(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 311 872**

(51) Int. Cl.:
G10L 15/08 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **04805044 .7**
(96) Fecha de presentación : **28.12.2004**
(97) Número de publicación de la solicitud: **1831870**
(97) Fecha de publicación de la solicitud: **12.09.2007**

(54) Título: **Sistema y procedimiento de reconocimiento vocal automático.**

(73) Titular/es: **Loquendo S.p.A.**
Via Arrigo Olivetti 6
Torino, IT

(45) Fecha de publicación de la mención BOPI:
16.02.2009

(72) Inventor/es: **Colibro, Daniele y**
Vair, Claudio

(45) Fecha de la publicación del folleto de la patente:
16.02.2009

(74) Agente: **Ponti Sales, Adelaida**

DESCRIPCIÓN

Sistema y procedimiento de reconocimiento vocal automático.

Sector técnico de la invención

La presente invención se refiere en general al reconocimiento de voz automático, y en particular a medidas de confianza en el reconocimiento de voz automático.

Antecedentes

Como es sabido, los sistemas de reconocimiento de voz automáticos (ASRs, siglas de *automatic speech recognition systems*) están diseñados para convertir una representación digital de una señal de voz, transformando el discurso en una secuencia de texto de palabras, que hace una suposición sobre el contenido léxico de la propia señal de voz. El proceso de reconocimiento automático utiliza modelos acústicos estocásticos, y por lo tanto, el resultado obtenido, en términos de secuencia de palabras reconocidas, puede verse afectado por una tasa de errores residuales diferente de cero. Además, el ámbito de las formulaciones reconocidas por un sistema de reconocimiento de voz automático está en cualquier caso ligado a un vocabulario limitado, formalizado mediante un modelo estadístico de lenguaje o de gramáticas libres de contexto, que pueden ser reducidas a autómatas finitos (este es el caso, por ejemplo, de una gramática que describe la manera de pronunciar una fecha o una hora). Por lo tanto, la pronunciación de palabras que están fuera del vocabulario o de las formulaciones no previstas da lugar a errores de reconocimiento. Es por lo tanto deseable un sistema de reconocimiento de voz automático que disponga de una medida de la fiabilidad de las palabras reconocidas.

Con este propósito, los sistemas de reconocimiento de voz automáticos proporcionan lo que en la literatura es conocido como medida de confianza, que es un indicador de fiabilidad comprendido entre 0 y 1, y que puede aplicarse a las palabras individuales reconocidas y/o a su secuencia. En el caso de un error de reconocimiento, la medida de confianza debería tomar valores bajos, y en cualquier caso, valores inferiores a aquellos que se obtienen en ausencia de errores. Se puede fijar un umbral en los valores de confianza medidos para evitar propuestas de resultados que no son tan fiables.

Los sistemas de reconocimiento de voz automáticos más avanzados permiten el reconocimiento con vocabularios flexibles, que son definidos por el usuario y descritos mediante formalismos adecuados. Para lograr este resultado, los modelos de voz utilizados para el reconocimiento están compuestos de las unidades acústico-fonéticas (APUs, siglas de "acoustic-phonetic units") elementales o subpalabras, cuya composición secuencial permite la representación de cualquier palabra de un determinado lenguaje. Las herramientas matemáticas utilizadas para describir la evolución temporal de la voz son los llamados modelos ocultos de Markov (HMMs, siglas de "*hidden Markov Models*"), y cada unidad acústico-fonética elemental está representada por un modelo oculto de Markov, que está formado por estados que describen la evolución temporal de estas. Las palabras a reconocer, que se describen como secuencias de las unidades acústico-fonéticas elementales, se obtienen concatenando modelos ocultos de Markov constituyentes individuales.

Además de describir la evolución temporal de la voz, los modelos ocultos de Markov permiten la generación de las probabilidades de emisión, también conocidas como probabilidades de salida, de los estados acústicos que los forman, dados los vectores de observación que transforman la información de la señal de voz. La secuencia de las probabilidades, junto con su evolución temporal, permite obtener el resultado de reconocimiento.

La probabilidad de emisión de los estados acústicos de los modelos ocultos de Markov se puede obtener utilizando un modelo estadístico característico, en el que las distribuciones de los vectores de observación, que resumen el contenido de información de la señal de voz en cantidades discretas de tiempo, más adelante designadas como ventanas, son, por ejemplo, representadas por mezclas de distribuciones Gaussianas multivariantes. Mediante algoritmos de capacitación, conocidos en la literatura como algoritmos de Segmento K y de avance-retroceso, es posible estimar el valor medio de la varianza de las distribuciones Gaussianas de las mezclas, a partir de bases de datos de voz ya registradas o y glosadas. Para una descripción más detallada de la teoría, los algoritmos y la implementación de modelos ocultos de Markov, se puede recurrir a Huang X., Acero A., y Hon H.W., *Spoken Lenguaje Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall*, Capítulo 8, páginas 377-413, 2001.

Otro procedimiento para obtener la probabilidad de emisión de modelos ocultos de Markov es usar modelos discriminadores, que tienden a resaltar las peculiaridades de los modelos individuales comparados con otros. Una técnica conocida en la literatura, y que es ampliamente utilizada, es la de las redes neuronales artificiales (ANNs, siglas de "*Artificial neural networks*"), que representan sistemas no lineales capaces de producir, tras seguir una capacitación, las probabilidades de emisión de los estados acústicos, dado el vector de observación acústico. En general, los sistemas de reconocimiento de este tipo son designados como HMM-NN híbridos.

El detalle de las unidades acústico-fonéticas elementales utilizadas como componentes para la composición de los modelos de palabras pueden depender del tipo de modelización de las probabilidades de emisión de estados de Markov. En general, cuando se recurre a modelos caracterizadores (mezclas de Distribuciones Gaussianas multivariantes), se adoptan unidades acústico-fonéticas elementales contextuales. Como ejemplo, están los trifonos, que representan los

fonemas básicos de un determinado lenguaje que están especializados en las palabras y sus contextos izquierdo y derecho (fonemas adyacentes). En el caso de las unidades acústico-fonéticas elementales capacitadas empleando la discriminación representativa (por ejemplo, con la técnica ANN), puede ser menos necesaria la contextualización; en este caso, se pueden emplear fonemas independientes del contexto o composición de unidades estacionarias (es decir, la parte estacionaria de los fonemas independientes de contexto) y unidades de transición (es decir, difonos de transición entre fonemas), tal como se describe en "*Acoustic-Phonetic Modelling for Flexible Vocabulary Speech Recognition*", de L. Fissore, F. Ravera, P. Laface, Proc. de EUROSPEECH, páginas. I 799-802, Madrid, España, 1995.

