(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 327 468**

(51) Int. Cl.:
G10L 15/06 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **07756698 .2**
(96) Fecha de presentación : **06.02.2007**
(97) Número de publicación de la solicitud: **1979894**
(97) Fecha de publicación de la solicitud: **15.10.2008**

(54) Título: **Reconocimiento de voz con adaptación del hablante basándose en la clasificación del tono.**

(30) Prioridad: **21.02.2006 US 358001**

(45) Fecha de publicación de la mención BOPI:
29.10.2009

(45) Fecha de la publicación del folleto de la patente:
29.10.2009

(73) Titular/es: **Sony Computer Entertainment Inc.**
2-6-21, Minami-Aoyama
Minato-ku, Tokyo 107-0062, JP

(72) Inventor/es: **Chen, Ruxin**

(74) Agente: **Elzaburu Márquez, Alberto**

DESCRIPCIÓN

Reconocimiento de voz con adaptación del hablante basándose en la clasificación del tono.

Campo del invento

Esta aplicación se refiere al reconocimiento de voz y más específicamente a los sistemas de reconocimiento de voz que se adaptan a hablantes basándose en la altura de tono.

Antecedentes del invento

Las tecnologías de reconocimiento de voz y habla permiten que los ordenadores y otros dispositivos electrónicos equipados con una fuente de entrada de sonido, tal como un micrófono, interpretar el habla humana, por ejemplo para transcripción o como un método alternativo de interacción con un ordenador. Se están desarrollando soportes lógicos de reconocimiento de habla para uso en dispositivos electrónicos de consumo tales como teléfonos móviles, consolas para juegos, ordenadores personales y asistentes digitales personales. En un algoritmo típico de reconocimiento de habla, una señal de dominio de tiempo que representa el habla humana es partida en varias ventanas de tiempo y cada ventana es convertida en una señal de dominio de frecuencia, por ejemplo por la transformación rápida de Fourier (FFT). Esta señal de frecuencia o de dominio espectral es comprimida a continuación tomando un logaritmo de la señal de dominio espectral y después realizando otra FFT. Del espectro comprimido (denominado también un cepstrum), se puede usar un modelo estadístico para determinar los fenómenos y el contexto dentro del habla representado por la señal. El cepstrum puede ser visto como información sobre la tasa de cambio de las diferentes bandas espectrales dentro de la señal de habla. Para aplicaciones de reconocimiento de habla el espectro es usualmente transformado primero usando las bandas de Frecuencia Mel. El resultado se denomina los Coeficientes Cepstral de Frecuencia Mel o MFCCs. Una frecuencia f en hercios (ciclos por segundo) puede convertirse en una frecuencia mel *m* de acuerdo con: m=(1127,01048 Hz)$\log_e$(1+f/700). Igualmente una frecuencia mel *m* puede convertirse en una frecuencia *f* en hercios usando f=(700 Hz)($e^{m/1127,01048}$-1).

En el reconocimiento de voz el espectro es a menudo filtrado usando un conjunto de funciones de filtro de forma triangular. Las funciones de filtro dividen el espectro en un conjunto de bandas que se solapan parcialmente y que se encuentran entre una frecuencia mínima f_{min} y una frecuencia máxima f_{max}. Cada función de filtro está centrada en una determinada frecuencia dentro de una gama de frecuencias de interés. Cuando se ha convertido a la escala de frecuencias mel cada función de filtro puede ser expresada como un conjunto de baterías de filtros mel, en donde cada batería de filtros MFB está dada por:

$$MFB_i = \left(\frac{mf - mf_{min}}{mf_{max} - mf_{max}} \right) i$$

en donde el índice i se refiere al número de batería de filtros y mf_{min} y mf_{max} son las frecuencias mel que corresponden a f_{min} y f_{max}.

