puede ser

sustituida por la relación entre la parte alta y la baja del espectro:

$$T_n = \frac{\frac{2}{K}\sum_{k=1}^{K/2} X_k^2(n) - \frac{2}{K}\sum_{k=K/2+1}^{K} X_k^2(n)}{\frac{1}{K}\sum_{k=1}^{K} X_k^2(n)} \quad (6)$$

Las ecuaciones (2) y (3) se pueden sustituir por adiciones sobre secciones de frecuencia $X_k(n)$ en lugar de muestras de entrada $x_m(n)$, lo que da:

$$E_n = 10 \cdot \log_{10}\left(\frac{1}{K}\sum_{k=1}^{K} X_k^2(n)\right) \quad (7)$$

y

$$E_n = \left(\frac{1}{K}\sum_{k=1}^{K} X_k^2(n)\right)^{\alpha}, \quad (8)$$

respectivamente.

De manera similar, la ecuación (4) puede ser sustituida por:

$$\Delta E_n = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(X_k^2(n) - X_k^2(n-1)\right)^2} \quad (9)$$

o por:

$$\Delta E_n = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\log\left\{X_k^2(n)\right\} - \log\left\{X_k^2(n-1)\right\}\right)^2} \quad (10)$$

La descripción anterior se ha enfocado en las tres medidas de características T_n, E_n, ΔE_n para clasificar señales de audio. Sin embargo, se pueden añadir otras medidas de características manejadas del mismo modo. Un ejemplo de una medida de paso (frecuencia fundamental) $\hat{P}$ que se puede calcular haciendo máxima la función de auto-correlación:

$$\hat{P}_n = \arg_p \, máx\left(\sum_{m=P+1}^{M} x_m(n) x_{m-P}(n)\right) \quad (11)$$

Es también posible realizar la estimación de paso en el dominio cepstrum. Los coeficientes cepstrales $c_m(n)$ se obtienen por medio de la Transformación de Fourier Discreta (DFT) inversa del espectro de magnitud logarítmica. Esto se puede expresar en los siguientes pasos: realizar una DFT sobre el vector de forma de onda; en el vector de frecuencia resultante, tomar el valor absoluto y a continuación el logaritmo; finalmente, la Transformación de Fourier Discreta (DFT) inversa da el vector de coeficientes cepstrales. El lugar del pico en este vector es una estimación del dominio de frecuencia del periodo de paso. En notación matemática:

$$c_m(n) = IDFT\left\{\log\left|DFT\left\{x_m(n)\right\}\right|\right\}$$

$$\hat{P}_n = \arg_p \, máx(c_p(n)) \quad (12)$$

La figura 6 es un diagrama de bloques que ilustra un ejemplo de realización de un clasificador de audio. Esta realización es una ejecución en dominio de tiempo, pero también podría ser ejecutada en el dominio de frecuencia utilizando las secciones de frecuencia en lugar de muestras de audio. En la realización de la figura 6, el clasificador 12 de audio incluye un extractor 14 de características, un comparador 16 de medidas de características y un

clasificador 18 de marcos. El extractor 14 de características puede ser configurado para ejecutar las ecuaciones descritas anteriormente para determinar al menos T_n, E_n, ΔE_n. El comparador 16 de medidas de características está configurado para comparar cada medida de característica predeterminada con al menos un intervalo correspondiente de características predeterminado. El clasificador 18 de marcos está configurado para calcular, para cada intervalo de características, una medida de fracción que represente el número total de medidas de características correspondientes que caigan dentro del intervalo de características, y para clasificar el último de los marcos consecutivos como habla si cada medida de fracción se sitúa dentro de un intervalo de fracción correspondiente, y como no-habla, de otro modo.

La figura 7 es un diagrama de bloques que ilustra un ejemplo de realización del comparador 16 de medidas de características en el clasificador 12 de audio de la figura 6. Un comparador 20 de intervalos de características que recibe las medidas de características extraídas, por ejemplo T_n, E_n, ΔE_n, está configurado para determinar si las medidas de características se sitúan dentro de intervalos de características predeterminados, por ejemplo los intervalos dados en la Tabla 1 anterior. Estos intervalos de características se obtienen de un generador 22 de intervalos de características, ejecutado, por ejemplo, como una tabla de consulta. El intervalo de características que depende del parámetro auxiliar E_n^{MAX} es obtenido actualizando la tabla de consulta con E_n^{MAX} para cada nuevo marco. El valor E_n^{MAX} es determinado por un seguidor 24 de máximo de señal configurado para seguir el máximo de la señal, por ejemplo de acuerdo con la ecuación (5) anterior.

La figura 8 es un diagrama de bloques que ilustra un ejemplo de realización de un clasificador 18 de marcos en el clasificador 12 de audio de la figura 6. El calculador 26 de fracciones recibe las decisiones binarias (una decisión por cada intercalo de características) desde el comparador 16 de medidas de características y está configurado para calcular, para cada intervalo de características, una medición de fracción (en el ejemplo, $\Phi_1 - \Phi_2$) que representa el número total de mediciones de características correspondientes que caen dentro del intervalo de características. Un ejemplo de realización del calculador 26 de fracciones se ilustra en la figura 9. Estas medidas de fracciones son enviadas a un selector de clase 28 configurado para clasificar el último marco de audio como habla si cada medida de fracción se sitúa dentro de un correspondiente intervalo de fracciones, y como no-habla en caso contrario. Un ejemplo de realización del selector 28 de clase se ilustra en la figura 10.

La figura 9 es un diagrama de bloques que ilustra un ejemplo de realización de un calculador 26 de fracciones del clasificador 18 de marcos de la figura 8. Las decisiones binarias procedentes del comparador 16 de medidas de características son enviadas a una memoria temporal 30 de decisiones, que almacena las últimas *N* decisiones para cada intervalo de características. Un calculador 32 de fracción por cada intervalo de características determina cada medida de fracción contando el número de decisiones para la correspondiente característica que indica habla y dividiendo esta cuenta por el número total de decisiones *N*. Una ventaja de esta realización es que la memoria temporal de decisiones sólo tiene que almacenar decisiones binarias, lo que hace la ejecución sencilla y que reduce esencialmente el cálculo de fracciones a un proceso de cuenta sencillo.

La figura 10 es un diagrama de bloques que ilustra un ejemplo de realización de un selector de clase 28 del clasificador 18 de marcos de la figura 8. Las medidas de fracciones procedentes del calculador 26 de fracciones son enviadas a un calculador 34 de intervalos de fracciones, que está configurado para determinar si cada medida de fracción se sitúa dentro de un intervalo correspondiente de fracciones, y para dar salida a una decisión binaria correspondiente. Los intervalos de fracciones son obtenidos de un almacenamiento 36 de intervalos de fracciones, que almacena, por ejemplo, los intervalos de fracciones de la columna 7 de la Tabla 1 anterior. Las decisiones binarias procedentes del calculador 34 de intervalos de fracciones, son enviadas a una puerta lógica Y (AND) 38, que está configurada para clasificar el último marco como habla si todas ellas indican habla, y como no-habla en caso contrario.

Los pasos, funciones, procedimientos y/o bloques descritos en esta memoria pueden ser ejecutados en hardware utilizando cualquier tecnología convencional, tal como la tecnología de circuitos discretos o circuitos integrados, incluyendo tanto circuitos electrónicos de finalidad general como circuitos específicos de aplicaciones.

