(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 371 094**

(51) Int. CI.:
G10L 15/06 (2006.01)
G10L 15/14 (2006.01)
G10L 15/12 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **07014802 .8**
(96) Fecha de presentación: **22.03.2002**
(97) Número de publicación de la solicitud: **1850324**
(97) Fecha de publicación de la solicitud: **31.10.2007**

(54) Título: **SISTEMA DE RECONOCIMIENTO DE LA VOZ QUE USA ADAPTACIÓN IMPLÍCITA AL ORADOR.**

(30) Prioridad:
28.03.2001 US 821606

(45) Fecha de publicación de la mención BOPI:
27.12.2011

(45) Fecha de la publicación del folleto de la patente:
27.12.2011

(73) Titular/es:
QUALCOMM, INCORPORATED
5775 MOREHOUSE DRIVE
SAN DIEGO, CA 92121-1714, US

(72) Inventor/es:
Malayath, Narendranath;
Dejaco, Andrew P.;
Chang, Chienchung;
Jalil, Suhail;
Bi, Ning y
Garudadri, Harinath

(74) Agente: **Carpintero López, Mario**

DESCRIPCIÓN

Sistema de reconocimiento de la voz que usa adaptación implícita al orador

ANTECEDENTES

Campo

La presente invención se refiere al procesamiento de señales de voz. De manera particular, la presente invención se refiere a un procedimiento y aparato novedosos de reconocimiento de voz para conseguir prestaciones mejoradas a través de un entrenamiento sin supervisión.

Antecedentes

El reconocimiento de voz representa una de las técnicas más importantes para dotar a una máquina con inteligencia simulada para reconocer órdenes habladas del usuario y para facilitar la interfaz humana con la máquina. Los sistemas que emplean técnicas para recuperar un mensaje lingüístico a partir de una señal de voz acústica, se denominan sistemas de reconocimiento de voz (VR). La **FIG. 1** muestra un sistema de VR básico que tiene un filtro **102** de preénfasis, una unidad **104** de extracción de características acústicas (AFE) y un motor **110** de apareo de patrones. La unidad AFE **104** convierte una serie de muestras digitales de voz en un conjunto de valores de medida (por ejemplo, las componentes de frecuencia extraídas) denominados un vector de características acústicas. El motor **110** de apareo de patrones aparea una serie de vectores de características acústicas con las plantillas contenidas en un modelo acústico **112** de VR. Los motores de apareo de patrones de VR emplean por lo general técnicas de distorsión dinámica en el tiempo (DTW) o el Modelo de Markov Oculto (HMM). Tanto DTW como HMM son bien conocidas en la técnica y se describen con detalle en el documento de Rabiner, L. R., y Juang, B. H., "FUNDAMENTOS DEL RECONOCIMIENTO DE VOZ", Prentice-Hall, 1993. Cuando una serie de características acústicas coincide con una plantilla en el modelo acústico **112**, la plantilla identificada se usa para generar un formato de salida deseado, tal como una secuencia identificada de palabras lingüísticas correspondientes a la voz de entrada.

Como se ha hecho notar anteriormente el modelo acústico **112** generalmente es un modelo HMM o un modelo DTW. Se puede pensar en un modelo acústico DTW como en una base de datos de plantillas asociadas con cada una de las palabras que necesitan ser reconocidas. En general, una plantilla DTW consiste en una secuencia de vectores de características que han sido promediados sobre muchas muestras de la palabra asociada. El apareo de patrones del DTW generalmente implica la localización de una plantilla almacenada que tiene una distancia mínima a la secuencia del vector de características de entrada que representa la voz de entrada. Una plantilla usada en un modelo acústico basado en HMM contiene una descripción estadística detallada de la emisión vocal de la voz asociada. En general, una plantilla HMM almacena una secuencia de vectores de media, vectores de varianza y un conjunto de probabilidades de transición. Estos parámetros se usan para describir las estadísticas de una unidad de voz y se estiman a partir de muchos ejemplos de la unidad de voz. El apareo de patrones de HMM generalmente implica la generación de una probabilidad para cada plantilla del modelo basada en la serie de vectores de características de entrada asociados a la voz de entrada. La plantilla que tenga la probabilidad más alta se selecciona como la emisión vocal de entrada más probable.

El "entrenamiento" se refiere al proceso de recoger muestras de voz de un segmento o sílaba particular del habla de uno o más oradores con el fin de generar plantillas en el modelo acústico **112**. Cada plantilla en el modelo acústico está asociada a una palabra o segmento de voz particulares denominados una clase de emisión vocal. Puede haber múltiples plantillas en el modelo acústico asociadas a la misma clase de emisiones vocales. "Prueba" se refiere al procedimiento para aparear las plantillas del modelo acústico con una secuencia de vectores de características extraídos de la voz de entrada. Las prestaciones de un sistema dado dependen en gran medida del grado de coincidencia entre la voz de entrada del usuario final y el contenido de la base de datos, y por ello de la coincidencia entre las plantillas de referencia creadas a través del entrenamiento y las muestras de voz usadas para la prueba de VR.

Los dos tipos comunes de entrenamiento son el entrenamiento supervisado y el entrenamiento no supervisado. En el entrenamiento supervisado, la clase de emisión vocal asociada a cada conjunto de vectores de características de entrenamiento se conoce a priori. Al hablante que proporciona la voz de entrada a menudo se le proporciona un guión de palabras o segmentos de voz correspondientes a las clases de emisión vocal predeterminadas. Los vectores de características resultantes de la lectura del guión pueden ser incorporados entonces en las plantillas del modelo acústico asociado a las clases de emisiones vocales correctas.

En el entrenamiento no supervisado, la clase de emisión vocal asociada a un conjunto de vectores de características de entrenamiento no se conoce a priori. La clase de emisión vocal debe ser identificada de manera correcta antes de que se pueda incorporar un conjunto de vectores de características de entrenamiento en la plantilla correcta del modelo acústico. En el entrenamiento no supervisado, un error al identificar la clase de emisión vocal para un conjunto de vectores de características de entrenamiento puede conducir a una modificación de la plantilla del modelo acústico erróneo. Dicho error por lo general degrada, en lugar de mejorar, el funcionamiento del reconocimiento de la voz. Con el fin de evitar dichos errores, cualquier modificación del modelo acústico basada en el entrenamiento no supervisado se debe hacer por lo general de una manera muy conservadora. Se incorpora un conjunto de vectores de características de entrenamiento en el modelo acústico solamente si hay una confianza relativamente alta de que la clase de emisión vocal se ha identificado de manera correcta. Dicho conservadurismo necesario hace que la construcción de un modelo acústico SD a través de un entrenamiento no supervisado sea un proceso muy lento. Hasta que se construya de esta manera el modelo acústico SD, las prestaciones de la VR probablemente serán inaceptables para la mayoría de los usuarios.

De manera óptima, el usuario final proporciona vectores de características acústicas de voz tanto durante el

entrenamiento como durante la prueba, de forma que el modelo acústico **112** coincidirá notablemente con la voz del usuario final. Un modelo acústico individualizado que esté personalizado para un único hablante también se denomina un modelo acústico dependiente del orador (SD). La generación de un modelo acústico SD por lo general requiere que el usuario final proporcione una gran cantidad de muestras de entrenamiento supervisadas. Primero, el usuario debe proporcionar muestras de entrenamiento para una gran variedad de clases de formas de habla. Además, con el fin de conseguir las mejores prestaciones, el usuario final debe proporcionar múltiples plantillas, que representen una variedad de posibles entornos acústicos para cada clase de emisión vocal. Debido a que la mayoría de los usuarios son incapaces o están poco dispuestos a proporcionar la voz de entrada necesaria para generar un modelo acústico SD, muchos sistemas de VR existentes, en lugar de esto, usan modelos acústicos generalizados que están entrenados usando la voz de muchos hablantes "representativos". Se hace referencia a dichos modelos acústicos como modelos acústicos independientes del orador (SI), y están diseñados para tener las mejores prestaciones sobre una amplia gama de usuarios. Los modelos acústicos SI, sin embargo, pueden no estar optimizados para cualquier usuario único. Un sistema de VR que use un modelo acústico SI no funcionará tan bien para un usuario específico como un sistema de VR que use un modelo acústico SD personalizado para ese usuario. Para algunos usuarios, tales como los que tienen fuertes acentos extranjeros, las prestaciones de un sistema de VR que use un modelo acústico SI pueden ser tan pobres que no puedan usar de manera efectiva los servicios de VR en absoluto.