Un procedimiento ampliamente utilizado en el estado de la técnica para calcular la confianza consiste en utilizar las así llamadas probabilidades *a posteriori*, que son cantidades derivadas de las probabilidades de emisión de los modelos ocultos de Markov. El logaritmo de lasprobabilidades *a posteriori*, calculadas para cada trama, pueden promediarse ponderando todas las tramas con el mismo peso o también ponderando todos los fonemas con el mismo peso, tal como se describe en Rivlin, Z. *Et al*., "*A Phone-Dependent Confidence Measure for Utterance Rejection*", *Proc. of the IEEE International Conference on Acoustics, Speech y Signal Processing*, Atlanta, GA, páginas 515-517 (Mayo de 1996). Se pueden utilizar criterios similares en sistemas híbridos HMM-NN, que son capaces de producir directamente probabilidades *a posteriori* de los estados/fonemas. En Bernardis G. *Et al*., "*Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems*", *Proceedings of the International Conference on Spoken Language Processing*, páginas 775-778, Sydney, Australia (Dic. 1998), se proporciona una comparación entre diferentes maneras de promediar probabilidades *a posteriori* paraobtener medidas de confianza a partir de fonemas o de palabras.

Otra técnica ampliamente utilizada plantea normalizar lasprobabilidades *a posteriori*, o directamente las probabilidades de emisión, que coinciden en el cálculo de la confianza mediante un factor que no tiene en cuenta las restricciones de reconocimiento léxicas y gramaticales. La comparación entre las dos cantidades, es decir, el resultado obtenido aplicando las restricciones y el resultado obtenido disminuyendo las restricciones, proporciona información útil para determinar la confianza. De hecho, si las dos cantidades tienen valores comparables, significa que la introducción de las restricciones de reconocimiento no ha producido ninguna distorsión particular con respecto a lo que habría ocurrido sin restricciones de reconocimiento. Por lo tanto, puede considerarse que el resultado de reconocimiento es fiable, y su confianza debería tener valores elevados, cercanos a su límite superior. En cambio, cuando, el resultado con restricciones es considerablemente peor que el resultado sin restricciones, se puede deducir que el reconocimiento no es fiable en la medida en que el sistema de reconocimiento de voz automático habría producido un resultado diferente del obtenido como consecuencia de la aplicación de las restricciones. En este caso, la medida de confianza debería producir valores bajos, cercanos a su límite inferior.

Se han propuesto varias realizaciones de esta técnica en la literatura. En Gillick M. *Et al*., "*A Probabilistic Approach to Confidence Estimation and Evaluation*", *Proc. of the IEEE International Conference on Acoustics, Speech y Signal Processing, Munich, Germany*, páginas 879-882 (Mayo de 1997), se adopta la diferencia entre cantidades conocidas como resultado acústico y mejor resultado, donde los dos términos se obtienen respectivamente promediando el resultado acústico (con restricciones) y el mejor resultado (sin restricciones), obtenido para cada trama con los modelos ocultos de Markov acústicos en el intervalo de tiempo correspondiente a las palabras. De la misma manera, en US 5,710,866 de Alleva *et al*., la confianza se calcula como la diferencia entre un resultado acústico forzado y un resultado acústico no forzado, y esta diferencia se calcula para cada trama, de manera que se puede utilizar para ajustar el resultado acústico forzado empleado durante el reconocimiento. Weintraub, M. *Et al*., "*Neural Network Based Measures of confidence for word recognition*", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, páginas 887-890 (Mayo de 1997), propone una familia de medidas de confianza basada en características acústicas que difieren según los tipos de modelos empleados como factor de normalización y según el nivel en que se combinan los logaritmos de las probabilidades de los modelos ocultos de Markov de las distintas ventanas (nivel de palabra, nivel telefónico, nivel de estado telefónico). Los modelos empleados para la normalización pueden ser de fonemas independientes de contexto o de mezclas Gaussianas (GMMs, siglas de "*Gaussian Mixture Models*") y permiten obtener el resultado en el caso de ausencia de restricciones de reconocimiento.

En Andorno M. *Et al*., "*Experiments in Confidence Scoring for word and Sentence Verification*", *Proc. of the International Conference on Spoken Language Processing*, páginas 1377 a 1380, Denver, CO (Septiembre de 2002) se propone otra formulación más, que se refiere a un sistema HMM-NN híbrido. En este caso, la confianza se obtiene como ratio entre el resultado acústico no forzado y el resultado acústico forzado. El numerador se calcula como el promedio, sobre el número de tramas de la palabra, de los logaritmos de las mejores probabilidades *a posteriori* de entre todos los estados de los modelos acústicos, mientras que el denominador está representado por el promedio de las probabilidades *a posteriori* sobre la secuencia de estados que se produce con la así llamada Alineación de Viterbi. Para una descripción más detallada de la Alineación de Viterbi, se puede hacer referencia al ya mencionado "*Spoken Language Processing: A Guide to Theory, Algorythm, and system Development, chapter 8*".

Asimismo, se conocen en el estado de la técnica técnicas para mejorar la medida de confianza. En US 6,421,640, por ejemplo, se adapta la confianza mediante un ajuste específico para el usuario o la frase pronunciada, antes de ser comparado con el umbral para decidir si se propone el resultado de reconocimiento al usuario.

En US 6,539,353 la mejora en la calidad de la medida de confianza se obtiene aplicando una ponderación específica a las medidas de confianza de las sub-palabras, que son entonces combinadas para obtener la medida de confianza de la palabra.

La variedad de unidades acústico-fonéticas elementales, incluso en el mismo sistema de reconocimiento de voz automático, provoca una falta de homogeneidad en las probabilidades de emisión (probabilidades *a posteriori*), que son afectadas por el detalle de las unidades acústico-fonéticas elementales, por su incidencia en las palabras de un determinado lenguaje, por las características de los sonidos que representan, por la cantidad de material de capacitación disponible para su estimación, y así sucesivamente. Puesto que las medidas de confianza conocidas en la literatura se derivan de las probabilidades de emisión, la variabilidad de estas últimas produce inestabilidad y falta de homogeneidad en la propia confianza.

Objeto y resumen de la invención

Por lo tanto, el objeto de la presente invención es hacer que la medida de confianza sea homogénea en términos de capacidad de rechazo e independiente de los contenidos de las gramáticas y de los vocabularios de reconocimiento, del lenguaje de los modelos acústicos y, en general, de las restricciones de reconocimiento.

Este objeto es alcanzado por la presente invención gracias a que se trata de un procedimiento y un sistema de reconocimiento de voz automático, tal como se define en las reivindicaciones 1 y 16, respectivamente, y a un producto de programa informático, tal como se define en la reivindicación 17.

La presente invención alcanza el objeto mencionado empleando una medida de confianza basada en contribuciones diferenciales calculadas para cada trama de la ventana de análisis como diferencia entre el resultado acústico no forzado y el resultado acústico forzado, y promediado sobre todo el intervalo de reconocimiento. Esto hace posible actuar sobre la contribución diferencial individual de la suma, aplicándole una función de normalización respectiva, que hace que el comportamiento de las distintas unidades acústico-fonéticas elementales sea homogéneo en términos de capacidad de rechazo, a medida que varían el lenguaje, el vocabulario y la gramática. La solución propuesta permite que las distribuciones de las medidas de confianza sean invariantes con respecto al lenguaje, el vocabulario y la gramática, y homogénea entre objetos de reconocimiento y lenguajes diferentes. Esto facilita enormemente el desarrollo de aplicaciones en las etapas iniciales de su desarrollo, puesto que no precisan de calibración específica alguna para cada sesión de reconocimiento individual.