Alternativamente, algunos de los pasos, funciones, procedimientos y/o bloques descritos en esta memoria pueden ser realizados en software para su ejecución por un dispositivo de tratamiento adecuado, tal como un microprocesador, Procesador de Señales Digital (DSP) y/o cualquier dispositivo lógico programable apropiado, tal como un dispositivo de Serie de Puertas Programables de Campo (FPGA).

Se ha de entender también que puede ser posible reutilizar las capacidades de tratamiento general del codificador. Esto se puede hacer, por ejemplo, reprogramando el software existente o añadiendo nuevos componentes de software.

La figura 11 es un diagrama de bloques de un ejemplo de realización de un clasificador 12 de audio. Esta realización está basada en un procesador 100, por ejemplo un microprocesador, que ejecuta un componente 110 de software para determinar las medidas de características, un componente 120 de software para comparar medidas de características con intervalos de características, y un componente 130 de software para la clasificación de marcos. Estos componentes de software son almacenados en la memoria 150. El procesador 100 comunica con la memoria

a través de un bus del sistema. Las muestras de audio $x_m(n)$ son recibidas por un controlador 160 de entrada/salida (I/O) que controla un bus de entrada/salida, al cual están conectados el procesador 100 y la memoria 150. En esta realización, las muestras recibidas por el controlador 160 de I/O son almacenadas en la memoria 150, donde son tratadas por los componentes de software. El componente 110 de software puede ejecutar la funcionalidad del bloque 14 de las realizaciones descritas anteriormente. El componente 120 de software puede ejecutar la funcionalidad del bloque 16 de las realizaciones descritas anteriormente. El componente 130 de software puede ejecutar la funcionalidad del bloque 18 de las realizaciones descritas anteriormente. A la decisión de habla/no-habla, obtenida del componente 130 de software, se le da salida desde la memoria 150 por el controlador 160 de I/O por el bus de I/O.

La figura 12 es un diagrama de bloques que ilustra otro ejemplo de una disposición de codificador de audio que utiliza un clasificador 12 de audio. En esta realización, el codificador 10 comprende un codificador 50 de habla y un codificador 52 de música. El clasificador de audio controla un conmutador 54 que dirige las muestras de audio al codificador apropiado 50 ó 52.

La figura 13 es un diagrama de bloques que ilustra un ejemplo de una disposición de codec de audio que utiliza una decisión de habla/no-habla procedente del clasificador 12 de audio. Esta realización utiliza un post-filtro 60 para mejora del habla. La post-filtración se describe en [3] y [4]. En esta realización, la decisión de habla/no-habla procedente del clasificador 12 de audio es transmitida a un lado de recepción junto con la señal codificada procedente del codificador 10. La señal codificada es descodificada en un descodificador 60 y la señal descodificada es filtrada posteriormente en el post-filtro 62. La decisión de habla/no-habla es utilizada para seleccionar un método correspondiente de post-filtración. Además de seleccionar un método de post-filtración, la decisión de habla/no-habla puede ser usada también para seleccionar el método de codificación, como se indica con la línea de rayas hacia el codificador 10.

La figura 14 es un diagrama de bloques que ilustra un ejemplo de dispositivo de comunicación de audio que utiliza una disposición de codificador de audio de acuerdo con la presente tecnología. La figura ilustra una disposición 70 de codificador de audio en una estación de móvil. Un micrófono 72 está conectado a un bloque 74 amplificador y muestreador. Las muestras procedentes del bloque 74 son almacenadas en una memoria temporal 76 de marcos y son hechas seguir a la disposición 70 de codificador de audio en una base de marco-a-marco. Las señales codificadas son a continuación enviadas a una unidad de radio 78 para codificación de canal, modulación y amplificación de potencia. Las señales de radio obtenidas son finalmente transmitidas por una antena.

Aunque la mayor parte de las realizaciones ejemplares anteriores han sido ilustradas en el dominio de tiempo, se apreciará que pueden ser también ejecutadas en el dominio de frecuencia, por ejemplo para codificadores de transformación. En este caso, el extractor 14 de características estará basado, por ejemplo, en alguna de las ecuaciones (6)-(10). Sin embargo, una vez que han sido determinadas las medidas de características, se pueden utilizar los mismos elementos como en ejecuciones en el dominio de tiempo.

Con una realización basada en las ecuaciones (1), (2), (4), (5) y la Tabla 1, se obtuvo el siguiente rendimiento para clasificación de señales de audio:

% de habla clasificado erróneamente como música	5,9
% de música clasificado erróneamente como habla	1,8

La clasificación de audio descrita anteriormente es particularmente apropiada para sistemas que transmiten señales de audio codificadas en tiempo real. La información proporcionada por el clasificador puede ser usada para conmutar entre tipos de codificadores (por ejemplo un codificador de Predicción Lineal Excitado por Código (CELP) cuando se detecta una señal de habla , y un codificador de transformación, tal como n codificador de Transformación de Coseno Discreto Modificado (MDCT) cuando se detecta una señal de música), o parámetros codificadores. Además, las decisiones de clasificación pueden ser usadas también para controlar módulos de tratamiento específico de señales activas, tales como post-filtros de mejora de habla.

Sin embargo, la clasificación de audio descrita se puede usar también en aplicaciones fuera de línea, como una parte de un algoritmo de minería o extracción de datos, o módulos de tratamiento específicos de habla/música, tales como igualadores de frecuencias, control de estrépito, etc.

Los expertos en la técnica entenderán que se pueden hacer diversos cambios y modificaciones en la presente tecnología sin apartarse del alcance de la misma, que está definida por las reivindicaciones adjuntas.

Referencias

[1] E. Scheirer y M. Slaney, “"Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator”, ICASSP ’97 Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Volumen 2, páginas1331-1334, 1997.

[2] K. El-Maleh, M. Klein, G.Petrucci, P. Kabal, “"Speech/music discrimination for multimedia applications”, disponible

en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.3453&r ep=rep1&type=pdf.

[3] J-H. Chen, A. Gersho, “"Adaptive Postfiltering for Quality Enhancement of Coded Speech”, IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, Enero de 1993, páginas 59-71.

[4] WO 98/39768 A1

[5] US 7 127 392 B

Abreviaturas

CELP	Predicción Lineal Excitada por Código
DFT	Transformación de Fourier Discreta
DSP	Procesador de Señales Digital
FPGA	Serie de Puertas Programables de Campo
IDFT	Transformación de Fourier Discreta Inversa
LSFs	Frecuencias Espectrales de Línea
MDCT	Transformación de Coseno Discreta Modificada.

REIVINDICACIONES

1. Un método de clasificación de señales de audio basado en marcos o cuadros, **caracterizado por** los pasos de:

determinar (S1), para cada uno de un número predeterminado de marcos consecutivos, medidas de características que representan al menos las siguientes características:

- un coeficiente de auto-correlación (T_n),
- una energía de señal de marco (E_n) en un dominio comprimido,
- una variación de energía entre marcos;

comparar (S2) cada medida de característica determinada con al menos un correspondiente intervalo predeterminado de características; calcular (S3), para cada intervalo de características, una medida de fracción (Φ_1 – Φ_5) que representa el número total de medidas correspondientes de características (T_n, E_n, ΔE_n) que caen dentro del intervalo de características;

clasificar (S4) el último de los marcos consecutivos como habla si cada medida de fracción se sitúa dentro de un intervalo de fracción correspondiente, y como no-habla en caso contrario.