De manera óptima, un modelo acústico SD sería generado para cada usuario independiente. Como se ha expuesto con anterioridad, la construcción de modelos acústicos SD que usen entrenamiento supervisado no es práctica. Pero el uso de entrenamiento no supervisado para generar un modelo acústico SD puede llevar mucho tiempo, durante el que las prestaciones de VR basadas en un modelo acústico SD parcial pueden ser muy pobres. Existe una necesidad en la técnica de un sistema de VR que funcione razonablemente bien antes y durante la generación de un modelo acústico SD usando el entrenamiento no supervisado.

Se reclama atención al documento EP-A-1 011 094, que se refiere a un procedimiento para impedir la adaptación a palabras mal reconocidas en sistemas de reconocimiento automático de voz en línea o no supervisados. Se usan medidas de confianza o bien la reacción del usuario se interpreta para decidir si debería usarse o no un fonema reconocido, varios fonemas, una palabra, varias palabras o una emisión vocal completa para la adaptación del conjunto modelo independiente del orador al conjunto modelo adaptado al orador y, en caso de que se ejecute una adaptación, con cuánta energía debería realizarse la adaptación con esta emisión vocal reconocida, o parte de esta emisión vocal reconocida. Además, se propone una verificación de las prestaciones de la adaptación al orador para asegurar que la tasa de reconocimiento nunca disminuye, sino que sólo aumenta o permanece en el mismo nivel.

Resumen

De acuerdo a la presente invención, se proporciona un procedimiento para realizar el reconocimiento de la voz, según lo que se declara en la reivindicación 1. Realizaciones adicionales se reivindican en las reivindicaciones dependientes.

Los procedimientos y aparatos descritos en este documento están dirigidos a un novedoso y mejorado sistema de reconocimiento de voz (VR) que utiliza una combinación de modelos acústicos independientes del orador (SI) y dependientes del orador (SD). Al menos un modelo acústico SI se usa en combinación con al menos un modelo acústico SD para proporcionar un nivel prestaciones del reconocimiento de voz que al menos iguale el de un modelo acústico puramente SI. El sistema híbrido revelado SI / SD de VR usa de manera continua entrenamiento no supervisado para actualizar las plantillas acústicas en uno o más modelos acústicos SD. El sistema híbrido de VR usa después los modelos acústicos SD actualizados, solos o en combinación con al menos un modelo acústico SI, para proporcionar prestaciones de VR mejoradas durante la prueba del VR.

La palabra "ejemplar" se usa en este documento para significar "que sirve como un ejemplo, caso o ilustración". Cualquier realización descrita como una "realización ejemplar" no ha de interpretarse necesariamente como preferida o ventajosa con respecto a otra realización.

Breve descripción de los dibujos

Las características, objetos y ventajas del procedimiento y del aparato descritos en la presente se harán más evidentes a partir de la siguiente descripción detallada expuesta a continuación, cuando se considere junto con los dibujos, en los que idénticos caracteres de referencia identifican de manera correspondiente en todo el documento, y en los que:

la **FIG. 1** muestra un sistema básico de reconocimiento de voz;

la **FIG. 2** muestra un sistema de reconocimiento de voz de acuerdo a una realización ejemplar;

la **FIG. 3** muestra un procedimiento para realizar el entrenamiento no supervisado.

la **FIG. 4** muestra un enfoque ejemplar para generar una plantilla de apareo combinada usada en el entrenamiento no supervisado.

la **FIG. 5** es un diagrama de flujo que muestra un procedimiento para realizar el reconocimiento (prueba) de voz usando plantillas de apareo tanto independientes del orador (SI) como dependientes del orador (SD);

la **FIG. 6** muestra un enfoque para generar una plantilla de apareo combinada a partir tanto de plantillas de apareo independientes del orador (SI) como dependientes del orador (SD); y

Descripción detallada

La **FIG. 2** muestra una realización ejemplar de un sistema de reconocimiento de voz (VR) híbrido como se podría implementar dentro de una estación remota inalámbrica **202**. En una realización ejemplar, la estación remota **202** se comunica a través de un canal inalámbrico (que no se muestra) con una red de comunicaciones inalámbrica (que no se muestra). Por ejemplo, la estación remota **202** puede ser un teléfono inalámbrico que comunique con un sistema de teléfono inalámbrico. Alguien que sea experto en la técnica reconocerá que las técnicas descritas en este documento se pueden aplicar de igual manera a un sistema de VR que sea fijo (no portátil) o que no implique un canal inalámbrico.

En la realización mostrada, las señales de voz de un usuario se convierten en señales eléctricas en un micrófono (MIC) **210** y se convierten en muestras digitales de voz en un convertidor de analógico a digital (ADC) **212**. El flujo de muestras digitales se filtra después usando un filtro **214** de preénfasis (PE), por ejemplo un filtro de respuesta finita al impulso (FIR) que atenúa los componentes de la señal de baja frecuencia.

Las muestras filtradas se analizan después en una unidad **216** de extracción de las características acústicas (AFE). La unidad AFE **216** convierte las muestras digitales de voz en vectores de características acústicas. En una realización ejemplar, la unidad AFE **216** realiza una transformada de Fourier sobre un segmento de muestras digitales consecutivas para generar un vector de intensidades de señal correspondientes a diferentes compartimentos de frecuencia. En una realización ejemplar, los compartimentos de frecuencia tienen anchos de banda variables de acuerdo a una escala de tonos del habla. En una escala de tonos del habla, el ancho de banda de cada compartimento de frecuencia soporta una relación respecto a la frecuencia central del compartimento, de forma que los compartimentos de frecuencia más alta tienen bandas de frecuencia más anchas que los compartimentos de frecuencia más baja. La escala de tonos del habla se describe en el documento de Rabiner, L. R. y Juang, B. H., “FUNDAMENTOS DEL RECONOCIMIENTO DE VOZ”, Prentice-Hall, 1993 y es bien conocida en la técnica.

En una realización ejemplar, cada vector de características acústicas se extrae de una serie de muestras de voz recogidas durante un intervalo de tiempo fijo. En una realización ejemplar, estos intervalos de tiempo se solapan. Por ejemplo, las características acústicas se pueden obtener a partir de intervalos de 20 ms de datos de voz que comienzan cada diez milisegundos, de forma que cada dos intervalos consecutivos comparten un segmento de 10 ms. Alguien que sea experto en la técnica reconocería que, en lugar de esto, los intervalos de tiempo podrían no ser solapados o podrían tener una duración no fijada sin apartarse del alcance de las realizaciones descritas en este documento.

Los vectores de características acústicas generados por la unidad AFE **216** se entregan a un motor **220** de VR, que realiza el apareo de patrones para caracterizar el vector de características acústicas en base al contenido de uno o más modelos acústicos, **230**, **232** y **234**.

En la realización ejemplar mostrada en la **FIG. 2**, se muestran tres modelos acústicos: un modelo 230 de Markov Oculto (HMM) independiente del orador, un modelo 232 de Distorsión Dinámica en el Tiempo (DTW) independiente del orador, y un modelo acústico 234 dependiente del orador (SD). Alguien que sea experto en la técnica reconocerá que se pueden usar diferentes combinaciones de modelos acústicos SI en realizaciones alternativas. Por ejemplo, una estación remota 202 podría incluir solamente el modelo acústico SIHMM 230 y el modelo acústico SD 234 y omitir el modelo acústico SIDTW 232. De manera alternativa, una estación remota 202 podría incluir un único modelo acústico SIHMM 230, un modelo acústico SD 234 y dos modelos acústicos SIDTW diferentes 232. Además, alguien que sea experto en la técnica reconocerá que el modelo acústico SD 234 puede ser del tipo HMM o del tipo DTW o una combinación de los dos. En una realización ejemplar, el modelo acústico SD 234 es un modelo acústico DTW.