La elección de f_{min} y f_{max} determina las baterías de filtros que son usadas por un algoritmo de reconocimiento de voz. Típicamente, f_{min} y f_{max} son fijadas por el modelo de reconocimiento de voz que se está usando. Un problema con el reconocimiento de voz es que los diferentes hablantes pueden tener longitudes diferentes del tracto vocal y producir señales de voz con las correspondientes gamas de frecuencia diferentes. Para compensar lo anterior, los sistemas de reconocimiento de voz pueden realizar una normalización del tracto vocal de la señal de voz antes del filtrado. A modo de ejemplo, la normalización puede usar una función del tipo:

$$f' = f + \frac{1}{\pi \ arctg \ \alpha \left(\frac{sen\ (2\pi f)}{1 - \alpha \cos\ (2\pi f)} \right)}$$

en donde *f'* es la frecuencia normalizada y α es un parámetro que ajusta una curvatura de la función de normalización.

Los componentes de una señal de habla que tiene N bandas diferentes de frecuencia mel pueden ser representados como un vector A que tiene N componentes. Cada componente del vector A es un coeficiente de frecuencia mel de la señal de habla. La normalización del vector A típicamente implica una matriz de transformación del tipo:

F'=[M]·F+B, en donde [M] es una matriz NxN dada por:

$$[M]=\begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1N} \\ M_{21} & M_{22} & \cdots & M_{2N} \\ M & M & M & M \\ M_{N1} & M_{N2} & \cdots & M_{NN} \end{bmatrix}$$

y B es un vector de polarización dado por:

$$B=\begin{bmatrix} B_1 \\ B_2 \\ M \\ B_N \end{bmatrix}$$

F' y F son vectores de la forma:

$$F=\begin{bmatrix} F_1 \\ F_2 \\ M \\ F_N \end{bmatrix}, \qquad F'=\begin{bmatrix} F'_1 \\ F'_2 \\ M \\ F'_N \end{bmatrix}$$

en donde los coeficientes M_{ij} de la matriz y los componentes B_i del vector se calculan fuera de línea para maximizar la probabilidad de una secuencia de habla observada en un sistema HMM. Usualmente, para una trama dada y una característica dada F', la probabilidad observada es la calculada por una función Gaussiana:

$$Gaussiana \quad (F'_0 \ldots F'_n)=\frac{1}{\sqrt{\delta}}\exp\left(-\sum_i \frac{(F'_i-\mu_{Hi})^2}{2\cdot\sigma^2_{Hi}}\right)$$

Cada componente del vector normalizado F' es un componente de la frecuencia mel de la señal de habla.

Se sabe que los hablantes hombre y mujer producen señales de voz caracterizadas por coeficientes de frecuencia mel diferentes (MFCC). En la técnica anterior, los sistemas de reconocimiento de voz han realizado un entrenamiento para diferenciar si el hablante es hombre o mujer y ajustar el modelo acústico usado en el reconocimiento de voz basado en si el hablante es hombre o mujer. Típicamente, el modelo acústico se ensaya haciendo que un número, por ejemplo 10, de hablantes hombre y un número igual de hablantes mujer digan las mismas palabras para producir muestras de voz. Los análisis de características basados en muestras de voz dicen las mismas palabras para producir muestras de voces. Los análisis de características basados en muestras de voz se combinan conjuntamente en un supermodelo de reconocimiento de voz.

Una desventaja importante de la normalización anterior es que el vector F puede tener hasta 40 componentes. Por lo tanto, la matriz [M] podría tener hasta 1.600 coeficientes. El cálculo de un número tan alto de coeficientes puede emplear mucho tiempo para adaptar el algoritmo de reconocimiento de voz.

Además, como los sistemas y métodos de reconocimiento de voz de la técnica anterior usan valores fijos de f_{min}, f_{max}, mf_{min} y mf_{max} para filtrado y normalización, no tienen en cuenta adecuadamente las variaciones en la longitud del tracto vocal entre hablantes. Por lo tanto, la precisión en el reconocimiento de habla puede ser menor que la óptima. Por tanto, existe la necesidad de sistemas y métodos de reconocimiento de voz que superen tales desventajas.