2. El método de la reivindicación 1, en el que las medidas de características que representan el coeficiente de auto-correlación (T_n) y la energía de señal de marco (E_n) en el dominio comprimido son determinadas en el dominio de tiempo.

3. El método de la reivindicación 2, en el que la medida de característica que representa el coeficiente de auto-correlación está dada por:

$$T_n = \frac{\sum_{m=1}^{M} x_m(n) x_{m-1}(n)}{\sum_{m=2}^{M} x_m^2(n)}$$

en la que

$x_m(n)$ indica muestra m en el marco n,

M es el número total de muestras en cada marco.

4. El método de la reivindicación 2 o la 3, en el que la medida de característica que representa la energía de la señal de marco en el dominio comprimido está dada por:

$$E_n = 10 \cdot \log_{10}\left(\frac{1}{M}\sum_{m=1}^{M} x_m^2(n)\right)$$

en la que

$x_m(n)$ indica muestra m

M es el número total de muestras en un marco.

5. El método de la reivindicación 1, en el que las medidas de características que representan el coeficiente de auto-correlación (T_n) y la energía de la señal de marco (E_n) en el dominio comprimido son determinadas en el dominio de frecuencia.

6. El método de cualquiera de las reivindicaciones precedentes 1-5, en el que la medida de característica que representa la variación de energía de la señal de marco entre marcos adyacentes está dada por:

$$\Delta E_n = \frac{\left\| E_n - E_{n-1} \right\|}{E_n - E_{n-1}}$$

7. El método de cualquiera de las reivindicaciones precedentes 1-6, que incluye el paso de determinar una medida de característica adicional que representa la variación espectral (SD_n) entre marcos.

8. El método de cualquiera de las reivindicaciones precedentes 1-7, que incluye el paso de terminar una medida

de característica adicional que representa la frecuencia fundamental $(\hat{P})$.

9. El método de cualquiera de las reivindicaciones precedentes 1-8, en el que un intervalo de características correspondiente a la energía (E_n) de la señal de marco en el dominio comprimido está dada por $\{0{,}62E_n^{MAX}, \Omega\}$, donde Ω es un límite superior de energía y E_n^{MAX} es un parámetro auxiliar dado por:

$$E_n^{MAX} = (1-\mu)E_{n-1}^{MAX} + \mu E_n$$

$$\mu = \begin{cases} 0{,}557 & si \quad E_n \geq E_{n-1}^{MAX} \\ 0{,}038 & si \quad E_n < E_{n-1}^{MAX} \\ 0{,}001 & si \quad E_n < 0{,}62E_{n-1}^{MAX} \end{cases}$$

donde E_n representa la energía de la señal de marco en el dominio comprimido en el marco *n*.

10. Un clasificador (12) de audio para la clasificación de señales de audio basada en marcos, **caracterizado por:**

un extractor (14) de características configurado para determinar, para cada uno de un número predeterminado de marcos consecutivos, medidas de características que representen al menos las siguientes características:

* un coeficiente de auto-correlación (T_n),

* energía (E_n) de la señal de marco en un dominio comprimido,

* variación de energía de la señal entre marcos;

un comparador (16) de medidas de características configurado para comparar cada medida de característica determinada (T_n, E_n, ΔE_n) con al menos un correspondiente intervalo de características predeterminado;

un clasificador (18) de marcos configurado para calcular, para cada intervalo de características, una medida de fracción ($\Phi_1 - \Phi_5$) que representa el número total de medidas de características correspondientes que caen dentro del intervalo de características, y para clasificar el último de los marcos consecutivos como habla si cada medida de fracción se sitúa dentro de un correspondiente intervalo de fracciones, y como no-habla en caso contrario.

11. El clasificador de audio de la reivindicación 10, en el que el extractor (14) de características está configurado para determinar las medidas de características que representan energía (E_n) de la señal de marco en el dominio comprimido y el coeficiente de auto-correlación (T_n) en el dominio de tiempo.

12. El clasificador de audio de la reivindicación 11, en el que el extractor (14) de características está configurado para determinar la medida de característica que representa el coeficiente de auto-correlación de acuerdo con:

$$T_n = \frac{\sum_{m=1}^{M} x_m(n)x_{m-1}(n)}{\sum_{m=2}^{M} x_m^2(n)}$$

en la que

$x_m(n)$ indica muestra *m* en el marco *n*

M es el número total de muestran en cada marco.

13. El clasificador de audio de la reivindicación 11 o la 12, en el que el extractor (14) de características está configurado para determinar la medida de características que representa la energía de la señal de marco en el dominio comprimido de acuerdo con:

$$E_n = 10 \cdot \log_{10}\left(\frac{1}{M}\sum_{m=1}^{M} x_m^2(n)\right)$$

en la que

$x_m(n)$ indica muestra m

M es el número total de muestras en un marco.

14. El clasificador de audio de la reivindicación 10, en el que el extractor (14) de características está configurado para determinar las medidas de características que representan la energía (E_n) de las señales de marco en el dominio comprimido y el coeficiente de auto-correlación (T_n) en el dominio de frecuencia.

15. El clasificador de audio de cualquiera de las reivindicaciones precedentes 10-14, en el que el extractor (14) de características está configurado para determinar la medida de características que representa la variación de energía entre marcos de acuerdo con:

$$\Delta E_n = \frac{\left\| E_n - E_{n-1} \right\|}{E_n + E_{n-1}}$$

en la que E_n representa la energía de la señal de marco en el dominio comprimido en el marco n.

16. El calificador de audio de cualquiera de las reivindicaciones precedentes 10-15, en el que el extractor (14) de características está configurado para determinar una medida de característica adicional que represente la frecuencia fundamental $(\hat{P})$.

17. El clasificador de audio de cualquiera de las reivindicaciones precedentes 10-16, en el que el comparador (16) de medidas de características está configurado (20, 22) para generar un intervalo $\left\{0{,}62E_n^{MAX}, \Omega\right\}$ de características correspondiente a la energía (E_n) de la señal de marco en el dominio comprimido, donde Ω es un límite superior de energía y E_n^{MAX} es un parámetro auxiliar dado por:

$$E_n^{MAX} = (1-\mu)E_{n-1}^{MAX} + \mu E_n$$

$$\mu = \begin{cases} 0{,}557 & si \quad E_n \geq E_{n-1}^{MAX} \\ 0{,}038 & si \quad E_n < E_{n-1}^{MAX} \\ 0{,}001 & si \quad E_n < 0{,}62E_{n-1}^{MAX} \end{cases}$$

donde E_n representa la energía de la señal de marco en el dominio comprimido en el marco n.

18. El clasificador de audio de cualquiera de las reivindicaciones precedentes 10-17, en el que el clasificador (18) de marcos incluye un calculador (26) de fracciones configurado para calcular, para cada intervalo de características, una medida de fracción ($\Phi_1 - \Phi_5$) que representa el número total de medidas de características correspondientes que caen dentro del intervalo de características;

un selector (28) de clase configurado para clasificar el último de los marcos consecutivos como habla si cada medida de fracción se sitúa dentro de un intervalo de fracción correspondiente, y como no-habla en caso contrario.