La función de normalización aplicada a los términos diferenciales individuales está constituida por una familia de distribuciones acumulativas, una para cada contribución diferencial a la medida de confianza. Cada función puede estimarse de manera simple a partir de un conjunto de datos de capacitación y es específica para cada estado de la unidad acústico-fonética elemental. Por lo tanto, la solución propuesta no requiere consideraciones heurísticas o suposiciones *a priori* y hace posible obtener todas las cantidades necesarias para obtener directamente la medida de confianza diferencial a partir de los datos de capacitación.

Breve descripción de los dibujos

Para una mejor comprensión de la presente invención, se describirá a continuación una realización preferida, ofrecida únicamente a modo de ejemplo y en ningún caso limitativa, con referencia a los dibujos adjuntos, en los cuales:

- La figura 1 muestra un diagrama de bloque de un sistema de reconocimiento de voz automático según la presente invención;

- Las figuras 2 y 3 muestran curvas de rechazo en sistemas de reconocimiento de voz automáticos del estado de la técnica y de la presente invención, respectivamente; y

- La figura 4 muestra un diagrama de bloque para la estimación de los parámetros de distribuciones acumulativas.

Descripción detallada de realizaciones preferidas de la invención

La siguiente discusión se presenta para permitir a un experto en la materia utilizar la invención. Varias modificaciones de las realizaciones se harán fácilmente aparentes para los expertos en la materia, y los principios generales presentados se podrán aplicar a otras realizaciones y aplicaciones sin salir del alcance de la presente invención tal como se define mediante las reivindicaciones adjuntas. De este modo, la presente invención no se limita a las realizaciones mostradas, si no que se le debe atribuir el alcance más amplio consecuente con los principios y características aquí descritas y tal como se define en las reivindicaciones adjuntas.

La figura 1 muestra un diagrama de bloque de un sistema de reconocimiento de voz automático que implementa la medida de confianza diferencial según la invención.

Tal como se muestra en la figura 1, el sistema de reconocimiento de voz automático, indicado en su totalidad como 1, incluye:

- un convertidor analógico digital 2 que recibe una señal de voz analógica proveniente de un dispositivo de entrada (no mostrado), que podría ser, por ejemplo, un micrófono de un PC conectado a una tarjeta de sonido o a un auricular

de un teléfono, conectado a una red móvil o fija, y que proporciona una señal de voz digitalizada formada por una secuencia de muestras digitales;

• un extractor de vectores 3 que recibe la señal de voz digitalizada proveniente del convertidor digital-analógico 2, y que suministra, para ventanas fijas (por ejemplo, de 10 ms), un vector de observación, que es una representación vectorial compacta del contenido de información del discurso. El extractor de vectores 3 calcula convenientemente, para cada trama, los así llamados coeficientes Mel-scale Cepstrum, que se describen en el ya mencionado "*Spoken Language Processing: A Guide to Theory, Algorythm, and System Development*", páginas 316-318;

• un módulo de cálculo de probabilidades 4 que recibe los vectores de observación proveniente del extractor de vectores 3 y los modelos acústicos de las unidades acústico-fonéticas elementales, almacenados en el módulo de almacenamiento 5 (siendo los modelos acústicos autómatas de Markov, cuyas probabilidades de emisión pueden ser determinadas por mezclas de Distribuciones Gaussianas multivariantes o empleando la técnica de redes neuronales artificiales, en particular un sistema HMM-NN híbrido), y que proporciona las probabilidades de emisión de las unidades acústico-fonéticas elementales;

• un motor de búsqueda 6 que recibe las probabilidades de emisión provenientes del módulo de cálculo de probabilidades 4 y las restricciones de conocimiento esquematizadas como una red de estados finitos (FSN, siglas de *finite state network*), que describen todos los caminos de reconocimiento admisibles, y almacenados en el módulo de almacenamiento 7, y que proporciona una secuencia de palabras hipótesis. En particular, el núcleo del motor de búsqueda 6 implementa una técnica muy conocida en la literatura como programación dinámica, una de cuyas implementaciones es la antes mencionada Alineación de Viterbi, y que realiza un alineamiento temporal dinámico entre los vectores de observación y los estados de las unidades acústico-fonéticas elementales;

• un módulo de medida de confianza 8 que recibe las probabilidades de emisión provenientes del módulo de cálculo de probabilidades 4 y los parámetros de distribuciones acumulativas calculados a partir de datos de capacitación, tal como se describe más adelante con referencia a la figura 4, almacenados en el módulo de almacenamiento 9, y que proporciona una medida de confianza para cada palabra de hipótesis; y

• un módulo de gestión de los resultados 10 que recibe las palabras hipótesis provenientes del motor de búsqueda 6 y las medidas de confianza provenientes del módulo de medida de confianza 8, y que proporciona un resultado de reconocimiento completo y, opcionalmente, una lista de reconocimientos alternativos (N-mejores reconocimientos), que son los menos probables pero potencialmente útiles para la aplicación lógica.

Durante el proceso de reconocimiento, se utilizan las probabilidades de emisión de las unidades acústico-fonéticas elementales para determinar la secuencia más probable de estados acústicos. Junto con las probabilidades de emisión, para cada trama t de la ventana de reconocimiento, se calcula una contribución diferencial $C_{i,t}$ a la medidade confianza mediante el módulo de medida de confianza 8, como diferencia entre un resultado acústico forzado $SCV_{i,t}$ para elestado acústico i al nivel de la trama de voz de entrada t, y un resultado acústico no forzado SCL_t en la misma trama de voz de entrada t, y ponderada con una función de distribución acumulativa $F_{\xi i}$ de la propia diferencia, calculada a partir de los datos de voz de capacitación:

$$C_{i,t} = F_{\xi i}\,(SCV_{i,t} - SCL_t) \qquad \textbf{(Ecuación 1)}$$

donde el resultado acústico forzado puede definirse como la probabilidad de que la voz de entrada se corresponda con la hipótesis de reconocimiento, el resultado acústico no forzado está definido como la probabilidad de que la voz de entrada se corresponda con la hipótesis de reconocimiento sin restricciones FSN, y la función de distribución acumulativa $F_{\xi i}(x)$ es la probabilidad de que la variable ξ_i sea menor o igual a x. La contribución diferencial $C_{i,t}$ y la función de distribución acumulativa $F_{\xi}i(x)$ están relacionadas con el estado acústico i, que es un elemento de una cadena de un modelo oculto de Markov, que a su vez está constituido por uno o más estados, dependiendo del tipo de sub-palabras empleado para modelizar una unidad acústico-fonética determinada.