Posteriores disposiciones de la técnica anterior se discuten en el documento EP-A-0.866.442, la ponencia "Normalización de hablantes usando los procedimientos eficientes de distorsión de frecuencia", Li Lee y otros, Proc 1996 IEEE Conferencia Internacional sobre procedimientos de Acústica, Habla y Señales, vol 1, Conf 21, páginas 353-356, y la ponencia "Normalización de hablantes basada en establecimiento de escalas no uniformes", Proc 2002 IEEE Conferencia Internacional sobre Procedimientos de Acústica, Habla y Señales, vol 4, páginas 1-589.

Compendio del invento

Los diferentes aspectos referentes al invento están definidos en las reivindicaciones que se acompañan.

Las desventajas asociadas con la técnica anterior están superadas, o al menos atenuadas por realizaciones del presente invento dirigidas a métodos y sistemas de reconocimiento de voz. De acuerdo con realizaciones del invento se obtiene una señal de voz a partir de una pronunciación de un hablante. Se determina una altura de tono de ejecución a partir de la señal de voz de la pronunciación. El hablante es clasificado basándose en la altura de tono de ejecución y se ajustan uno o más parámetros de modelos acústicos basándose en una clasificación del hablante. A continuación se realiza un análisis de reconocimiento de voz de la pronunciación basándose en los parámetros del modelo acústico.

Breve descripción de los dibujos

A continuación se describen las realizaciones del invento, solamente a modo de ejemplo, haciendo referencia a los dibujos que se acompañan, en los que:

la Figura 1 es un diagrama de flujos que ilustra un algoritmo de reconocimiento de voz de acuerdo con una realización del presente invento,

la Figura 2 es un diagrama de bloques que ilustra un sistema de reconocimiento de voz de acuerdo con una realización del presente invento.

Descripción de las realizaciones específicas

Aunque la descripción detallada que sigue contiene muchos detalles específicos para los fines de ilustración, cualquier persona con un conocimiento normal de la técnica observará que muchas variaciones y alteraciones en los siguientes detalles están dentro del alcance del invento. Por lo tanto, las realizaciones del invento que se describen más adelante se exponen sin ninguna pérdida de generalidad a, y sin imponer limitaciones sobre, el invento reivindicado.

De acuerdo con una realización del presente invento, un método de reconocimiento de voz 100 puede seguir su curso como está ilustrado en la Figura 1A. En 102, se obtiene una señal de voz a partir de una pronunciación de un hablante. La señal de voz puede ser obtenida de cualquier forma convencional, por ejemplo usando un micrófono y un codificador digital de la forma de la onda para poner la señal de voz en formato digital. La señal de voz puede obtenerse por sobremuestreo de la señal de voz en una frecuencia de muestreo que es mayor que la frecuencia de análisis de una característica de trabajo. En particular, la frecuencia de muestreo puede ser mayor que la tasa de muestreo de habla durante un tiempo de ensayo. A modo de ejemplo, y sin limitación, si la señal de voz está caracterizada por una frecuencia de análisis de característica de trabajo de 12 kilohercios, la señal puede ser muestreada a una frecuencia de muestreo de, por ejemplo, 16-22 kilohercios.

En 104, se determina un valor de altura de tono de ejecución p_{run} de la pronunciación. Existen varias formas de determinar la altura de tono de ejecución p_{run}. Por ejemplo, p_{run} puede ser una altura de tono media móvil $p_{avg}(t)$ que puede ser calculada a lo largo de una ventana de tiempo dada que incluye el tiempo t por:

(Ecuación 1)

$$p_{avg}(t) = \frac{1}{NP} \sum_{i} p(t_i),$$

en donde la suma se realiza sobre una cantidad de NP medidas de altura de tono tomadas en los momentos t_i = {t-(NP-1), t-(NP-2), ..., t} durante la ventana de tiempo para probabilidades de altura de tono por encima de un umbral predeterminado. Una forma sencilla de calcular la probabilidad de altura de tono es

$$prob\ (pitch) = \frac{correlatio\ n\left(\frac{1}{pitch}\right)}{correlatio\ n(0)}$$