Como se ha descrito anteriormente, el motor **220** de VR realiza un apareo de patrones para determinar el grado de coincidencia entre los vectores de características acústicas y el contenido de uno o más modelos acústicos **230**, **232** y **234**. En una realización ejemplar, el motor **220** de VR genera plantillas de apareo en base al apareo de los vectores de características acústicas con las diferentes plantillas acústicas en cada uno de los modelos acústicos **230**, **232** y **234**. Por ejemplo, el motor **220** de VR genera plantillas de apareo HMM en base a un conjunto de vectores de características acústicas con múltiples plantillas HMM en el modelo acústico SIHMM **230**. Igualmente, el motor **220** de VR genera plantillas de apareo DTW en base al apareo de los vectores de características acústicas con múltiples plantillas DTW en el modelo acústico SIDTW **232**. El motor **220** de VR genera plantillas de apareo en base al apareo de los vectores de características acústicas con las plantillas del modelo acústico SD **234**.

Como se ha descrito anteriormente, cada plantilla en un modelo acústico está asociada a una clase de emisión vocal. En una realización ejemplar, el motor **220** de VR combina plantillas para las plantillas asociadas a la misma clase de emisión vocal para crear una plantilla de apareo combinada, para su uso en el entrenamiento sin supervisión. Por ejemplo, el motor **220** de VR combina plantillas de SIHMM y SIDTW obtenidas a partir de la correlación de un conjunto de entrada de vectores de características acústicas para generar una plantilla SI combinada. En base a esa plantilla de apareo combinada, el motor **220** de VR determina si almacena o no el conjunto de entrada de vectores de características acústicas como una plantilla SD en el modelo acústico SD **234**. En una realización ejemplar, el entrenamiento sin supervisión para actualizar el modelo acústico **234** se realiza usando exclusivamente plantillas de apareo SI. Esto evita errores aditivos que en cualquier otro caso podrían resultar del uso de un modelo acústico SD evolutivo **234** para el autoentrenamiento sin supervisión. A continuación se describe con mayor detalle un procedimiento ejemplar para realizar este entrenamiento sin supervisión.

Además del entrenamiento sin supervisión, el motor **220** de VR usa los diversos modelos acústicos (**230**, **232**, **234**) durante la prueba. En una realización ejemplar, el motor **220** de VR recupera las plantillas de apareo de los modelos acústicos (**230**, **232**, **234**) y genera plantillas de apareo combinadas para cada clase de emisión vocal. Las plantillas de apareo combinadas se usan para seleccionar la clase de emisión vocal que mejor se adapte a la voz de entrada.

El motor **220** de VR agrupa entre sí las clases de emisión vocal consecutivas según sea necesario para reconocer palabras o frases completas. El motor **220** de VR proporciona entonces información acerca de la palabra o la frase reconocida a un procesador **222** de control, que usa la información para determinar la respuesta apropiada a la información o comando de voz. Por ejemplo, en respuesta a la palabra o frase reconocida, el procesador **222** de control puede proporcionar realimentación al usuario a través de una pantalla o de otra interfaz de usuario. En otro ejemplo, el procesador **222** de control puede enviar un mensaje a través de un módem inalámbrico **218** y una antena **224** a una red inalámbrica (que no se muestra), iniciando una llamada de un teléfono móvil a un número de teléfono de destino asociado a la persona cuyo nombre fue pronunciado y reconocido.

El módem inalámbrico **218** puede transmitir señales a través de cualquiera entre una gran variedad de tipos de canal inalámbrico, incluyendo CDMA, TDMA o FDMA. Además, el módem inalámbrico **218** puede ser sustituido por otros tipos de interfaces de comunicaciones que comunican sobre un canal inalámbrico sin apartarse del alcance de las realizaciones descritas. Por ejemplo, la estación remota **202** puede transmitir información de señalización a través de cualquiera entre una gran variedad de tipos de canal de comunicaciones, incluyendo módems de línea terrestre, T1/E1, RDSI, DSL, Ethernet o incluso pistas sobre una placa de circuitos impresos (PCB).

La **FIG. 3** es un diagrama de flujo que muestra un procedimiento ejemplar para realizar el entrenamiento sin supervisión. En la etapa **302**, se muestrean datos analógicos de voz en un convertidor de analógico a digital (ADC) (**212** en la **FIG. 2**). El flujo de muestras digitales se filtra después en la etapa **304** usando un filtro de preénfasis (PE) (**214** en la **FIG. 2**). En la etapa **306**, se extraen los vectores de características acústicas de entrada de las muestras filtradas en una unidad de extracción de características acústicas (AFE) (**216** en la **FIG. 2**). El motor de VR (**220** en la **FIG. 2**) recibe los vectores de características acústicas de entrada desde la unidad AFE **216** y realiza el apareo de patrones de los vectores de características acústicas de entrada frente al contenido de los modelos acústicos SI (**230** y **232** en la **FIG. 2**). En la etapa **308**, el motor 220 de VR genera plantillas de apareo a partir de los resultados del apareo de patrones. El motor **220** de VR genera plantillas de apareo SIHMM mediante el apareo de los vectores de características acústicas de entrada con el modelo acústico SIHMM **230**, y genera plantillas de apareo SIDTW mediante el apareo de los vectores de características acústicas de entrada con el modelo acústico SIDTW **232**. Cada plantilla acústica en los modelos acústicos SIHMM y SIDTW (**230** y **232**) está asociada a una clase de emisión vocal particular. En la etapa **310**, las plantillas SIHMM y SIDTW se combinan para formar plantillas de apareo combinadas.

La **FIG. 4** muestra la generación de plantillas de apareo combinadas para su uso en el entrenamiento sin supervisión. En la realización ejemplar mostrada, la plantilla de apareo combinada independiente del orador S_{COMB_SI} para una clase particular de emisión vocal es una suma ponderada de acuerdo a la ecuación 1, como se muestra, en la que:

$SIHMM_T$ es la plantilla de apareo SIHMM para la clase de emisión vocal de destino;

$SIHMM_{NT}$ es la siguiente mejor plantilla de apareo para una plantilla en el modelo acústico SIHMM que está asociada a una clase de emisión vocal no de destino (una clase de emisión vocal distinta a la clase de emisión vocal de destino);

$SIHMM_G$ es la plantilla de apareo SIHMM para la clase de emisión vocal “inservible”;

$SIDTW_T$ es la plantilla de apareo para la clase de emisión vocal de destino;

$SIDTW_{NT}$ es la siguiente mejor plantilla de apareo para una plantilla en el modelo acústico SIDTW que está asociada a una clase de emisión vocal no de destino; y

$SIDTW_G$ es la plantilla de apareo SIDTW para la clase de emisión vocal “inservible”.

Las distintas plantillas de apareo individuales $SIHMM_n$ y $SIDTW_n$ pueden verse como representantes de un valor de distancia entre una serie de vectores de características acústicas de entrada y una plantilla en el modelo acústico. Cuanto mayor es la distancia entre los vectores de características acústicas de entrada y una plantilla, mayor es la plantilla de apareo. Una coincidencia próxima entre una plantilla y los vectores de características acústicas de entrada produce una plantilla de apareo muy bajo. Si la comparación de una serie de vectores de características acústicas de entrada con dos plantillas asociadas a diferentes clases de emisión vocal produce dos plantillas de apareo que son casi iguales; entonces el sistema de VR puede ser incapaz de reconocer qué clase de emisión vocal es la “correcta”.

$SIHMM_G$ y $SIDTW_G$ son plantillas de apareo para las clases de emisión vocal “inservible”. La plantilla o las plantillas asociadas a la clase de emisión vocal inservible se denominan plantillas de información inservible y no corresponden a una palabra o frase específica. Por esta razón, tienden a estar igualmente incorrelacionadas con respecto a toda la voz de entrada. Las plantillas de apareo de información inservible son útiles como una clase de medición de fondo del ruido en un sistema de VR. Generalmente, una serie de vectores de características acústicas de entrada debería tener un grado mucho mejor de apareo con una plantilla asociada a una clase de emisión vocal de destino que con la plantilla de información inservible antes de que pueda reconocerse de manera segura la clase de emisión vocal.

Antes de que el sistema de VR pueda reconocer de manera segura una clase de emisión vocal como la “correcta”, los vectores de características acústicas de entrada deberían tener un grado más alto de coincidencia con las plantillas asociadas a esa clase de emisión vocal que con las plantillas de información inservible o con las plantillas asociadas a otras clases de emisión vocal. Las plantillas de apareo combinadas generadas a partir de una gran variedad de modelos acústicos pueden discriminar de manera segura entre clases de emisión vocal que las plantillas de apareo basadas solamente en un modelo acústico. En una realización ejemplar, el sistema de VR usa dichas plantillas de apareo combinadas para determinar si sustituye o no una plantilla en el modelo acústico SD (**234** en la **FIG. 2**) por una obtenida a partir de un nuevo conjunto de vectores de características acústicas de entrada.