Hay varios algoritmos bien conocidos para calcular los resultados acústicos forzados y no forzados, tales como el algoritmo de avance-retroceso o el algoritmo de Viterbi, que son tratados con más detalle en el ya mencionado "*Spoken Lenguaje Processing: A Guide to Theory, Algorythm, and System Development, chapter 8*".

Cada función de distribución acumulativa es creciente monótona, con los siguientes límites: $F_{\xi i}(\infty) = 1$ y $F_{\xi i}(0) = 0$. Otra propiedad útil es que la función de distribución acumulativa $F_{\xi i}$ aplicada a la variable aleatoria correspondiente ξ_i produce una nueva variable aleatoria $\eta = F_{\xi i}(x)$ con una distribución uniforme entre 0 y 1. Puesto que los términos elementales individuales $C_{i,t}$, que son ejemplos de la nueva variable aleatoria distribuida uniformemente η, están en el rango [0,1] y puesto que están relacionados con la comparación entre el resultado acústico forzado $SCV_{i,t}$ yel resultado acústico no forzado $SCL_{i,t}$, se pueden considerar a todos los efectos como las componentes elementales de la nueva medida de confianza.

Expresando el resultado acústico forzado $SCV_{i,t}$ y el resultado acústico no forzado $SCL_{i,t}$ como una función de las probabilidades logarítmicas de una emisión de estados acústicos S_i, dado el vector de observación o_t en la ventana t:

$$SCV_{i,t} = \log P(S_i \mid o_t) \qquad \text{(Ecuación 2)}$$

$$SCL_t = \log\left[\max_{1 \le j \le N_s} P(S_j \mid o_t)\right] \qquad \text{(Ecuación 3)}$$

donde N_s es el número total de estados de modelos ocultos de Markov del sistema de reconocimiento.

La forma desarrollada de la contribución diferencial a la medida de confianza se puede calcular de la manera siguiente:

$$SCL_t = \log\left[\max_{1 \le j \le N_s} P(S_j \mid o_t)\right] \qquad \text{(Ecuación 4)}$$

En la siguiente descripción, las probabilidades *a posteriori* generadas por una red neuronal artificial de un sistema HMM-NN híbrido o derivadas de las probabilidades a partir de un modelo oculto de Markov estándar se obtienen como probabilidades de emisión.

El numerador de la ecuación 4 es la probabilidad de emisión del estado S_i en la ventana *t* disponible para el algoritmo de programación dinámico. El denominador de la ecuación 4 es, en cambio, la mejor probabilidad de emisión disponible, ligada al resultado acústico no forzado. Con el fin de calcular este último, puede ser necesario un modelo específico, tal como, por ejemplo, una red de fonemas completamente conectados, para ejecutarlo en paralelo con el reconocimiento principal con una etapa de Viterbi sin restricciones léxicas. Otra posibilidad es deducir la mejor probabilidad a partir de un subconjunto de los modelos utilizados para el reconocimiento principal, escogiendo, para cada ventana de tiempo, el modelo más prometedor una vez que se han suavizado las restricciones de reconocimiento.

Finalmente, se calcula la medida de confianza C_{DC}, la cual, en rasgos generales, puede definirse como una medida de lo cercano que está el resultado acústico forzado del mejor resultado coincidente posible para la trama de voz de entrada, como la media temporal de las contribuciones diferenciales definidas por la ecuación 4, sobre la secuencia de los estados, designados por S_{i*}, la cual, de entre todas las secuencias posibles, representa el resultado de reconocimiento:

$$C_{DC} = \frac{1}{T}\sum_{t=1}^{T} F_{t,i*}\left\{\log \frac{P(S_{i*} \mid o_t)}{\max_{1 \le j \le N_s} P(S_j \mid o_t)}\right\} \qquad \text{(Ecuación 5)}$$

donde *T* es el número de tramas de la voz de entrada.

A partir de la ventana de tiempo en la cual se calcula el promedio y la secuencia de estados acústicos de las unidades acústico-fonéticas elementales correspondiente, la medida de confianza adquiere importancia para palabras individuales reconocidas, para una secuencia de estas o para todo el resultado de reconocimiento.

Puesto que la medida de confianza se calcula como promedio de las variables aleatorias distribuidas uniformemente entre 0 y 1, también toma valores en el rango [0, 1] y es adecuada, por lo tanto, para su uso directo como medida de confianza. Asimismo, hay que destacar que, considerando C_{DC} como una variable aleatoria, la distribución de la nueva medida de confianza no es de tipo uniforme, sino más bien de tipo Gaussiano, con un promedio de 0,5, tal como se demuestra con el teorema central del límite.

Las figuras 2 y 3 muestran curvas de rechazo para series de gramáticas de reconocimiento en Inglés Británico, cuyas curvas de rechazo representan la distribución acumulativa del porcentaje R% de reconocimientos rechazados, con un umbral de rechazo R_{TH} igual a la confianza en las abscisas. En particular, en las figuras 2 y 3, A designa la curva relacionada con una gramática Booleana, B designa la curva relacionada con una gramática actual, C designa la curva relacionada con una gramática de fechas, D designa la curva relacionada con una gramática de dígitos, y E designa la curva relacionada con una gramática de números.

Además, La figura 2 muestra curvas de rechazo obtenidas calculando la confianza según una técnica del estado de la técnica, mientras que La figura 3 muestra curvas de rechazo obtenidas calculando la confianza según la presente invención.

Las figuras 2 y 3 representan cada una dos conjuntos de curvas: el conjunto de la izquierda, con una tasa de rechazo que enseguida aumenta a confianzas bajas, corresponde a reconocimientos de formulaciones que quedan fuera del ámbito, mientras que el conjunto de la derecha representa, en cambio, las tasas de rechazo que se obtienen con formulaciones cubiertas por las gramáticas y que son perfectamente reconocidas. Las curvas permiten la calibración de los umbrales de rechazo de la aplicación, estimando de antemano el porcentaje de reconocimientos erróneos propuestos erróneamente como correctos (aceptaciones falsas) y el porcentaje de reconocimientos correctos no considerados como tales (rechazos erróneos). Como se puede apreciar inmediatamente, la confianza diferencial calculada según la presente invención hace que las curvas de rechazo sean claramente más estables y homogéneas a medida que varían las gramáticas.

Tal como ya se ha descrito con referencia a la figura 1, los parámetros de las distribuciones acumulativas proporcionadas por el módulo de medida de confianza se calculan a partir de datos de voz de capacitación. La figura 4 muestra los elementos implicados en el proceso de cálculo.