donde

$$correlatio\ n(t) = \sum_{i} signal\ (t+i)\, signal\ (i)$$

es la correlación de la señal de habla de análisis. Alternativamente, la altura de tono de ejecución p_{run} puede estar relacionada con la altura de tono actual, por ejemplo, por:

$$\text{(Ecuación 2)} \quad p_{run}(t) = c.p_{run}(t-1)+(1-c.p(t)), \text{ para } t>0$$

y

$$p_{run}(0)= p(0), \text{ para } t=0$$

en donde c es una constante entre 0 y 1 y $p(t)$ es un valor actual de la altura de tono en el momento t. El valor de la constante c está relacionado con el tamaño de la ventana. Por ejemplo, un valor de c=0 corresponde a no ventana (en cuyo caso $p_{run}(t)= p(t)$) y un valor de c=1 corresponde con una ventana infinita (en cuyo caso $p_{run}(t)=p(t-1)$). Adviértase que para valores de t>0, los valores de altura de tono para momentos anteriores a t contribuyen al valor de la altura de tono $p_{run}(t)$. Esto puede ser ilustrado con un ejemplo numérico en el que c=0,6. En tal caso, la ecuación 2 da:

$$p_{run}(0)=p(0)$$

$$p_{run}(1) = 0,6c.p_{run}(0)+(1-c).p(1) = 0,6\cdot p(0)+0,4\ p(1)$$

$$p_{run}(2) = 0,6c.p_{run}(1)+(1-c).p(2)) = 0,6\cdot(0,6p(0)+0,4\ p(1)) + 0,4\cdot p(2)$$

En algunas realizaciones del invento $p_{run}(t)$ puede ser calculado de acuerdo con la Ecuación 2 si la probabilidad de altura de tono está por encima de algún umbral, por ejemplo, por encima de aproximadamente 0,4.

A modo de ejemplo, la clasificación del hablante realizada en 106 de la Figura 1A puede estar basada en la edad del hablante y/o en el género. Por ejemplo, a partir de los datos de entrenamiento se puede determinar que la altura de tono media de hombres, mujeres y niños se encuentra en diferentes gamas. El hablante puede ser clasificado a partir de la gama en la que se encuentra la altura de tono actual de la señal de voz. A modo de ejemplo, un hablante varón adulto tiene una altura de tono media entre aproximadamente 120 Hz y aproximadamente 160 Hz, un hablante hembra adulto tiene una altura de tono media entre aproximadamente 180 Hz y aproximadamente 220 Hz, y un hablante niño tiene una altura de tono media mayor de 220 Hz. Si la altura de tono actual es 190 Hz, el hablante sería clasificado como un hablante hembra. En algunos casos la altura de tono media de un hablante puede estar incluida como una característica en el vector F.

Una vez que el hablante ha sido clasificado, los parámetros del modelo acústico pueden ser seleccionados de acuerdo con lo indicado en 108. Estos parámetros se usan después en un análisis de reconocimiento de voz en 110. La elección de parámetros depende del tipo de modelo acústico usado en el análisis de reconocimiento de voz. Por ejemplo, el análisis de reconocimiento de voz puede filtrar la señal de voz usando un conjunto de funciones de filtro. Las funciones de filtro, por ejemplo las funciones de filtro de forma triangular, dividen el espectro en un conjunto de bandas que se solapan parcialmente. Cada análisis de reconocimiento de voz usa una batería de filtros definida por una frecuencia máxima f_{max} diferente y una frecuencia mínima f_{min} diferente. Las f_{max} y f_{min} pueden ser frecuencias en la escala de Hertz o alturas de tono en la escala mel. La frecuencia máxima f_{max} se refiere a un límite superior de la gama de frecuencias de la batería de filtros, y la frecuencia mínima f_{min} se refiere a un límite inferior de la gama de frecuencias de la batería de filtros. Los valores de los parámetros f_{max} y f_{min} pueden ajustarse dinámicamente en cualquier momento durante el análisis de reconocimiento de voz, por ejemplo, para cualquier ventana de tiempo durante el análisis de reconocimiento de voz. El análisis de reconocimiento de voz produce una probabilidad de reconocimiento P, de reconocimiento de una o más unidades de habla. Las unidades de habla pueden ser frases, palabras, o subunidades de palabras, tales como fonemas.