Los factores de ponderación (W_1,..., W_6) se seleccionan para proporcionar las mejores prestaciones de entrenamiento sobre todos los entornos acústicos. En una realización ejemplar, los factores de ponderación (W_1,..., W_6) son constantes para todas las clases de emisión vocal. En otras palabras, el W_n usado para crear la plantilla de apareo combinada para una primera clase de emisión vocal de destino es el mismo que el valor W_n usado para crear la plantilla de apareo combinada para otra clase de emisión vocal de destino. En una realización alternativa, los factores de ponderación varían en base a la clase de emisión vocal de destino. Otras formas de combinación mostradas en la **FIG. 4** serán obvias para alguien que sea experto en la técnica, y deben verse dentro del alcance de las realizaciones descritas en este documento. Por ejemplo, se pueden usar también más de seis o menos de seis entradas ponderadas. Otra variación obvia sería generar una plantilla de apareo combinada basada en un tipo de modelo acústico. Por ejemplo, se podría generar una plantilla de apareo combinada en base a $SIHMM_T$, $SIHMM_{NT}$ y $SIHMM_G$. O se podría generar una plantilla de apareo combinada en base a $SIDTW_T$, $SIDTW_{NT}$ y $SIDTW_G$.

En una realización ejemplar, W_1 y W_4 son números negativos, y un valor mayor (o menos negativo) de S_{COMB} indica un mayor grado de coincidencia (distancia menor) entre una clase de emisión vocal de destino y una serie de vectores de características acústicas de entrada. Alguien que sea experto en la técnica apreciará que los signos de los factores de ponderación pueden ser fácilmente redispuestos de forma que un grado mayor de coincidencia corresponda a un valor menor sin apartarse del alcance de las realizaciones descritas.

Volviendo de nuevo a la **FIG. 3**, en la etapa **310**, se generan plantillas de apareo combinadas para las clases de emisión vocal asociadas a las plantillas en los modelos acústicos HMM y DTW (**230** y **232**). En una realización ejemplar, se generan plantillas de apareo solamente para las clases de emisión vocal asociadas a las mejores *n* plantillas de apareo SIHMM y para las clases de emisión vocal asociadas a las mejores *m* plantillas de apareo SIDTW. Este límite puede ser deseable para conservar recursos de cómputo, incluso aunque se consuma una cantidad mucho mayor de potencia de cómputo a la vez que se generan las plantillas de apareo individuales. Por ejemplo, si *n*=*m*=3, se generan plantillas de apareo combinados para las clases de emisión vocal asociadas a los tres primeros SIHMM y las clases de emisión vocal asociadas a las tres primeras plantillas de apareo SIDTW. Según que las clases de emisión vocal asociadas a las tres primeras plantillas de apareo SIHMM sean o no las mismas que las clases de emisión vocal asociadas a las tres primeras plantillas de apareo SIDTW, este enfoque producirá entre tres y seis plantillas distintas de apareo combinadas.

En la etapa **312**, la estación remota **202** compara las plantillas de apareo combinadas con las plantillas de apareo combinadas almacenadas con las correspondientes plantillas (asociadas a la misma clase de emisión vocal) en el modelo acústico SD. Si la nueva serie de vectores de características acústicas de entrada tiene un grado mayor de coincidencia que los de una plantilla más antigua almacenada en el modelo SD para la misma clase de emisión vocal, entonces se genera una nueva plantilla SD a partir de la nueva serie de vectores de características acústicas de entrada. En una realización en la que un modelo acústico SD sea un modelo acústico DTW, la serie de vectores de características acústicas de entrada por sí sola constituye la nueva plantilla SD. La vieja plantilla se sustituye entonces por la nueva plantilla y la plantilla de apareo combinada asociada a la nueva plantilla se almacena en el modelo acústico SD para su uso en comparaciones futuras.

En una realización alternativa, el entrenamiento sin supervisión se usa para actualizar una o más plantillas en un modelo acústico de Markov Oculto dependiente del orador (SDHMM). Este modelo acústico SDHMM se podría usar bien en lugar de un modelo SDHMM o bien además de un modelo acústico SDDTW dentro del modelo acústico SD **234**.

En una realización ejemplar, la comparación en la etapa **312** incluye también la comparación de la plantilla de apareo combinada de una nueva plantilla SD probable con un umbral de entrenamiento constante. Incluso si no hubiese todavía ninguna plantilla almacenada en un modelo acústico SD para una clase de emisión vocal particular, no se almacenará una nueva plantilla en el modelo acústico SD a menos que haya una plantilla de apareo combinada que sea mejor (que indique un grado mayor de coincidencia) que el valor umbral de entrenamiento.

En una realización alternativa, antes de que se haya sustituido cualquier plantilla del modelo acústico SD, el modelo acústico SD se puebla por omisión con plantillas del modelo acústico SI. Dicha inicialización proporciona una aproximación alternativa para asegurar que las prestaciones del VR que usa el modelo acústico SD comenzarán al menos tan bien como las prestaciones del VR que usa solamente el modelo acústico SI. Cuantas más y más plantillas del modelo acústico SD se actualicen, las prestaciones del VR que use el modelo acústico SD sobrepasarán las prestaciones del VR que usa sólo el modelo acústico SI.

En una realización alternativa, el sistema de VR permite a un usuario realizar el entrenamiento supervisado. El usuario debe poner al sistema de VR en una modalidad de entrenamiento supervisado antes de realizar dicho entrenamiento supervisado. Durante el entrenamiento supervisado, el sistema de VR tiene un conocimiento a priori de la clase de emisión vocal correcta. Si la plantilla de apareo combinada para la voz de entrada es mejor que la plantilla de apareo combinada para la plantilla SD anteriormente almacenada para esa clase de emisión vocal, entonces la voz de salida se usa para formar una plantilla SD sustituta. En una realización alternativa, el sistema de VR permite al usuario forzar la sustitución de las plantillas SD existentes durante el entrenamiento supervisado.

El modelo acústico SD puede ser diseñado con espacio para múltiples (dos o más) plantillas para una única clase de emisión vocal. En una realización ejemplar, se almacenan dos plantillas en el modelo acústico SD para cada clase de emisión vocal. La comparación en la etapa **312** supone por lo tanto la comparación de la plantilla de apareo obtenida con una nueva plantilla con las plantillas de apareo obtenidas para ambas plantillas en el modelo acústico SD para la misma clase de emisión vocal. Si la nueva plantilla tiene una mejor plantilla de apareo que cualquiera de las antiguas plantillas del modelo acústico SD, entonces en la etapa **314** la plantilla del modelo acústico SD que tenga la peor plantilla de apareo es sustituida por la nueva plantilla. Si la plantilla de apareo de la nueva plantilla no es mejor que cualquier plantilla antigua, entonces se omite la etapa **314**. De manera adicional, en la etapa **312**, la plantilla de apareo obtenida con la nueva plantilla se compara con un umbral de plantilla de apareo. Así, hasta que se almacenen en el modelo acústico SD nuevas plantillas que tengan una plantilla de apareo que sea mejor que el

umbral, las nuevas plantillas son comparadas con este valor de umbral antes de que sean usadas para sobrescribir el contenido anterior del modelo acústico SD. Se anticipan y se consideran dentro del alcance de las realizaciones descritas en este documento variaciones obvias, tales como el almacenamiento de las plantillas del modelo acústico SD en orden clasificado de acuerdo a la plantilla de apareo combinada y la comparación de nuevas plantillas de apareo solamente con las más bajas. También se anticipan variaciones obvias acerca de los números de plantillas almacenadas en el modelo acústico para cada clase de emisión vocal. Por ejemplo, el modelo acústico SD puede contener más de dos plantillas para cada clase de emisión vocal, o puede contener diferentes números de plantillas para diferentes clases de emisiones vocales.