19. Una disposición de codificador de audio que incluye un clasificador (12) de audio de acuerdo con cualquiera de las reivindicaciones precedentes 10-18 para clasificar marcos de audio en habla/no-habla y seleccionar con ello un método de codificación correspondiente.

20. Un dispositivo de comunicación de audio que incluye una disposición (70) de codificador de audio de acuerdo con la reivindicación 19.

21. Una disposición de codificador-descodificador (codec) de audio que incluye un clasificador (12) de audio de acuerdo con cualquiera de las reivindicaciones precedentes 10-19 para clasificar marcos de audio en habla/no-habla para seleccionar un método de post-filtración correspondiente.

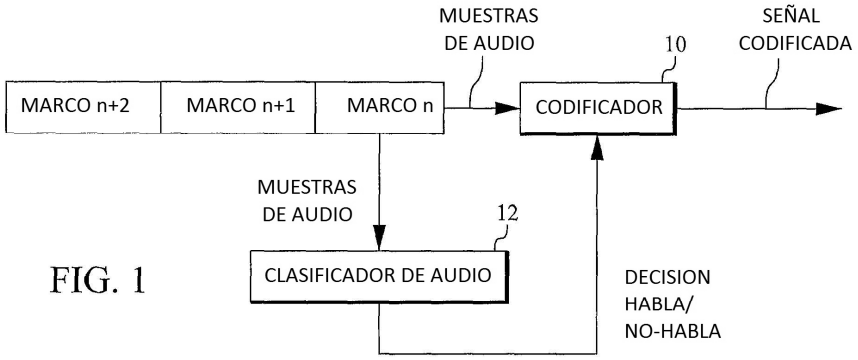


FIG. 1

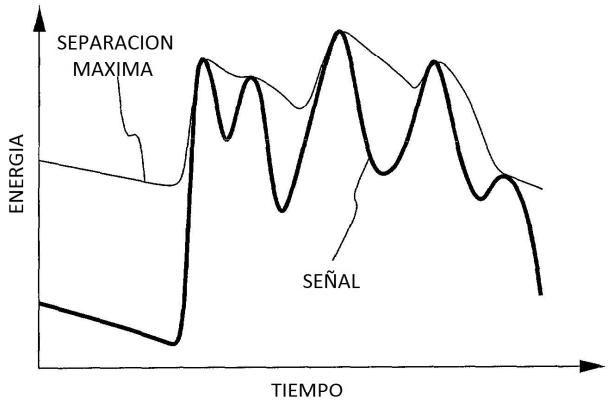


FIG. 2

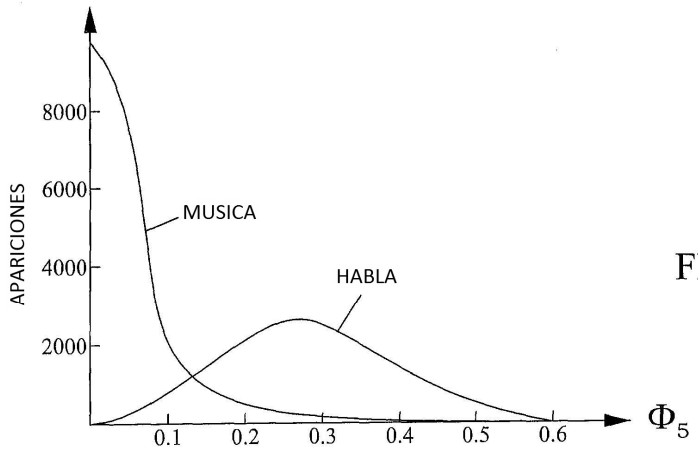


FIG. 3

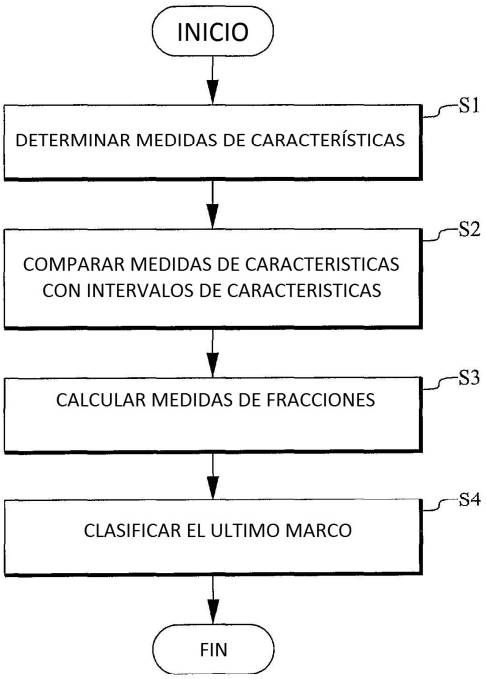

INICIO
DETERMINAR MEDIDAS DE CARACTERÍSTICAS
S1
COMPARAR MEDIDAS DE CARACTERISTICAS CON INTERVALOS DE CARACTERISTICAS
S2
CALCULAR MEDIDAS DE FRACCIONES
S3
CLASIFICAR EL ULTIMO MARCO
S4
FIN

FIG. 4

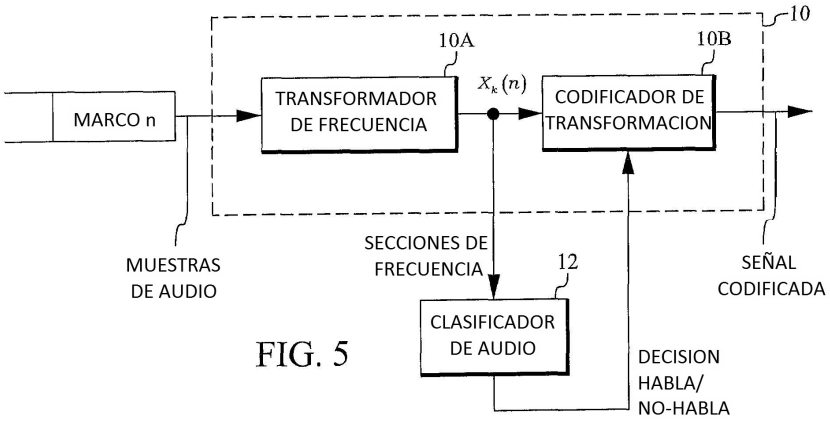

10
10A
10B
MARCO n
TRANSFORMADOR DE FRECUENCIA
$X_k(n)$
CODIFICADOR DE TRANSFORMACION
MUESTRAS DE AUDIO
SECCIONES DE FRECUENCIA
12
SEÑAL CODIFICADA
CLASIFICADOR DE AUDIO
DECISION HABLA/ NO-HABLA