A modo de ejemplo, una vez que el hablante ha sido clasificado como un varón, hembra o niño, los valores de f_{max} y f_{min} para análisis de reconocimiento de voz de la pronunciación pueden ser por lo tanto seleccionados. Por ejemplo, si se acepta que el hablante es un hombre, f_{max} puede ser aproximadamente 70 Hz y f_{min} puede ser aproximadamente 3.800 Hz. Si se supone que el hablante es una mujer, f_{max} puede ser aproximadamente 70 Hz y f_{min} puede ser aproximadamente 4.200 Hz. Si se supone que el hablante es un niño, f_{max} puede ser aproximadamente 90 Hz y f_{min} puede ser aproximadamente 4.400 Hz.

En 110, una probabilidad *P* de reconocimiento procede de un análisis de voz de la pronunciación basado en los parámetros del modelo de ajuste. A modo de ejemplo, y sin pérdida de generalidad, el análisis de reconocimiento de voz puede usar un Modelo de Markov Hidden (HMM) para determinar las unidades de habla en una señal de voz dada. Las unidades de habla pueden ser palabras, combinaciones de dos palabras o unidades de subpalabras y similares. El HMM puede estar caracterizado por:

L, que representa varios posibles estados del sistema;

M, que representa el número total de Gaussianos que existen en el sistema;

N, que representa el número de características distintas observables en un instante dado; pudiendo estas características ser espectrales (por ejemplo, dominio de frecuencia) o temporal (dominio de tiempo) de la señal del habla;

$A=\{a_{ij}\}$, una distribución de probabilidad de transición de estado, en la que cada a_{ij} representa la probabilidad de que el sistema pase del estado j^{th} en el instante t+1 si el sistema está inicialmente en el estado j^{th} en el instante t;

$B=\{b_j(k)\}$, una distribución de probabilidad de característica de observación del estado j^{th}, en el que cada $b_j(k)$ representa la distribución de probabilidad de los valores observados de la característica k^{th} cuando el sistema está en el estado j^{th}; y

$\pi=\{\pi_j\}$, una distribución de estados iniciales, en donde cada componente π_j representa la probabilidad de que el sistema esté en el estado j^{th} en algún momento inicial.

Los Modelos Markov Hidden pueden aplicarse a la señal de voz para resolver uno o más problemas básicos, que incluyen: (1) la probabilidad de una secuencia dada de observaciones obtenidas de la señal de voz; (2) dada la secuencia de observaciones, qué secuencia de estado correspondiente explica mejor la secuencia de observaciones; y (3) cómo ajustar el conjunto de parámetros modelo A, B, π para maximizar la probabilidad de una secuencia de observaciones dada.

La aplicación de HMMs para el reconocimiento de habla es descrita con detalle, por ejemplo, por Lawrence Rabiner en "Un curso de enseñanza sobre modelos de Markov Hidden y en Aplicaciones Seleccionadas sobre Reconocimiento de Habla" en las Actas del IEEE, Vol 77, Nº 2, 2 de febrero de 1989.

Los análisis de reconocimiento de habla realizados en 110 pueden caracterizar el habla mediante varios patrones reconocibles conocidos como fonemas. Cada uno de estos fonemas puede ser dividido en varias partes, por ejemplo, una parte inicial, media y final. Se ha observado que la parte media es típicamente la más estable ya que la parte inicial está a menudo afectada por el fonema anterior y la parte final está afectada por el fonema que sigue. Las diferentes partes de los fonemas están caracterizadas por características de dominio de frecuencia que pueden ser reconocidas por un análisis estadístico apropiado de la señal. El modelo estadístico usa a menudo funciones de distribución de probabilidad Gaussianas para predecir la probabilidad de cada estado diferente de las características que constituyen partes de la señal correspondientes a las diferentes partes de los diferentes fonemas. Un estado HMM puede contener uno o más Gaussianos. Un Gaussiano determinado para un estado posible dado, por ejemplo, el Gaussiano k^{th} puede ser representado por un conjunto de N valores medios μ_{ki} y por las varianzas σ_{ki}. En un algoritmo de reconocimiento de habla típico uno determina cuál de los Gaussianos para una ventana de tiempo dada es el mayor. A partir del Gaussiano mayor se puede deducir el fonema más probable para la ventana de tiempo.