Para la determinación de la forma y los parámetros de las distribuciones acumulativas son necesarias bases de datos de capacitación disponibles, representadas en la figura 4 por una línea de puntos y designadas por 11, que contienen grabaciones de discurso digitalizados (bloque 12) y las anotaciones de texto correspondientes (bloque 13). Las grabaciones se proporcionan al extractor de vectores 3, que a su vez suministra estos al módulo de cálculo de probabilidades 4. Las anotaciones, en cambio, son suministradas a un módulo autómata APU 14, que proporciona una secuencia de cadena única de las unidades acústico-fonéticas elementales que corresponde a la grabación bajo reconocimiento. La probabilidad de emisión calculada por el módulo de cálculo de probabilidades 4 a partir de los vectores de observación y los modelos acústicos, es suministrada, junto con la secuencia de cadena única de unidades acústico-fonéticas elementales, al motor de búsqueda 6, que realiza una alineación temporal dinámica entre los vectores de observación y los estados de la secuencia de cadena única de unidades acústico-fonéticas elementales. A partir de las probabilidades de emisión provenientes del módulo de cálculo de probabilidades 4 y la alineación entre vectores de observación y estados provenientes del motor de búsqueda 6, un módulo de cálculo de parámetros 15 calcula los parámetros de las distribuciones acumulativas, es decir, el argumento de la ecuación 4:

$$x_{i,t} = \log \frac{P(S_i \mid o_t)}{\max_{1 \le j \le N_s} P(S_j \mid o_t)} \qquad \text{(Ecuación 6)}$$

Los valores de la ecuación 6, uno para cada trama *t*, se consideran como realizaciones de una familia de variables aleatorias, cada una de las cuales es específica de un estado de una unidad acústico-fonética elemental.

Concretamente, el módulo de cálculo de parámetros hace la hipótesis de que las distribuciones de probabilidad pertenecen a una determinada familia (Distribuciones Gaussianas, distribuciones de Laplace, distribuciones gamma, etc.) y determina sus parámetros característicos. Si se consideran distribuciones Gaussianas para las variables aleatorias $x_{i,t}$, la función de densidad de probabilidad $f_{\xi i}(x)$ correspondiente es:

$$f_{\xi_i} = \frac{1}{\sqrt{2\pi\sigma_i}} e^{\frac{x_{i,t} - \mu_i}{\sigma_i}} \qquad \text{(Ecuación 7)}$$

los parámetros característicos de las cuales, es decir, valor medio y varianza, son:

$$\mu_i = \frac{1}{T} \sum_{t=1}^{T} x_{i,t} \qquad \text{(Ecuación 8)}$$

$$\sigma_i^2 = \frac{1}{T} \sum_{t=1}^{T} x_{i,t}^2 - \mu_i^2 \qquad \text{(Ecuación 9)}$$

Una realización particular de la presente invención se aplica a las unidades acústico-fonéticas elementales estacionarias y de transición, y a las probabilidades de emisión calculadas empleando un sistema HMM-NN híbrido optimizando las técnicas de aceleración descritas en WO 03/073416. La red neuronal recibe los vectores de observación del extractor de vectores, y procesa los vectores de observación para proporcionar indicadores de la actividad de cada estado acústico, que son los componentes de los modelos ocultos de Markov que modelizan a las unidades acústico-fonéticas elementales.

Los valores de salida proporcionados por la red neuronal n_i son filtrados mediante una función adecuada (función sigmoide), que permite el cálculo de las probabilidades de emisión (correspondientes a las probabilidades *a posteriori*), es decir, $P(S_i \mid o_t)$, que suman 1 para todo el conjunto de estados admisibles. Su definición se da en la siguiente ecuación:

$$P(S_i \mid o_t) = \frac{e^{n_i(o_t)}}{\sum_{j=1}^{N} e^{n_j(o_t)}} \qquad \text{(Ecuación 10)}$$

Sustituyendo la ecuación 10 en la ecuación 4, se obtiene:

$$C_{i,t} = F_{\xi_i}\left\{\log \frac{e^{n_i(o_t)}}{e^{\max_j n_j(o_t)}}\right\} = F_{\xi_i}\left[n_i(o_t) - \max_j n_j(o_t)\right] \qquad \text{(Ec. 11)}$$

donde n_i es el resultado acústico forzado, y $max\ n_j$ es el resultado acústico no forzado.

En esta realización, el resultado acústico no forzado se calcula directamente a partir de las probabilidades de emisión de los modelos de reconocimiento HMM-NN híbridos principales. La elección del estado acústico con la mejor probabilidad de emisión se limita a las unidades acústico-fonéticas elementales de tipo estacionario, cuyas salidas se calculan siempre, de conformidad con las técnicas de aceleración descritas en WO 03/073416.

En esta realización, la medida de confianza está definida por la siguiente ecuación:

$$C_{DC} = \frac{1}{T}\sum_{t=1}^{T} F_{\xi_{i^*}}\left[n_{i^*}(o_t) - \max_{j \in stat} n_j(o_t)\right] \qquad \text{(Ec. 12)}$$

donde *stat* representa el conjunto de unidades acústico-fonéticas estacionarias.

Finalmente, es obvio que son posibles numerosas modificaciones y variantes de la presente invención, todas al alcance de la invención, tal como se define en las reivindicaciones adjuntas.

En particular, la presente invención se podrá aplicar a sistemas de reconocimiento de voz automáticos que emplean cadenas de Markov y técnicas de alineación temporal dinámicas, y que están basadas en algoritmos de programación dinámica. Estos sistemas de reconocimiento de voz automáticos, de hecho, incluyen modelos ocultos de Markov discretos y continuos, con o sin reparto de distribución (con modelos ligados). Las probabilidades de emisión pueden ser representadas por distribuciones de probabilidad de varios tipos, que incluyen mezclas de Distribuciones Gaussianas multivariantes, distribuciones de Laplace, etc. Las probabilidades de emisión también se pueden obtener utilizando las técnicas de redes neuronales empleadas por sistemas HMM-NN híbridos.

La tipología de las unidades acústico-fonéticas elementales que se pueden utilizar con la medida de confianza según la invención incluye tanto las unidades libres de contexto, tales como los fonemas independientes de contexto, como las unidades de contexto, tales como los difonos, los trifonos o unidades más especializadas. La presente invención también puede aplicarse a las acústicas fonéticas de transición o estacionarias descritas en la publicación previamente mencionada "*Acoustic-Phonetic Modelling for Flexible Vocabulary Speech Recognition*".

Las medidas de confianza se podrán aplicar utilizando las distintas técnicas conocidas para obtener el resultado acústico no forzado. Estas incluyen la ejecución de un modelo específico, tal como, por ejemplo, una red de fonemas completamente conectados, que debe ser ejecutado en paralelo con el reconocimiento principal de una etapa de Viterbi

sin restricciones léxicas. Otra posibilidad es deducir el resultado acústico no forzado a partir de un subconjunto de los modelos utilizados para el reconocimiento principal, escogiendo el mejor modelo con restricciones suavizadas.