FIG. 5

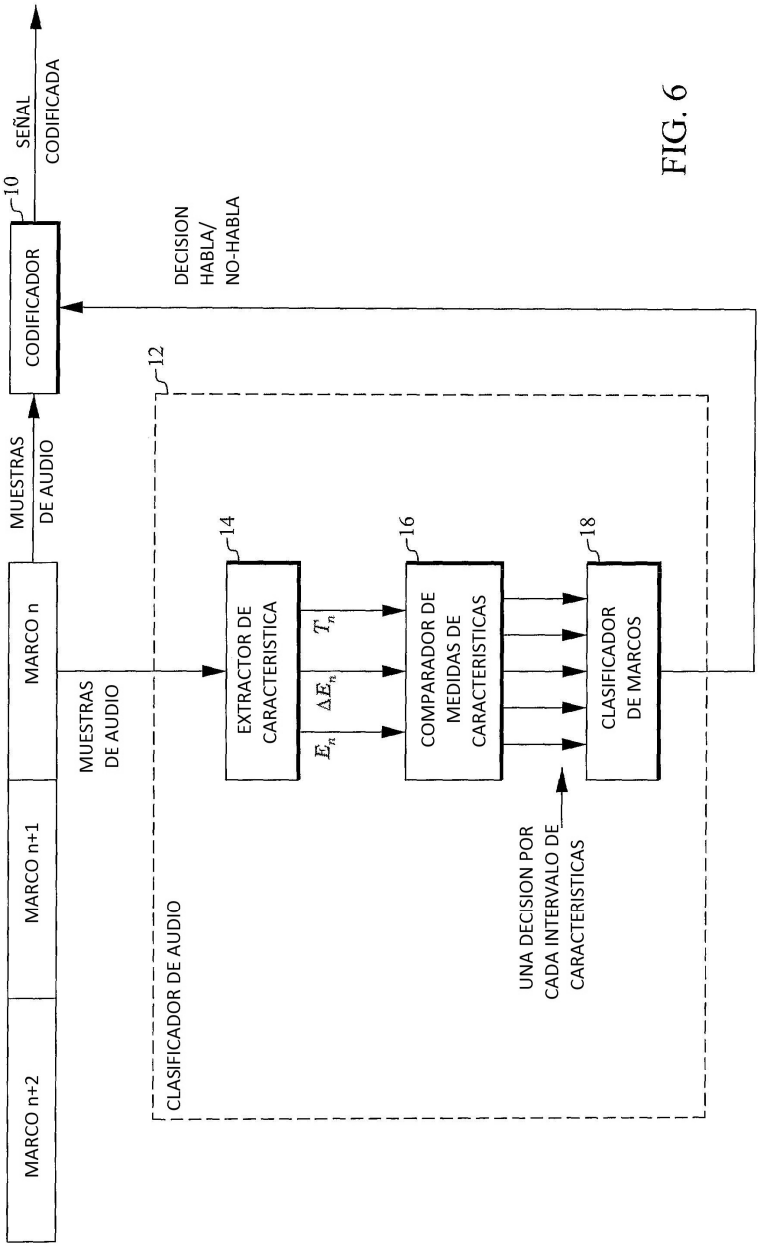

SEÑAL CODIFICADA
10
CODIFICADOR
DECISION HABLA/ NO-HABLA
12
MUESTRAS DE AUDIO
MARCO n
MARCO n+1
MARCO n+2
MUESTRAS DE AUDIO
14
EXTRACTOR DE CARACTERISTICA
T_n
ΔE_n
E_n
16
COMPARADOR DE MEDIDAS DE CARACTERISTICAS
18
CLASIFICADOR DE MARCOS
UNA DECISION POR CADA INTERVALO DE CARACTERISTICAS
CLASIFICADOR DE AUDIO

FIG. 6

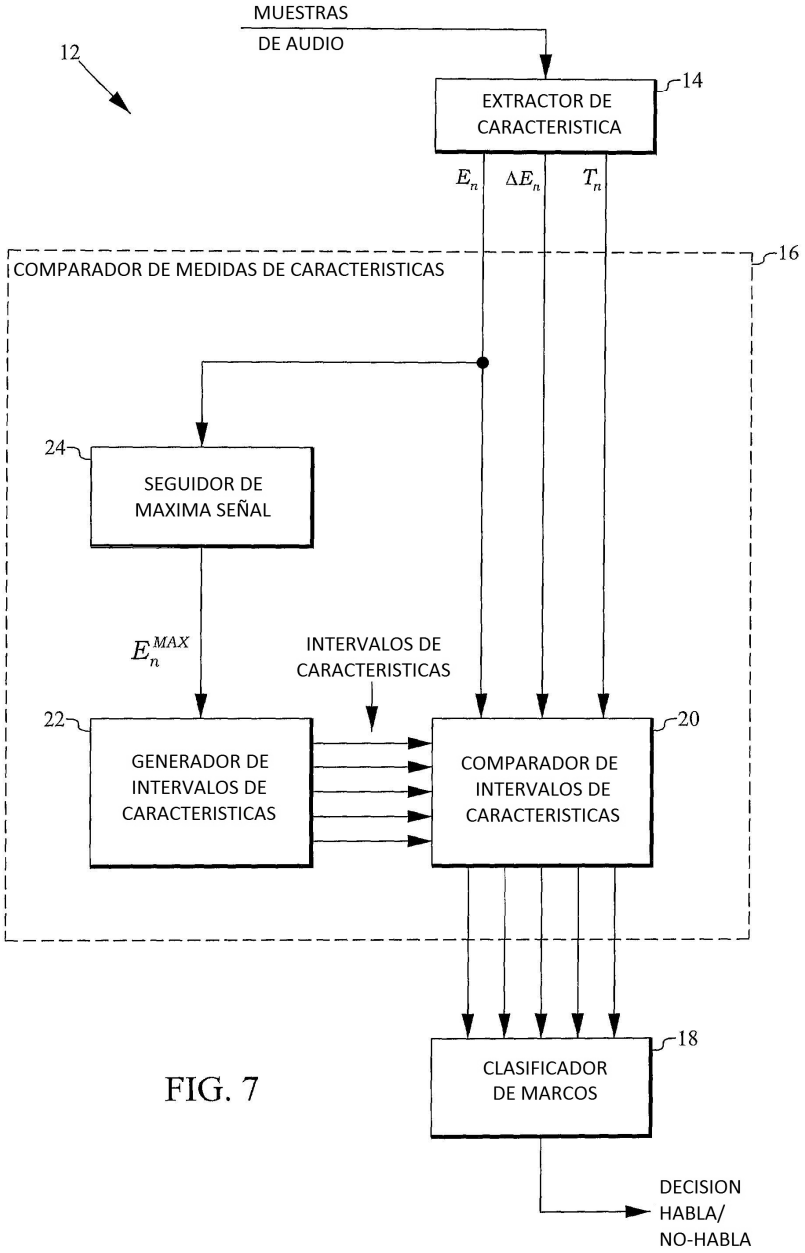

MUESTRAS
DE AUDIO
12
EXTRACTOR DE
CARACTERISTICA
14
E_n
ΔE_n
T_n
COMPARADOR DE MEDIDAS DE CARACTERISTICAS
16
24
SEGUIDOR DE
MAXIMA SEÑAL
E_n^{MAX}
INTERVALOS DE
CARACTERISTICAS
22
GENERADOR DE
INTERVALOS DE
CARACTERISTICAS
COMPARADOR DE
INTERVALOS DE
CARACTERISTICAS
20
CLASIFICADOR
DE MARCOS
18
DECISION
HABLA/
NO-HABLA