A modo de ejemplo, el análisis de reconocimiento de voz en 110 puede analizar una señal de dominio de tiempo para obtener N características de señal observables diferentes $x_0...x_n$, en donde n=N-1. La característica observada del sistema puede ser representada como un vector que tiene los componentes $x_0...x_n$. Estos componentes pueden ser características espectrales, cepstrales, o temporales de un señal de habla observada dada.

A modo de ejemplo, y sin limitación de las realizaciones del invento, los componentes $x_0...x_n$ pueden ser coeficientes cepstrales de frecuencia mel (MFFCs) de la señal de voz obtenida en 102. Un cepstrum es el resultado de tomar la transformación de Fourier (FT) del espectro de decibelios como si fuera una señal. El cepstrum de una señal de habla del dominio de tiempo puede ser definido verbalmente como la transformada de Fourier del log (con fase no desarrollada) de la transformación de Fourier de la señal del dominio de tiempo. El cepstrum de una señal del dominio de tiempo S(t) puede representarse matemáticamente como $FT(\log(FT(S(t)))+j2\pi q)$, en donde q es el entero necesario para desarrollar apropiadamente el ángulo o parte imaginaria de la función logarítmica compleja. Algorítmicamente, el cepstrum puede ser generado por la secuencia de operaciones: señal -> FT -> log -> desarrollo de fase -> FT -> cepstrum.

Hay un cepstrum complejo y un cepstrum real. El cepstrum real usa la función logarítmica definida de valores reales, mientras que el cepstrum complejo usa la función logarítmica definida de valores complejos también. El cepstrum complejo contiene información sobre la magnitud y fase del espectro inicial, que permite la reconstrucción de la señal. El cepstrum real solamente usa la información de la magnitud del espectro. A modo de ejemplo, y sin pérdida de generalidad, el análisis de reconocimiento de voz realizado en 110 puede usar el cepstrum real.

Ciertos patrones de combinaciones de componentes $x_0...x_n$ corresponden a unidades de habla (por ejemplo palabras o frases) o a subunidades, tales como sílabas, fonemas u otras subunidades de palabras. Cada unidad o subunidad puede ser considerada como un estado del sistema. La función de densidad de probabilidad $f_k(x_0...x_n)$ de un Gaussiano dado del sistema (el Gaussiano kth) puede ser cualquier tipo de función de densidad de probabilidad $f_k(x_0...x_n)$, por ejemplo una función Gaussiana que tenga la siguiente forma:

$$f_k(x_0 \ldots x_n) = \frac{1}{\sqrt{\delta_k}} \exp\left[-\sum_i \frac{(x_i - \mu_{ki})^2}{2 \bullet \sigma_{ki}^2}\right] \qquad (1)$$

en donde

$$\delta_k = \prod_i (2\pi \bullet \sigma_{ki}^2)$$

$$i=1...N,\ k=1...M.$$

En las ecuaciones anteriores "i" es un índice de característica y "k" es un índice de Gaussiano. En la ecuación (1) el subíndice k es un índice de la función Gaussiana. Puede haber de varios cientos a varios cientos de miles de Gaussianos usados por el algoritmo de reconocimiento de habla. La cantidad μ_{ki} es un valor medio de la característica x_i en el Gaussiano k^{th} del sistema. La cantidad σ_{ki}^2 es la varianza de x_i en el Gaussiano k^{th}. Uno o más Gaussianos pueden estar asociados con uno o más estados diferentes. Por ejemplo, puede haber L estados diferentes que contienen un número total de M Gaussianos en el sistema. La cantidad μ_{ki} es la media de todas las medidas de x_i que pertenecen a $f_k(x_0...x_N)$ a lo largo de todas las ventanas de tiempo de datos de preparación y σ_{ki} es la varianza de las medidas correspondientes usadas para calcular μ_{ki}.