La familia de funciones de distribución acumulativas usadas para uniformizar el comportamiento de las unidades acústico-fonéticas elementales se define sobre la base de la familia de funciones de densidad de probabilidad utilizada correspondiente, que puede tener una o más formas predefinidas, tales como Gaussiana, binomial, de Poisson, Gamma, etc., o ser totalmente estimada a partir de datos de voz de capacitación, utilizando procedimientos de adaptación. En particular, el ajuste de los datos es el análisis matemático de un conjunto de datos con el fin de analizar tendencias en los valores de los datos. Esto suele implicar un análisis de regresión de los valores de los datos para definir el conjunto de valores de parámetros que mejor caracteriza la relación entre los puntos de datos y un modelo teórico subyacente. En el caso anterior (capacitación), la etapa de estimación se limita a la determinación de parámetros característicos de las distribuciones (por ejemplo, el valor medio y la varianza en el caso de Distribuciones Gaussianas). En este ultimo caso, la etapa de capacitación tiene como finalidad agrupar los datos con los que se llevará a cabo el ajuste de las funciones paramétricas con un número de grados de libertad adecuado.

Referencias citadas en la descripción

Esta lista de referencias citadas por el solicitante está prevista únicamente para ayudar al lector y no forma parte del documento de patente europea. Aunque se ha puesto el máximo cuidado en su realización, no se pueden excluir errores u omisiones y la OEP declina cualquier responsabilidad en este respecto.

Documentos de patente citados en la descripción

- US 5710866 A, Alleva [0011]
- US 6421640 B [0013]
- US 6539353 B [0014]
- WO 03073416 A [0041] [0044]

Documentos que no son patentes citados en la descripción

- Spoken Language Processing: A Guide to Theory. **HUANG** X.; **ACERO** A.; **HON** H.W. Algorithm, and System Development. Prentice Hall, 2001, 377-413 [0006]
- L. **FISSORE**; F. **RAVERA**; P. **LAFACE**. Acoustic-Phonetic Modelling for Flexible Vocabulary Speech Recognition. *Proc. of EUROSPEECH*, 1995, I 799-802 [0008]
- **RIVLIN**, Z. *et al*. A Phone-Dependent Confidence Measure for Utterance Rejection. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1996, 515-517 [0009]
- **BERNARDIS** G. *et al*. Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition System. *Proceedings of the International Conference on Spoken Language Processing*, December 1998, 775-778 [0009]
- **GILLICK** M. *et al*. A Probabilistic Approach to Confidence Estimation and Evaluation. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1997, 879-882 [0011]
- **WEINTRAUB**, M. *et al*. Neural Network Based Measures of confidence for Word Recognition. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1997, 887-890 [0011]
- **ANDORNO** M. *et al*. Experiments in Confidence Scoring for Word and Sentence Verification. *Proc. of the International Conference on Spoken Language Processing*, September 2002, 1377-1380 [0012]
- Spoken Language Processing: A Guide to Theory. *Algorithm, and System Development*, 316-318 [0023]

REIVINDICACIONES

1. Procedimiento de reconocimiento de voz automático para identificar palabras de una señal de voz de entrada, que incluye:

• el suministro de al menos una hipótesis de reconocimiento sobre la base de dicha señal de voz de entrada, siendo dicha hipótesis de reconocimiento una hipótesis de palabra individual o hipótesis de secuencia de palabras individuales; y

• el cálculo de una medida de confianza (C_{DC}) para dicha hipótesis de reconocimiento, sobre la base de dicha señal de voz de entrada;

en el que el cálculo de una medida de confianza (C_{DC}) incluye:

• el cálculo de contribuciones diferenciales ($C_{i,t}$) sobre dicha medida de confianza (C_{DC}), cada una como una diferencia entre un resultado acústico forzado ($SCV_{i,t}$) y un resultado acústico no forzado (SCL_t), donde el resultado forzado puede definirse como la probabilidad de que la voz de entrada se corresponda con la hipótesis de reconocimiento, con las restricciones de conocimiento esquematizadas como conocimiento de red de estados finitos, y el resultado no forzado está definido como la probabilidad de que la voz de entrada se corresponda con la hipótesis de reconocimiento sin las restricciones de red de estados finitos; y

• el cálculo de dicha medida de confianza (C_{DC}) sobre la base de dichas contribuciones diferenciales ($C_{i,t}$);

caracterizado por el hecho de que el cálculo de dicha medida de confianza (C_{DC}) sobre la base de dichas contribuciones diferenciales ($C_{i,t}$) incluye:

• ponderar cada contribución diferencial ($C_{i,t}$) con una función de distribución acumulativa ($F_{\xi i}$) de la contribución diferencial ($C_{i,t}$).

2. Procedimiento de reconocimiento de voz automático según la reivindicación 1, en el que dichas funciones de distribución acumulativas ($F_{\xi i}$) de las contribuciones diferenciales ($C_{i,t}$) son tales que las distribuciones de las medidas de confianza (C_{DC}) sean homogéneas en términos de capacidad de rechazo, puesto que al menos uno de entre el lenguaje, el vocabulario y la gramática varía.

3. Procedimiento de reconocimiento de voz automático según la reivindicación 1 ó la 2, en el que cada contribución diferencial ($C_{i,t}$) se calcula según la siguiente ecuación:

$$C_{i,t} = F_{\xi i}\,(SCV_{i,t} - SCL_t)$$

donde $C_{i,t}$ es dicha contribución diferencial, $F_{\xi i}$ es dicha función de distribución acumulativa, $SCV_{i,t}$ es dicho resultado acústico forzado para el estado acústico i de un modelo oculto de Markov que modeliza una unidad acústico-fonética determinada, la cual, a su vez, modeliza un fenómeno de voz correspondiente, al nivel de la trama de voz de entrada t, y SCL_t es dicho resultado acústico no forzado al nivel de la trama de voz de entrada t.

4. Procedimiento de reconocimiento de voz automático según la reivindicación 3, en el que dicho resultado acústico forzado ($SCV_{i,t}$) se calcula según la siguiente ecuación:

$$SCV_{i,t} = \log P(S_i \mid o_t)$$

donde $P(S_i \mid O_t)$ es la probabilidad de una emisión de estados acústicos S_i, dado un vector de observación O_t extraído de la señal de voz de entrada al nivel de la trama de voz de entrada *t*.

5. Procedimiento de reconocimiento de voz automático según la reivindicación 4, en el que dicho resultado acústico no forzado (SCL_t) se calcula según la siguiente ecuación:

$$SCL_t = \log\left[\max_{1 \le j \le N_s} P(S_j \mid o_t)\right]$$

donde N_s es el número total de estados de los modelos ocultos de Markov.

6. Procedimiento de reconocimiento de voz automático según la reivindicación 5, en el que cada contribución diferencial ($C_{i,t}$) se calcula según la siguiente ecuación:

$$C_{i,t} = F_{\xi_i}\left\{\log \frac{P(S_i \mid o_t)}{\max\limits_{1 \le j \le N_s} P(S_j \mid o_t)}\right\}$$

7. Procedimiento de reconocimiento de voz automático según cualquiera de las reivindicaciones anteriores, en el que dicha medida de confianza (C_{DC}) se calcula como media temporal de dichas contribuciones diferenciales ponderadas ($C_{i,t}$) sobre la secuencia de los estados acústicos reconocidos (S_i*) de las unidades acústico-fonéticas elementales que forman dicho reconocimiento.