FIG. 7

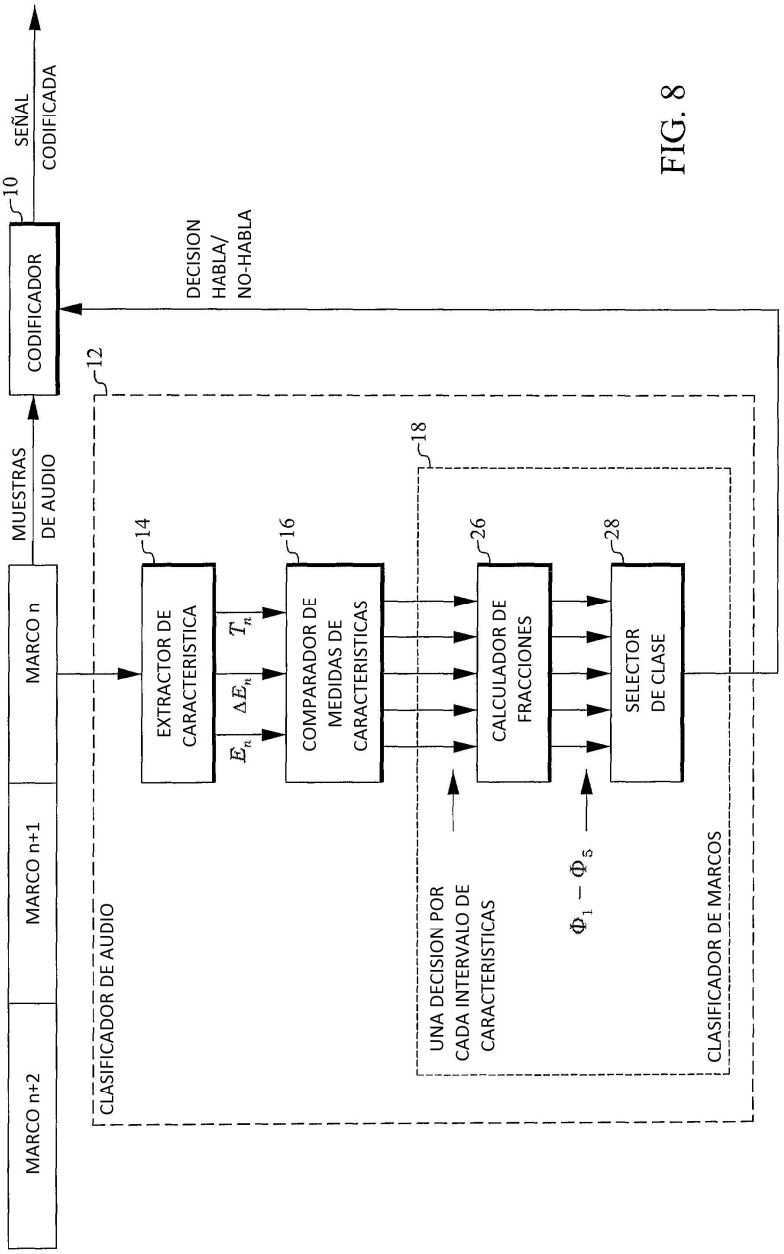

MARCO n+2
MARCO n+1
MARCO n
MUESTRAS DE AUDIO
CODIFICADOR
10
SEÑAL CODIFICADA
DECISION HABLA/ NO-HABLA
12
CLASIFICADOR DE AUDIO
14
EXTRACTOR DE CARACTERISTICA
E_n
ΔE_n
T_n
16
COMPARADOR DE MEDIDAS DE CARACTERISTICAS
18
UNA DECISION POR CADA INTERVALO DE CARACTERISTICAS
26
CALCULADOR DE FRACCIONES
$\Phi_1 - \Phi_5$
28
SELECTOR DE CLASE
CLASIFICADOR DE MARCOS

FIG. 8

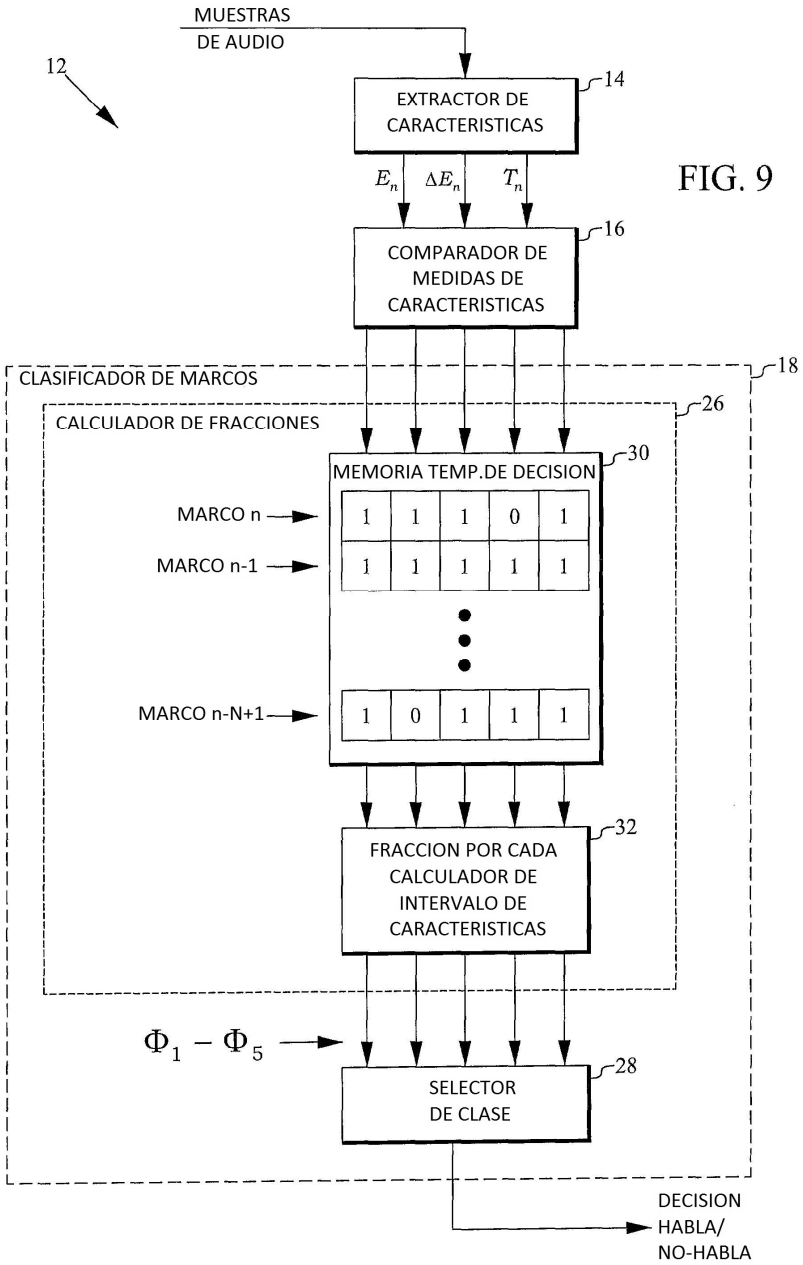

MUESTRAS
DE AUDIO
12
EXTRACTOR DE
CARACTERISTICAS
14
E_n
ΔE_n
T_n
COMPARADOR DE
MEDIDAS DE
CARACTERISTICAS
16
CLASIFICADOR DE MARCOS
18
CALCULADOR DE FRACCIONES
26
MEMORIA TEMP.DE DECISION
30
MARCO n
MARCO n-1
MARCO n-N+1
1 1 1 0 1
1 1 1 1 1
1 0 1 1 1
FRACCION POR CADA
CALCULADOR DE
INTERVALO DE
CARACTERISTICAS
32
$\Phi_1 - \Phi_5$
SELECTOR
DE CLASE
28
DECISION
HABLA/
NO-HABLA