La **FIG. 5** es un diagrama de flujo que muestra un procedimiento ejemplar para realizar la prueba del VR usando una combinación de modelos acústicos SI y SD. Las etapas **302**, **304**, **306** y **308** son las mismas que las descritas para la **FIG. 3**. El procedimiento ejemplar se diferencia del procedimiento mostrado en la **FIG. 3** en la etapa **510**. En la etapa **510**, el motor **220** de VR genera plantillas de apareo SD basadas en la comparación de los vectores de características acústicas de entrada con las plantillas en el modelo acústico SD. En una realización ejemplar, las plantillas de apareo SD se generan solamente para las clases de emisión vocal asociadas a las *n* mejores plantillas de apareo SIHMM y las *m* mejores plantillas de apareo SIDTW. En una realización ejemplar, $n = m = 3$. Según el grado de solapamiento entre los dos conjuntos de clases de emisiones vocales, esto dará como resultado la generación de plantillas de apareo SD para entre tres y seis clases de emisiones vocales. Como se ha expuesto con anterioridad, el modelo acústico SD puede contener múltiples plantillas para una clase de emisión vocal única. En la etapa **512**, el motor **220** de VR genera plantillas de apareo combinadas híbridas para su uso en la prueba del VR. En una realización ejemplar, estas plantillas de apareo combinadas híbridas se basan tanto en las plantillas individuales de apareo SI como en las plantillas individuales de apareo SD. En la etapa **514**, se selecciona la palabra o la emisión vocal que tiene la mejor plantilla de apareo combinada y se compara con un umbral de prueba. Una emisión vocal solamente se considera como reconocida si su plantilla de apareo combinada sobrepasa este umbral. En una realización ejemplar, los pesos $[W_1, ..., W_6]$ usados para generar plantillas combinadas para el entrenamiento (como se muestra en la **FIG. 4**) son iguales a los pesos $[W_1, ..., W_6]$ usados para generar plantillas combinadas para prueba (como se muestra en la **FIG. 6**), pero el umbral de entrenamiento no es igual al umbral de prueba.

La **FIG. 6** muestra la generación de plantillas de apareo combinadas híbridas realizada en la etapa **512**. La realización ejemplar mostrada funciona de manera idéntica al combinador mostrado en la **FIG. 4**, excepto que el factor de ponderación W_4 se aplica a DTW_T en lugar de a $SIDTW_T$ y el factor de ponderación W_5 se aplica a DTW_{NT} en lugar de $SIDTW_{NT}$. DTW_T (la plantilla de apareo de Distorsión Dinámica en el Tiempo para la clase de emisión vocal de destino) se selecciona a partir de la mejor de las plantillas SIDTW y SDDTW asociadas a la clase de emisión vocal de destino. De manera similar, DTW_{NT} (la plantilla de apareo de Distorsión Dinámica en el Tiempo para las restantes clases de emisión vocal no de destino) se selecciona a partir de las mejores plantillas SIDTW y SDDTW asociadas a clases de emisión vocal no de destino.

La plantilla híbrida SI/SD S_{COMB_H} para una clase de emisión vocal particular es una suma ponderada de acuerdo a la ecuación 2, como se muestra, donde $SIHMM_T$, $SIHMM_{NT}$, $SIHMM_G$ y $SIDTW_G$ son los mismos que en la ecuación 1. De manera específica, en la ecuación 2:

$SIHMM_T$ es la plantilla de apareo SIHMM para la clase de emisión vocal de destino;

$SIHMM_{NT}$ es la siguiente mejor plantilla de apareo para una plantilla en el modelo acústico SIHMM que está asociada a una clase de emisión vocal no de destino (una clase de emisión vocal distinta a la clase de emisión vocal de destino);

$SIHMM_G$ es la plantilla de apareo SIHMM para la clase de emisión vocal “inservible”;

DTW_T es la mejor plantilla de apareo DTW para las plantillas SI y SD correspondientes a las clases de emisión vocal de destino;

DTW_{NT} es la mejor plantilla de apareo DTW para las plantillas SI y SD correspondientes a las clases de emisión vocal no de destino; y

$SIDTW_G$ es la plantilla de apareo SIDTW para la clase de emisión vocal “inservible”.

De esta forma, la referencia híbrida SI/SD S_{COM_H} es una combinación de plantillas de apareo SI y SD individuales. La plantilla de apareo de combinación resultante no depende enteramente de ninguno de los modelos acústicos SI o SD. Si la plantilla de apareo $SIDTW_T$ es mejor que cualquier plantilla $SDDTW_T$, entonces la referencia híbrida SI/SD se calcula a partir de la mejor referencia $SIDTW_T$. De manera similar, si la plantilla de apareo $SDDTW_T$ es mejor que cualquier plantilla $SIDTW_T$, entonces la plantilla híbrida SI/SD se calcula a partir de la mejor plantilla $SDDTW_T$. Como resultado de esto, si las plantillas en el modelo acústico SD producen plantillas de apareo pobres, el sistema de VR puede reconocer aún la voz de entrada en base a las partes SI de las plantillas híbridas SI/SD. Dichas plantillas pobres de apareo SD podrían tener una gran variedad de causas, incluyendo diferencias entre entornos acústicos durante el entrenamiento y las pruebas, o quizá una entrada de pobre calidad usada para el entrenamiento.

En una realización alternativa, la plantillas SI son ponderadas de una manera menos pesada que las plantillas SD, o incluso pueden ser ignoradas por completo. Por ejemplo, DTW_T se selecciona entre las mejores plantillas SDDTW asociadas a la clase de emisión vocal de destino, ignorando las plantillas SIDTW para la clase de emisión vocal de destino. También, DTW_T se puede seleccionar a partir de las mejores plantillas bien de SIDTW o bien de SDDTW, asociadas a clases de emisión vocal no de destino, en lugar de usar ambos conjuntos de plantillas.

Aunque la realización ejemplar se describe usando solamente modelos acústicos SDDTW para la modelización dependiente del orador, la aproximación híbrida descrita en este documento es igualmente aplicable a un sistema de VR que use modelos acústicos SDHMM o incluso una combinación de modelos acústicos SDDTW y SDHMM. Por ejemplo, mediante la modificación de la aproximación mostrada en la FIG. 6, el factor de ponderación W_1 se podría

aplicar a una plantilla de apareo seleccionada entre las mejores plantillas $SIHMM_T$ y $SDHMM_T$. El factor de ponderación W_2 se podría aplicar a una plantilla de apareo seleccionada entre las mejores plantillas $SIHMM_{NT}$ y $SDHMM_{NT}$.

De esta forma, se describe en este documento un procedimiento y un aparato de VR que utilizan una combinación de modelos acústicos SI y SD para prestaciones mejoradas de VR durante el entrenamiento sin supervisión y la prueba. Los que sean expertos en la técnica comprenderán que la información y las señales pueden ser representadas usando cualquiera entre una gran variedad de diferentes tecnologías y técnicas. Por ejemplo, datos, instrucciones, órdenes, información, señales, bits, símbolos y segmentos a los que se puede hacer referencia a lo largo de toda la descripción anterior, pueden ser representados por medio de tensiones, corrientes, ondas electromagnéticas, campos magnéticos o partículas magnéticas, campos ópticos o partículas ópticas o cualquier combinación de los mismos. También, aunque las realizaciones se describen en primer lugar en términos de modelos acústicos tales como el modelo de Distorsión Dinámica en el Tiempo (DTW) o el modelo de Markov Oculto (HMM), las técnicas descritas se pueden aplicar a otros tipos de modelos acústicos tales como modelos acústicos de redes neuronales.

Los que sean expertos en la técnica apreciarán además que los varios bloques lógicos, módulos, circuitos y etapas de algoritmos ilustrativos descritos con respecto a las realizaciones reveladas en este documento se pueden implementar como hardware electrónico, software de ordenador o una combinación de ambos. Para ilustrar de manera clara esta capacidad de intercambiabilidad entre hardware y software, se han descrito anteriormente varios componentes, bloques, módulos, circuitos y etapas ilustrativos, generalmente en términos de su funcionalidad. Si se implementa dicha funcionalidad como hardware o software depende de la aplicación particular y de las restricciones del diseño impuestas sobre el sistema global. Los expertos pueden implementar la funcionalidad descrita de varias maneras para cada aplicación particular, pero dichas decisiones de implementación no se deberían interpretar como causantes de un alejamiento del alcance de la presente invención.