8. Procedimiento de reconocimiento de voz automático según la reivindicación 6 o la 7, en el que dicha medida de confianza (C_{DC}) se calcula según la siguiente ecuación:

$$C_{DC} = \frac{1}{T}\sum_{t=1}^{T} F_{\xi_{i^*}}\left\{\log \frac{P(S_{i^*} \mid o_t)}{\max\limits_{1 \le j \le N_s} P(S_j \mid o_t)}\right\}$$

donde T es el número de tramas de la señal de voz de entrada.

9. Procedimiento de reconocimiento de voz automático según cualquiera de las reivindicaciones anteriores 4 a 6, en el que dicha probabilidad de emisión es una probabilidad *a posteriori* calculada a partir de una probabilidad basada en un modelo oculto de Markov.

10. Procedimiento de reconocimiento de voz automático según cualquiera de las reivindicaciones anteriores 4 a 6, en el que dicha probabilidad de emisión es una probabilidad *a posteriori* calculada por una red neuronal artificial de un sistema de red neuronal del modelo oculto de Markov híbrido.

11. Procedimiento de reconocimiento de voz automático según cualquiera de las reivindicaciones anteriores 4 a 6, en el que dicha probabilidad de emisión se calcula por un sistema de red neuronal del modelo oculto de Markov híbrido según la siguiente ecuación:

$$P(S_i \mid o_t) = \frac{e^{n_i(o_t)}}{\sum\limits_{j=1}^{N} e^{n_j(o_t)}}$$

donde $n_i(o_t)$ es la salida de dicha red neuronal y es indicativa de la actividad del estado acústico S_i, dado el vector de observación o_t.

12. Procedimiento de reconocimiento de voz automático según la reivindicación 11, en el que dicha contribución diferencial ($C_{i,t}$) se calcula de la manera siguiente:

$$C_{i,t} = F_{\xi_i}\left\{\log \frac{e^{n_i(o_t)}}{e^{\max\limits_j n_j(o_t)}}\right\} = F_{\xi_i}\left[n_i(o_t) - \max_j n_j(o_t)\right]$$

13. Procedimiento de reconocimiento de voz automático según la reivindicación 12, en el que dicha medida de confianza (C_{DC}) se calcula de la manera siguiente:

$$C_{DC} = \frac{1}{T}\sum_{t=1}^{T} F_{\xi_{i^*}}\left[n_{i^*}(o_t) - \max_{j \in stat} n_j(o_t)\right]$$

donde T es el número de tramas de la señal de voz de entrada.

14. Procedimiento de reconocimiento de voz automático según cualquiera de las reivindicaciones anteriores, en el que cada función de distribución acumulativa ($F_{\xi i}$) está basada en una función de densidad de probabilidad con una de forma gausiana, una de forma binomial, una de forma de Poisson, y una de forma Gamma.

15. Procedimiento de reconocimiento de voz automático según cualquiera de las reivindicaciones anteriores 1 a 13, en el que cada función de distribución acumulativa ($F_{\xi i}$) es capacitada a partir de datos de voz de capacitación utilizando procedimientos de adaptación.

16. sistema de reconocimiento de voz automático para identificar palabras de una señal de voz de entrada, que incluye:

• un motor de búsqueda (6) configurado para suministrar al menos una hipótesis de reconocimiento a partir de dicha señal de voz de entrada, siendo dicha hipótesis de reconocimiento una palabra de hipótesis individual o una secuencia de palabras de hipótesis individuales; y

• un módulo de medida de confianza (8) configurado para calcular una medida de confianza (C_{DC}) para dicha hipótesis de reconocimiento, a partir de dicha señal de voz de entrada;

en el que dicho módulo de medida de confianza (8) está configurado para calcular dicha medida de confianza (C_{DC}) llevando a cabo las etapas siguientes:

• el cálculo de contribuciones diferenciales ($C_{i,t}$) sobre dicha medida de confianza (C_{DC}), cada una como una diferencia entre un resultado acústico forzado ($SCV_{i,t}$) y un resultado acústico no forzado (SCL_t) donde el resultado forzado puede definirse como la probabilidad de que la voz de entrada se corresponda con la hipótesis de reconocimiento, con las restricciones de conocimiento esquematizadas como conocimiento de red de estados finitos, y el resultado no forzado está definido como la probabilidad de que la voz de entrada se corresponda con la hipótesis de reconocimiento sin las restricciones de red de estados finitos;

• el cálculo de dicha medida de confianza (C_{DC}) sobre la base de dichas contribuciones diferenciales ($C_{i,t}$); **caracterizado** por el hecho de que el cálculo de dicha medida de confianza (C_{DC}) sobre la base de dichas contribuciones diferenciales ($C_{i,t}$) incluye:

• ponderar cada contribución diferencial ($C_{i,t}$) con una función de distribución acumulativa ($F_{\xi i}$) de la contribución diferencial ($C_{i,}t$).

17. Producto de programa informático que comprende un código de programa informático adaptado para implementar el procedimiento según cualquiera de las reivindicaciones anteriores 1 a 15.