FIG. 9

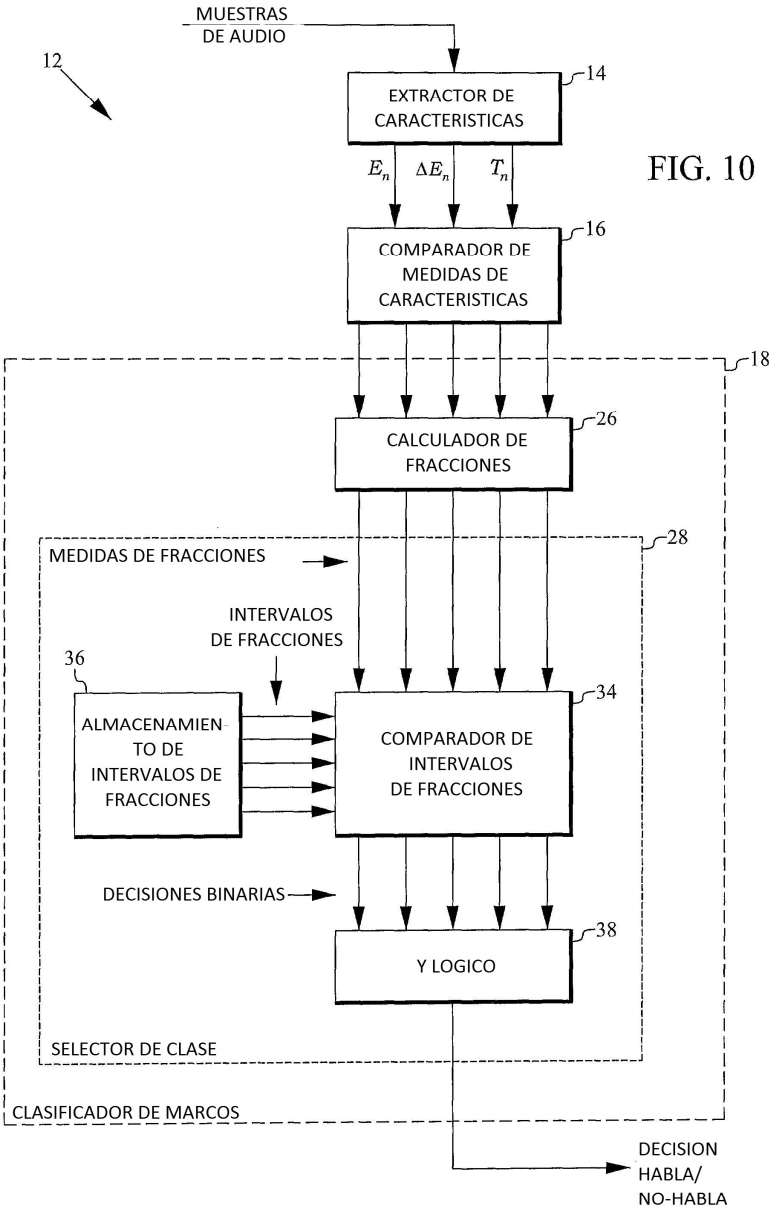

MUESTRAS DE AUDIO
12
14
EXTRACTOR DE CARACTERISTICAS
E_n
ΔE_n
T_n
16
COMPARADOR DE MEDIDAS DE CARACTERISTICAS
18
26
CALCULADOR DE FRACCIONES
28
MEDIDAS DE FRACCIONES
INTERVALOS DE FRACCIONES
36
ALMACENAMIENTO DE INTERVALOS DE FRACCIONES
34
COMPARADOR DE INTERVALOS DE FRACCIONES
DECISIONES BINARIAS
38
Y LOGICO
SELECTOR DE CLASE
CLASIFICADOR DE MARCOS
DECISION HABLA/ NO-HABLA

FIG. 10

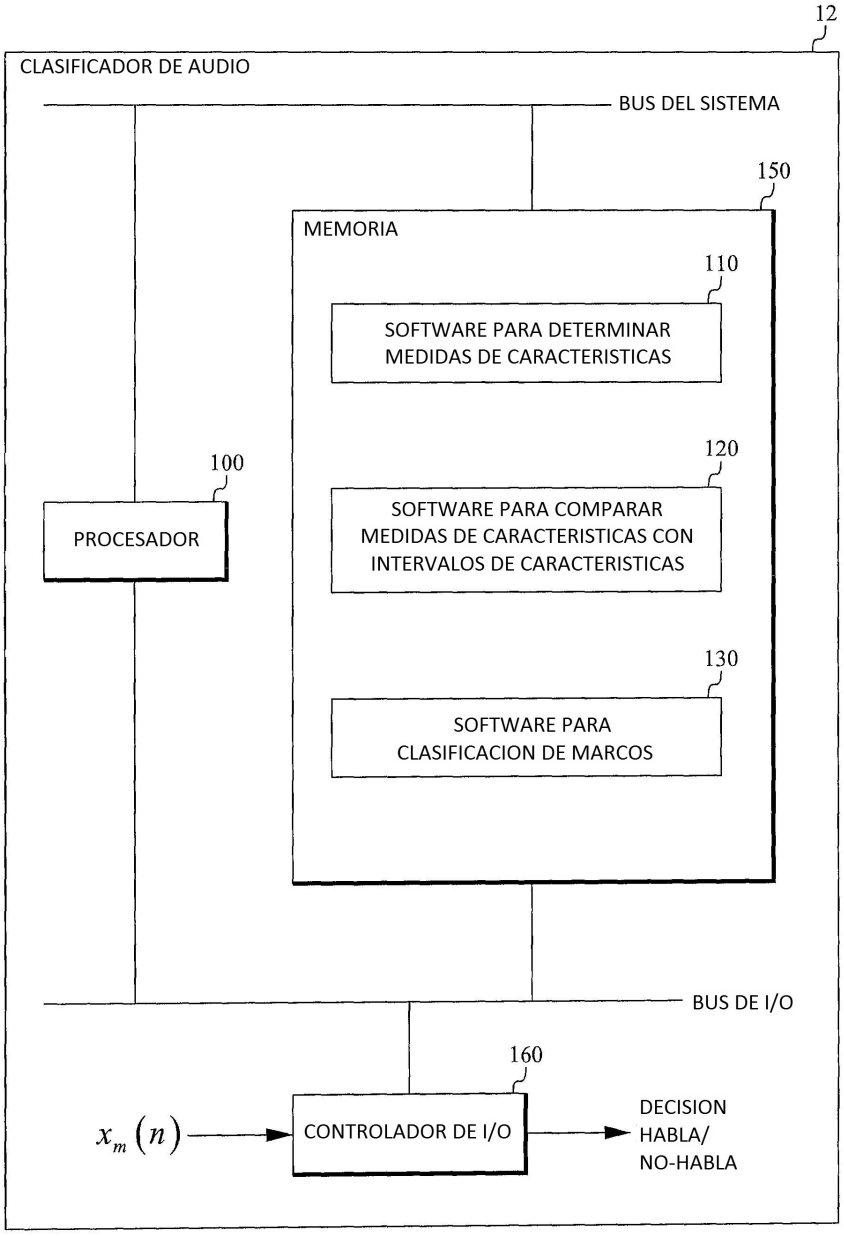
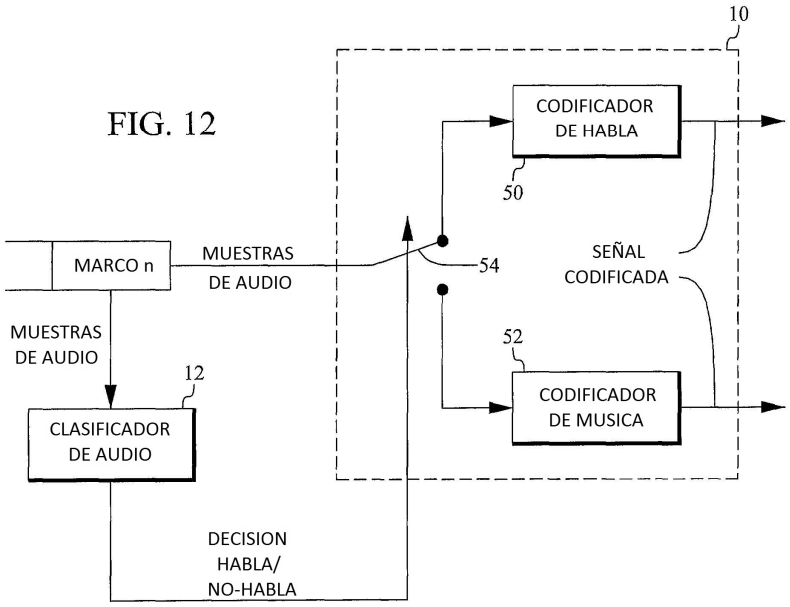