Los diversos bloques lógicos, módulos y circuitos ilustrativos descritos con relación a las realizaciones descritas en este documento se pueden implementar o realizar con un procesador de propósito general, un procesador de señales digitales (DSP), un circuito integrado específico de la aplicación (ASIC), una matriz de compuertas programable en el terreno (FPGA) u otro dispositivo lógico programable, compuerta discreta o lógica de transistores, componentes de hardware discretos o cualquier combinación de los mismos diseñada para realizar las funciones descritas en este documento. Un procesador de propósito general puede ser un microprocesador, pero como alternativa, el procesador puede ser cualquier procesador convencional, controlador, microcontrolador o máquina de estados. Un procesador también se puede implementar como una combinación de dispositivos de computación, por ejemplo, una combinación de un DSP y un microprocesador, una pluralidad de microprocesadores, uno o más microprocesadores junto con un núcleo de DSP o cualquier otra de tales configuraciones.

Las etapas de un procedimiento o algoritmo descrito con relación a las realizaciones descritas en este documento se pueden realizar directamente en hardware, en un módulo de software ejecutado por un procesador o en una combinación de los dos. Un módulo de software puede residir en memoria RAM, en memoria flash, en memoria ROM, en memoria EPROM, en memoria EEPROM, en registros, en disco duro, en un disco extraíble, en un CD-ROM o en cualquier otro formato de medio de almacenamiento conocido en la técnica. Un medio de almacenamiento ejemplar se acopla al microprocesador de forma tal que el microprocesador pueda leer la información de, y escribir información en, el medio de almacenamiento. Como alternativa, el medio de almacenamiento puede estar integrado en el procesador. El procesador y el medio de almacenamiento pueden residir en un ASIC. Como alternativa, el procesador y el medio de almacenamiento pueden residir como componentes discretos en un terminal de usuario.

La descripción anterior de las realizaciones descritas se proporciona para permitir a una persona experta en la técnica hacer o usar la presente invención. Varias modificaciones a estas realizaciones serán inmediatamente evidentes para los expertos en la técnica, y los principios genéricos definidos en este documento se pueden aplicar a otras realizaciones sin apartarse del alcance de la invención, según lo definido por las reivindicaciones.

REIVINDICACIONES

1. Un procedimiento para realizar el reconocimiento de voz que comprende:

realizar el apareo de patrones de un primer segmento de voz de entrada con al menos una primera plantilla acústica de un modelo acústico (230, 232) independiente del orador, para producir al menos una plantilla de apareo de patrones de entrada y para determinar una clase de emisión vocal reconocida, en el cual la clase de emisión vocal es una palabra o segmento de habla específico;

comparar dicha(s) plantilla(s) de apareo de patrones de entrada con una plantilla correspondiente asociada a al menos una segunda plantilla acústica proveniente del modelo acústico (234) del orador de la primera voz de entrada, la segunda plantilla acústica asociada a la clase de emisión vocal reconocida; y

determinar si se actualiza o no dicha(s) segunda(s) plantilla(s) acústica(s), en donde dicha(s) segunda(s) plantilla(s) acústica(s) se actualiza(n) si dicha(s) plantilla(s) de apareo de patrones de entrada es (son) mejor(es) que la correspondiente plantilla asociada a dicha(s) segunda(s) plantilla(s) acústica(s).

2. El procedimiento de la reivindicación 1, en el cual la realización del apareo de patrones comprende adicionalmente:

realizar el apareo de patrones (230) del modelo oculto de Markov (HMM) del primer segmento de voz de entrada con al menos una plantilla HMM, para generar al menos una plantilla HMM de apareo;

realizar el apareo de patrones (232) de la distorsión dinámica en el tiempo (DTW) con al menos una plantilla de DTW, para generar al menos una plantilla de apareo DTW; y

realizar al menos una suma ponderada de al menos una plantilla de apareo HMM y al menos una plantilla de apareo DTW para generar al menos una plantilla de apareo del patrón de entrada.

3. El procedimiento de la reivindicación 1, que comprende adicionalmente:

generar al menos una plantilla de apareo independiente del orador, realizando el apareo de patrones de un segundo segmento de voz de entrada con al menos una primera plantilla acústica, en donde al menos dicha primera plantilla acústica es independiente del orador;

generar al menos una plantilla de apareo dependiente del orador, realizando el apareo de patrones de un segundo segmento de voz de entrada con al menos dicha segunda plantilla acústica, en donde dicha al menos una plantilla acústica es dependiente del orador; y

combinar dicha al menos una plantilla de apareo independiente del orador con dicha al menos una plantilla de apareo dependiente del orador para generar al menos una plantilla de apareo combinada.

4. El procedimiento de la reivindicación 3, que comprende adicionalmente identificar la clase de emisión vocal asociada a la mejor entre dicha(s) plantilla(s) de apareo combinada(s).

5. El procedimiento de la reivindicación 1, en el cual el procedimiento es adicionalmente para realizar el entrenamiento y prueba del reconocimiento de voz sin supervisión, que comprende:

realizar en un motor (220) de reconocimiento de voz el apareo de patrones de la voz de entrada de un orador con el contenido de un modelo acústico (230, 232) independiente del orador, para producir plantillas de apareo de patrones independientes del orador;

comparar las plantillas de apareo de patrones independientes del orador con plantillas asociadas a plantillas de un modelo acústico (234) dependiente del orador por el motor (220) de reconocimiento de voz, en donde el modelo acústico dependiente del orador está personalizado para el orador; y

si las plantillas de apareo de patrones independientes del orador son mejores que las plantillas asociadas a plantillas del modelo acústico (234) dependiente del orador, generar una nueva plantilla para el modelo acústico (234) dependiente del orador, en base a las plantillas de apareo del patrón independiente del orador.

6. El procedimiento de la reivindicación 5, en el cual el modelo acústico (230, 232) independiente del orador comprende al menos un modelo acústico del modelo oculto de Markov (HMM).

7. El procedimiento de la reivindicación 5, en el cual el modelo acústico (230, 232) independiente del orador comprende al menos un modelo acústico de distorsión dinámica en el tiempo (DTW).

8. El procedimiento de la reivindicación 5, en el cual el modelo acústico (230, 232) independiente del orador comprende al menos un modelo acústico del modelo oculto de Markov (HMM) y al menos un modelo acústico de distorsión dinámica en el tiempo (DTW).

9. El procedimiento de la reivindicación 5, en el cual el modelo acústico (230, 232) independiente del orador incluye al menos una plantilla de información inservible, en donde la comparación incluye comparar la voz de entrada con dicha al menos una plantilla de información inservible.

10. El procedimiento de la reivindicación 5, en el cual el modelo acústico (234) dependiente del orador comprende al menos un modelo acústico de distorsión dinámica en el tiempo (DTW).

11. El procedimiento de la reivindicación 5, que comprende adicionalmente:

configurar el motor (220) de reconocimiento de voz para comparar un segundo segmento de voz de entrada con el contenido del modelo acústico independiente del orador y el modelo acústico dependiente del orador, para generar al menos una plantilla de apareo combinada, dependiente del orador e independiente del orador, e

identificar una clase de emisión vocal con la mejor plantilla combinada de apareo dependiente del orador e independiente del orador.

12. El procedimiento de la reivindicación 11, en el cual el modelo acústico (230, 232) independiente del orador comprende al menos un modelo acústico (230) del modelo oculto de Markov (HMM).

13. El procedimiento de la reivindicación 11, en el cual el modelo acústico (230, 232) independiente del orador comprende al menos un modelo acústico (232) de distorsión dinámica en el tiempo (DTW).

14. El procedimiento de la reivindicación 11, en el cual el modelo acústico (230, 232) independiente del orador comprende al menos un modelo acústico (230) del modelo oculto de Markov (HMM) y al menos un modelo acústico (232) de distorsión dinámica del tiempo (DTW).

15. El procedimiento de la reivindicación 11, en el cual el modelo acústico (234) dependiente del orador comprende al menos un modelo acústico de distorsión dinámica del tiempo (DTW).

16. El procedimiento de la reivindicación 1, que comprende adicionalmente:

realizar el apareo de patrones del segmento de voz de entrada con una plantilla acústica dependiente del orador, para generar al menos una plantilla de apareo dependiente del orador; y

combinar dicha al menos una plantilla de apareo independiente del orador con dicha al menos una plantilla de apareo dependiente del orador para generar al menos una plantilla de apareo combinada, en donde cada plantilla de apareo combinada corresponde a una clase de emisión vocal y depende de la plantilla de apareo del patrón independiente del orador para la clase de emisión vocal, y de la plantilla de apareo del patrón dependiente del orador para la clase de emisión vocal.