La probabilidad de que cada Gaussiano pueda ser calculado por la ecuación (1) para dar una probabilidad de reconocimiento correspondiente es $P_{r'}$. A partir del Gaussiano que tiene la probabilidad máxima uno puede construir un estado, palabra, fonema, carácter, etc más probable para esa determinada ventana de tiempo. Se debe observar que también es posible usar el estado más probable de una ventana de tiempo dada para ayudar a determinar el estado más probable de las ventanas de tiempo anteriores o posteriores, ya que éstas pueden determinar un contexto en el que ocurre el estado.

De acuerdo con realizaciones del presente invento, un método de reconocimiento (por ejemplo un método de reconocimiento de voz) del tipo representado en la Figura 1A o Figura 1B que funciona como se ha descrito antes puede ser realizado como parte de un aparato 200 de procesamiento de señal, como se ha representado en la Figura 2. El sistema 200 puede incluir un procesador 202 y una memoria 202 (por ejemplo RAM, DRAM, ROM, y similares). Además, el aparato de procesamiento de la señal 200 puede tener muchos procesadores 201 si se tiene que realizar un procesamiento paralelo. La memoria 202 incluye datos y códigos configurados como se ha descrito anteriormente. Específicamente, la memoria incluye datos que representan características 204 de la señal, y funciones de probabilidad 206, cada una de las cuales puede incluir códigos, datos o alguna combinación de códigos y datos.

El aparato 200 puede también incluir funciones de soporte bien conocidas 210, tal como los elementos entrada/salida (I/O) 211, suministros de potencia (P/S) 212, un reloj (CLK) 213 y memoria oculta 214. El aparato 200 puede opcionalmente incluir un dispositivo de almacenamiento masivo 215 tal como una unidad de disco, unidad de CD-ROM, unidad de cinta, o similar para almacenar programas y/o datos. El controlador puede también incluir opcionalmente una unidad de visualización 216 y una unidad de interfaz de usuario 218 para facilitar la interacción entre el controlador 200 y un usuario. La unidad de visualización 216 puede estar en la forma de un tubo de rayos catódicos (CRT) o de pantalla de panel plano que visualiza texto, números, símbolos gráficos o imágenes. La interfaz de usuario 218 puede incluir un teclado, un ratón, una palanca de control de mando, un lápiz óptico u otro dispositivo. Además, la interfaz de usuario 218 puede incluir un micrófono, una videocámara o cualquier otro dispositivo transductor de la señal para facilitar la captación directa de una señal para ser analizada. El procesador 201, la memoria 202 y los otros componentes del sistema 200 pueden intercambiar señales (por ejemplo, instrucciones de códigos y datos) entre sí a través de una vía de distribución 220 como se muestra en la Figura 2. Un micrófono 222 puede ser acoplado al aparato 200 a través de las funciones I/O 211.

Como se ha usado aquí, el término I/O generalmente se refiere a cualquier programa, operación o dispositivo que transfiera datos a o desde el sistema 200 y a o desde un dispositivo periférico. Cada transferencia es una salida de un dispositivo y una entrada a otro. Los dispositivos periféricos incluyen dispositivos de entrada solamente, tales como

teclados o ratones, dispositivos de salida solamente, tales como impresoras así como dispositivos tales como CD-ROM grabables que pueden actuar como un dispositivo de entrada y un dispositivo de salida. El término "dispositivo periférico" incluye dispositivos externos, tales como un ratón, un teclado, una impresora, un monitor, un micrófono, una cámara, una unidad de compresión externa o explorador así como dispositivos externos, tales como una unidad de CD-ROM, una unidad de CD-R o un modem interno u otro periférico tal como un lector/escritor de memoria instantáneo, una unidad de disco duro.