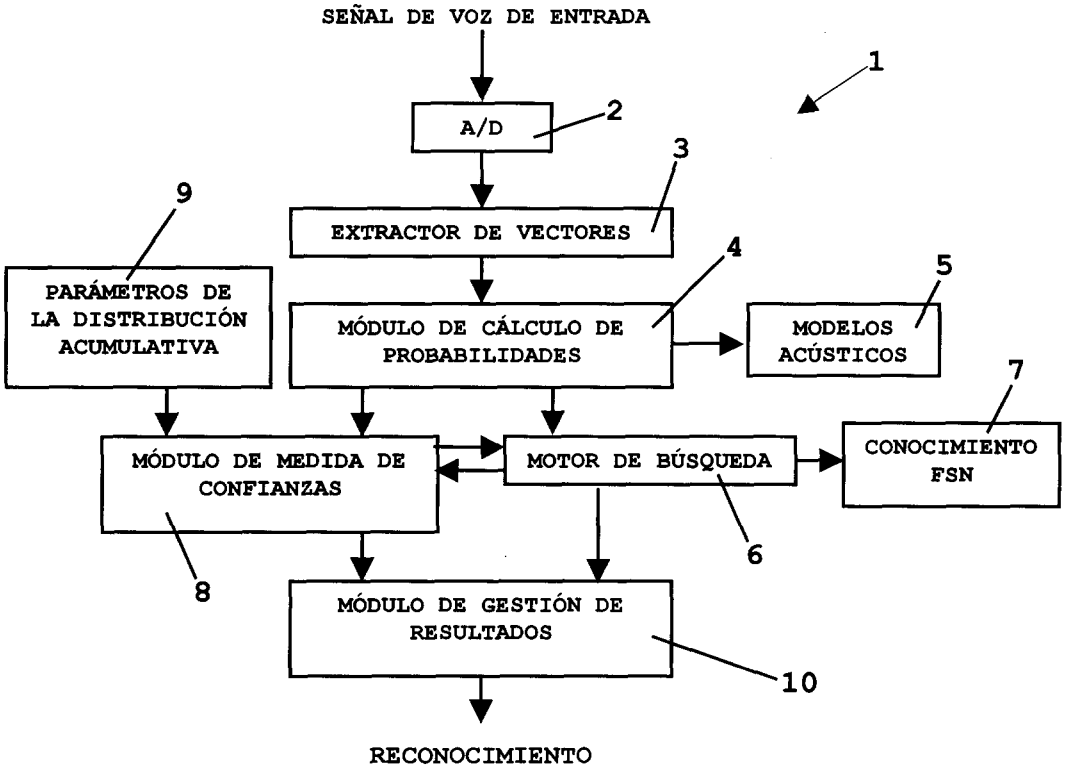

Figura 1

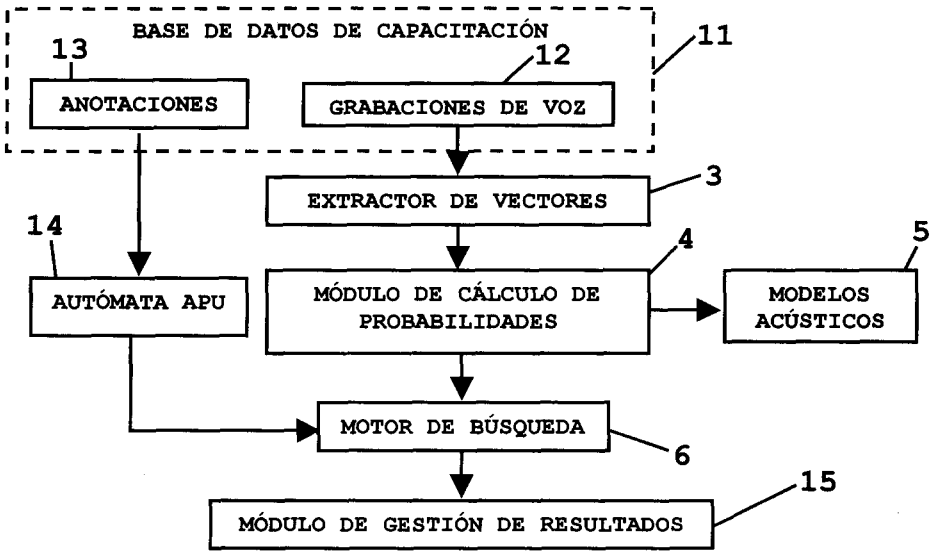

Figura 4

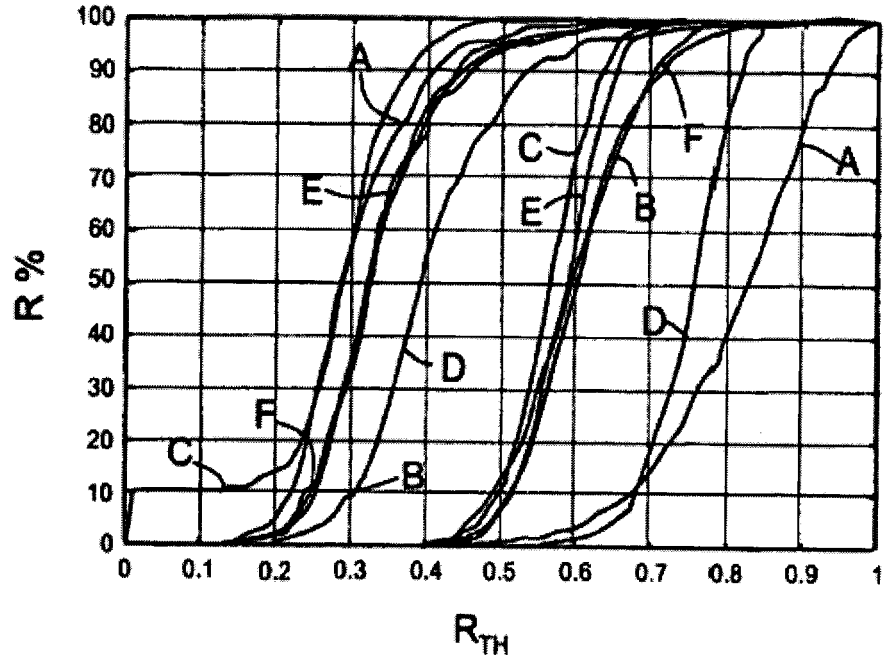
R %
100
90
80
70
60
50
40
30
20
10
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
A
E
C
F
B
D
R_{TH}

Fig.2

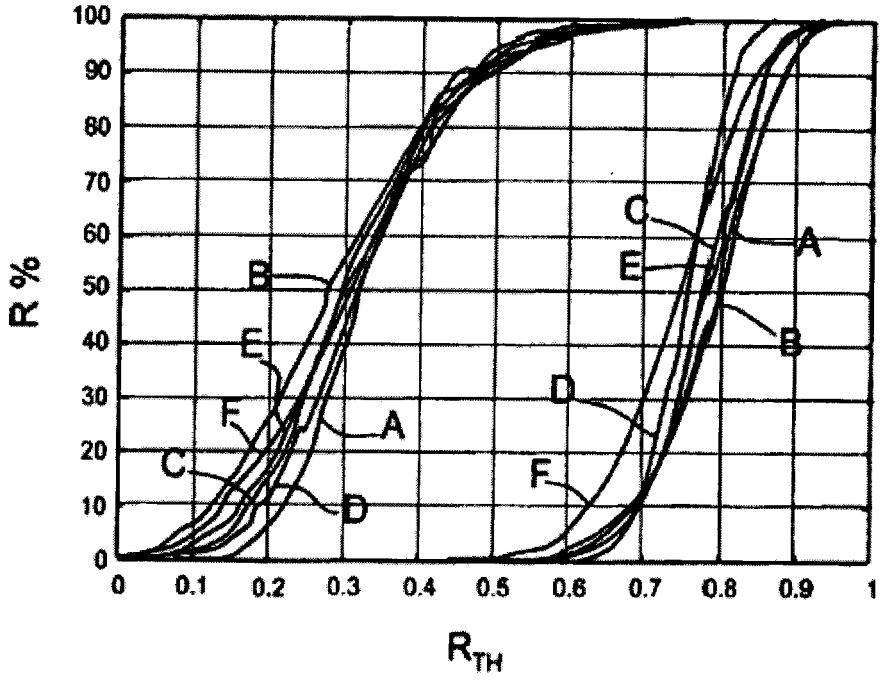
R %
100
90
80
70
60
50
40
30
20
10
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
B
E
F
C
A
D
R_{TH}

Fig.3