17. El procedimiento de la reivindicación 1, en el cual la etapa de realizar y la etapa de combinar son llevadas a cabo por un motor (220) de reconocimiento de voz.

18. Un producto de programa de ordenador que comprende instrucciones que, cuando son ejecutadas por un procesador, causan que el procesador realice un procedimiento de cualquiera de las reivindicaciones 1 a 17.


SEÑALES DE VOZ
102
FILTRO DE PREÉNFASIS
104
EXTRACCIÓN DE CARACTERÍSTICA ACÚSTICA (AFE)
110
CASACIÓN DE PATRONES ACÚSTICOS Y MODELADO DEL LENGUAJE
112
MODELO ACÚSTICO
CADENA ESTIMADA DE PALABRAS

FIG. 1

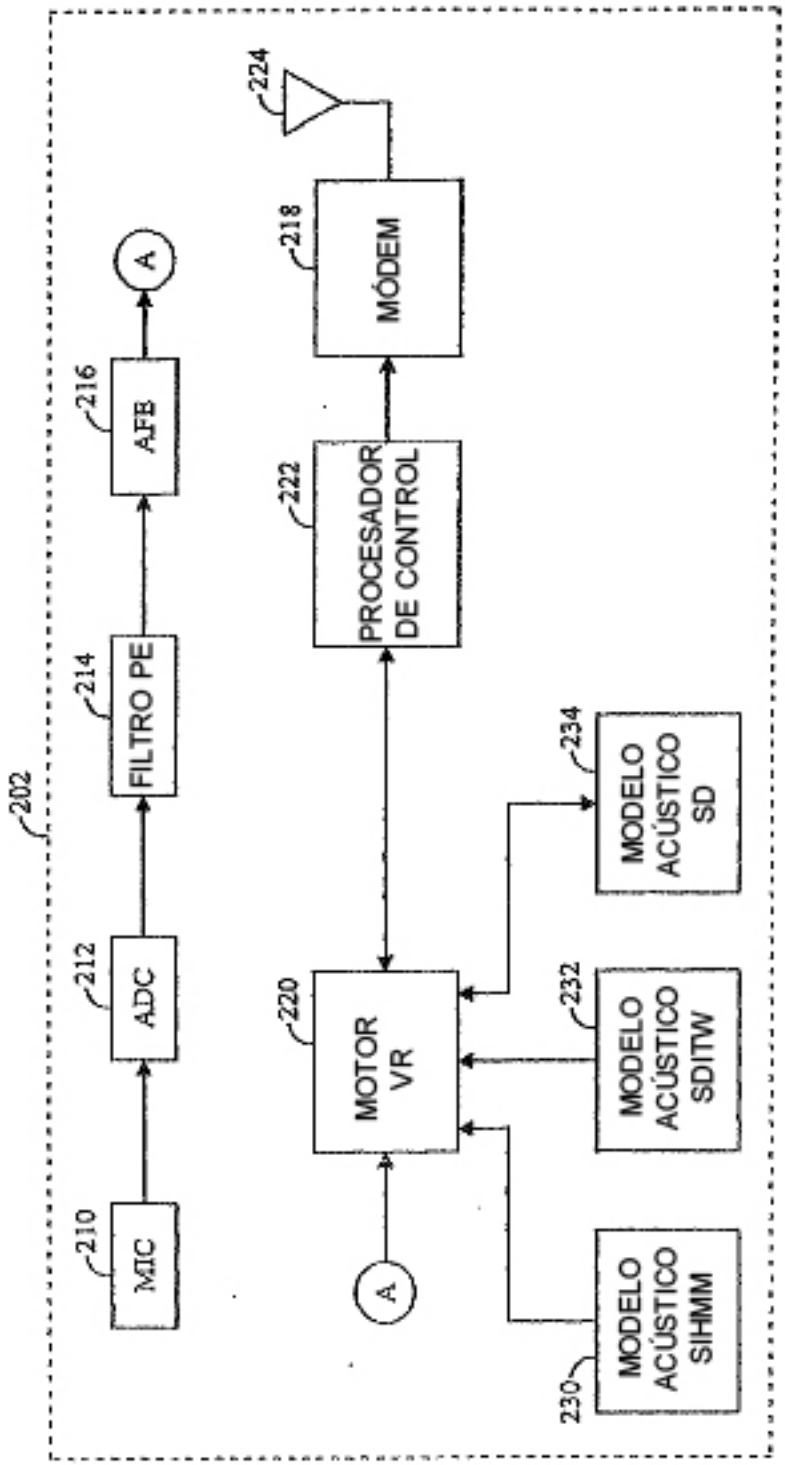
202
MIC
210
ADC
212
FILTRO PE
214
AFE
216
A
MOTOR VR
220
PROCESADOR DE CONTROL
222
MÓDEM
218
224
MODELO ACÚSTICO SIHMM
230
MODELO ACÚSTICO SDITW
232
MODELO ACÚSTICO SD
234
A

FIG. 2

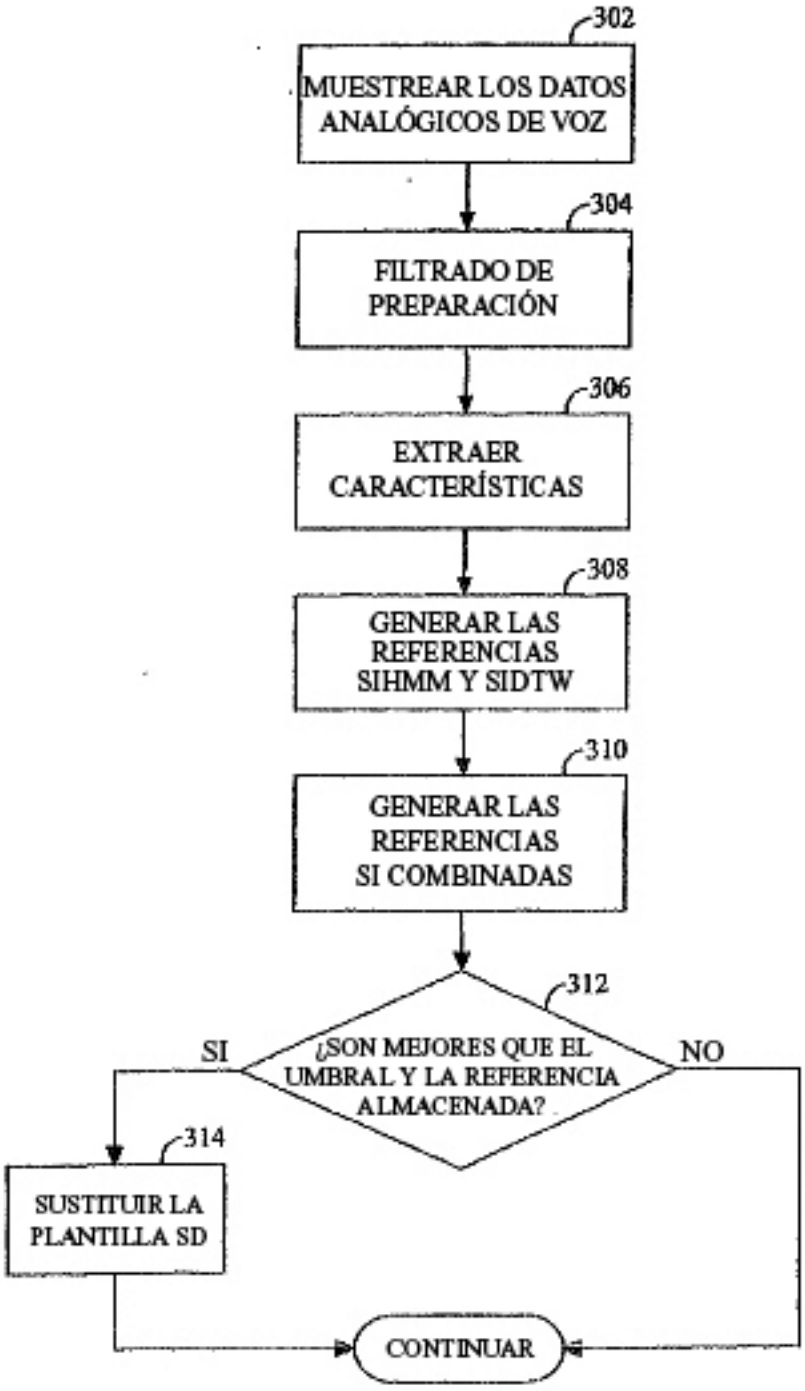
302
MUESTREAR LOS DATOS ANALÓGICOS DE VOZ
304
FILTRADO DE PREPARACIÓN
306
EXTRAER CARACTERÍSTICAS
308
GENERAR LAS REFERENCIAS SIHMM Y SIDTW
310
GENERAR LAS REFERENCIAS SI COMBINADAS
312
¿SON MEJORES QUE EL UMBRAL Y LA REFERENCIA ALMACENADA?
SI
NO
314
SUSTITUIR LA PLANTILLA SD
CONTINUAR