12
CLASIFICADOR DE AUDIO
BUS DEL SISTEMA
150
MEMORIA
110
SOFTWARE PARA DETERMINAR MEDIDAS DE CARACTERISTICAS
120
SOFTWARE PARA COMPARAR MEDIDAS DE CARACTERISTICAS CON INTERVALOS DE CARACTERISTICAS
130
SOFTWARE PARA CLASIFICACION DE MARCOS
100
PROCESADOR
BUS DE I/O
160
$x_m(n)$
CONTROLADOR DE I/O
DECISION HABLA/ NO-HABLA

FIG. 11

FIG. 12


10
CODIFICADOR DE HABLA
50
MARCO n
MUESTRAS DE AUDIO
54
SEÑAL CODIFICADA
MUESTRAS DE AUDIO
52
12
CODIFICADOR DE MUSICA
CLASIFICADOR DE AUDIO
DECISION HABLA/ NO-HABLA

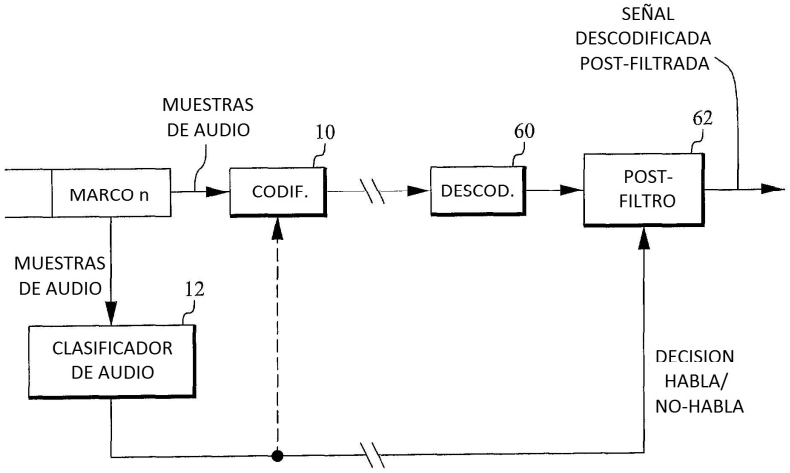

SEÑAL DESCODIFICADA POST-FILTRADA
MUESTRAS DE AUDIO
10
60
62
MARCO n
CODIF.
DESCOD.
POST-FILTRO
MUESTRAS DE AUDIO
12
CLASIFICADOR DE AUDIO
DECISION HABLA/ NO-HABLA

FIG. 13

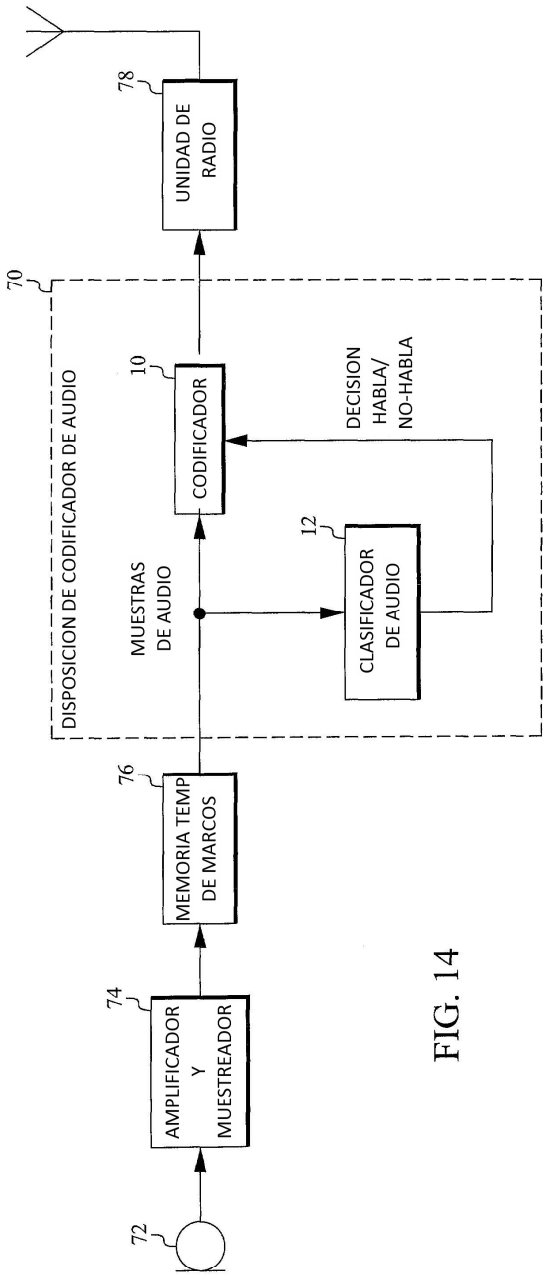

78
UNIDAD DE RADIO
70
DISPOSICION DE CODIFICADOR DE AUDIO
10
CODIFICADOR
DECISION HABLA/ NO-HABLA
12
CLASIFICADOR DE AUDIO
MUESTRAS DE AUDIO
76
MEMORIA TEMP DE MARCOS
74
AMPLIFICADOR Y MUESTREADOR
72

FIG. 14