FIG. 3

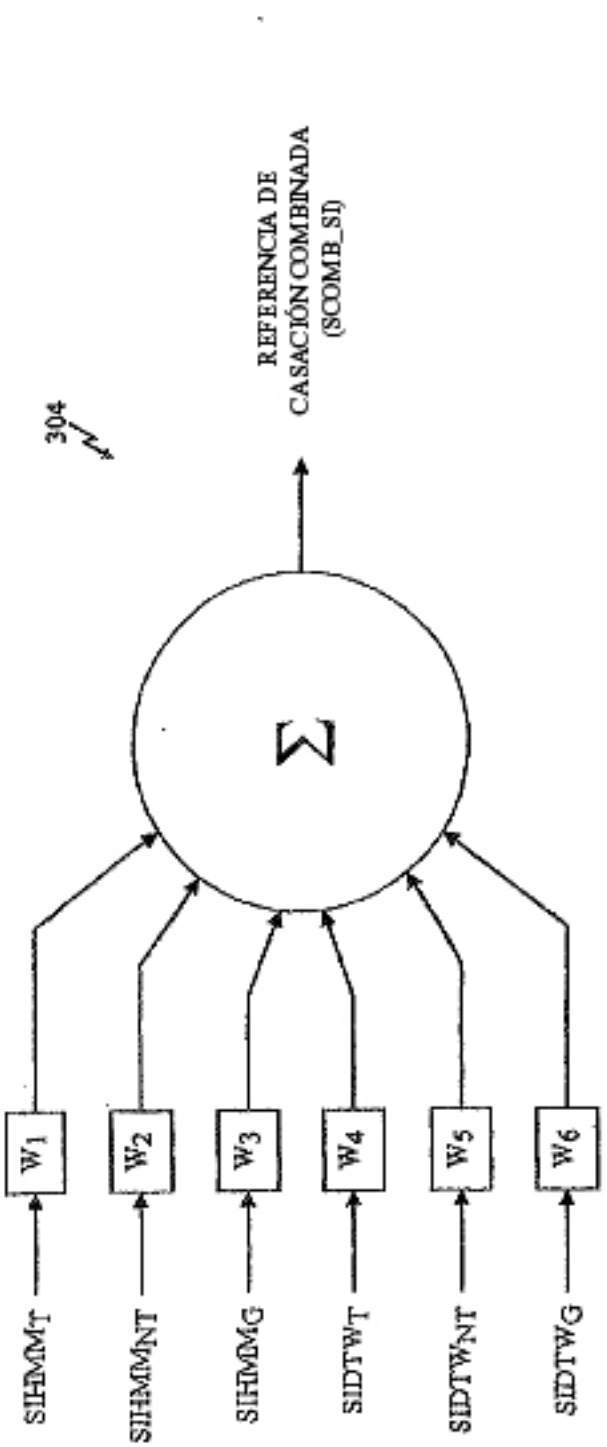

FIG. 4

ECUACIÓN 1: $S_{COMB\_SI} = (SIHMM_T * W_1) + (SIHMM_{NT} * W_2) + (SIHMM_G * W_3) + (SIDTW_T * W_4) + (SIDTW_{NT} * W_5) + (SIDTW_G * W_6)$

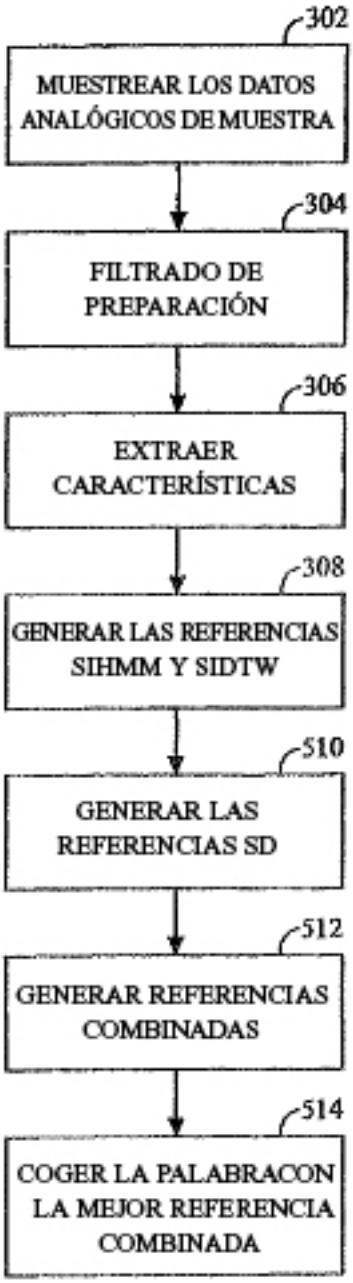
302
MUESTREAR LOS DATOS ANALÓGICOS DE MUESTRA
304
FILTRADO DE PREPARACIÓN
306
EXTRAER CARACTERÍSTICAS
308
GENERAR LAS REFERENCIAS SIHMM Y SIDTW
510
GENERAR LAS REFERENCIAS SD
512
GENERAR REFERENCIAS COMBINADAS
514
COGER LA PALABRACON LA MEJOR REFERENCIA COMBINADA

FIG. 5

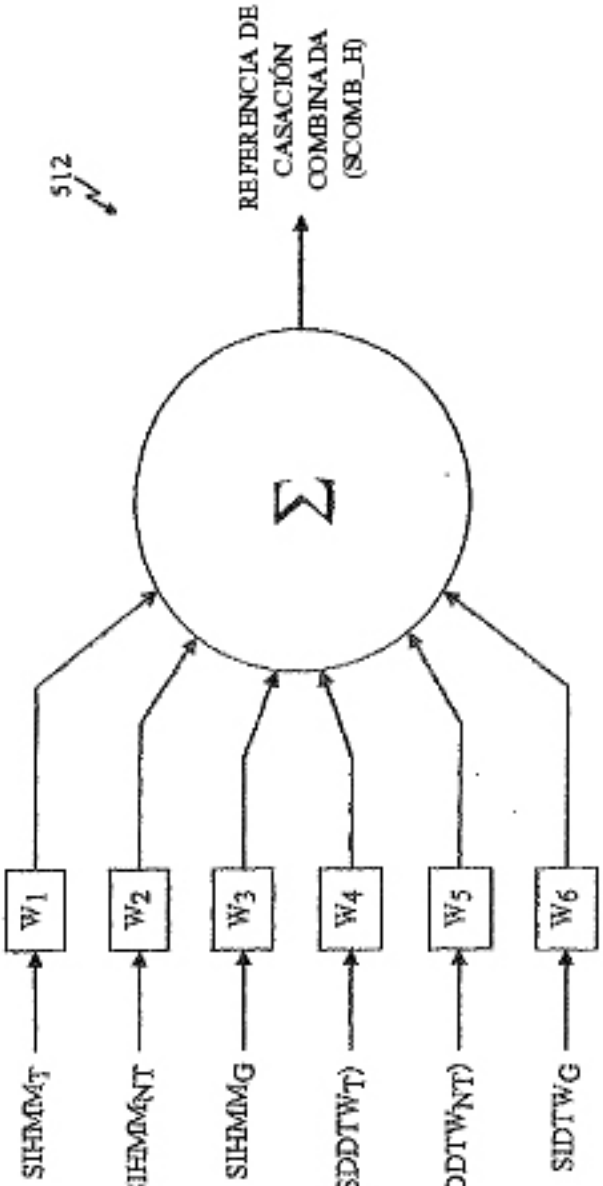

FIG. 6

ECUACIÓN 2 $SCOMB\_H = (SIHMM_T * W_1) + (SIHMM_{NT} * W_2) + (SIHMM_G * W_3) + (DTW_T * W_4) + (DTW_{NT} * W_5) + (SIDTW_G * W_6)$