El procesador 201 puede realizar el reconocimiento de señal de los datos de las señales 206 y/o la probabilidad en las instrucciones de códigos de un programa 204 almacenado y recuperado por la memoria 202 y ejecutado por el módulo procesador 201. Las partes de código del programa 203 pueden adaptarse a cualquiera de varios lenguajes de programación diferentes tales como Assembly, C++, JAVA o varios otros lenguajes. El módulo de procesador 201 forma un ordenador de uso general que se convierte en un ordenador de uso específico cuando ejecuta programas tales como el código de programa 204. Aunque el código de programa 204 se describe aquí como realizado en soporte lógico y ejecutado en un ordenador de uso general, los expertos en la técnica advertirán que el método de gestión de tareas podría alternativamente realizarse usando un soporte físico tal como un circuito integrado específico de la aplicación (ASIC) u otros circuitos de soporte físico. Como tal, se debería entender que las realizaciones del invento pueden realizarse, total o parcialmente, en soporte lógico, soporte físico o en una combinación de ambos.

En una realización, entre otras, el código de programa 204 puede incluir un conjunto de instrucciones leíbles por un procesador que realiza un método que tiene características en común con el método 100 de la Figura 1A o el método 110 de la Figura 1B. El programa 204 puede incluir generalmente una o más instrucciones que dirigen el procesador 201 para obtener una señal de voz para una pronunciación de un hablante; determinar una altura de tono de ejecución de la señal de voz de la pronunciación; clasificar el hablante basándose en la altura de tono de ejecución; ajustar uno o más parámetros de modelo acústico basándose en una clasificación del hablante; y realizar un análisis de reconocimiento de voz de la pronunciación basándose en los parámetros del modelo acústico.

A modo de ejemplo, el programa 204 puede ser parte de un programa total mayor tal como un programa para un juego de ordenador. En ciertas realizaciones del invento el código de programa 204 puede hacer que un hablante diga una palabra o frase (por ejemplo, el nombre del hablante) durante una fase de iniciación (por ejemplo, al comienzo de un juego) para proporcionar un ejemplo de habla. A partir de esta muestra el programa 204 puede avanzar como se ha descrito anteriormente con respecto a la Figura 1 para encontrar parámetros óptimos (por ejemplo, f_{min} y f_{max}) de ese hablante y ejecutar el reconocimiento de voz en 110 usando esos parámetros. Los parámetros pueden ser guardados después de la conclusión del programa y ser usados cuando ese hablante use el programa.

Las realizaciones del presente invento facilitan un reconocimiento de voz más consistente y más preciso. En un ejemplo de reconocimiento de voz que emplea la selección del parámetro de modelo acústico usando la clasificación del hablante basándose en la altura de tono con un hablante hembra único produjo el 94,8% de precisión de palabras. Un algoritmo de reconocimiento de voz convencional que no emplea selección de parámetro de modelo acústico usando la clasificación del hablante basándose en la altura de tono con un hablante hembra único consiguió solamente el 86,3% de precisión de palabras con el mismo hablante hembra.

Mientras que lo anterior es una descripción completa de la realización preferida del presente invento, es posible usar diferentes alternativas, modificaciones y equivalentes.

El alcance del invento está definido por las reivindicaciones anejas.

REIVINDICACIONES

1. Un método de reconocimiento de voz, comprendiendo el método:

obtener (102) una señal de voz de una pronunciación de un hablante;

determinar (104) una altura de tono de ejecución de la señal de la pronunciación;

clasificar (106) el hablante basándose en la altura de tono de ejecución;

ajustar (108) uno o más parámetros del modelo acústico basándose en la clasificación del hablante; y

realizar (110) un análisis de reconocimiento de voz de la pronunciación basándose en los parámetros del modelo acústico.

2. El método de la reivindicación 1 en el que la determinación de la altura de tono de ejecución incluye la determinación de una altura de tono media $p_{avg}(t)$ en el instante t dado por:

$$p_{avg}(t) = \frac{1}{NP} \sum_{t_j} (t_j),$$

en donde la suma está tomada a lo largo de un número *NP* de medidas de altura de tono tomadas en los instantes t_j durante una ventana de tiempo.

3. El método de la reivindicación 2 en el que cada una de las alturas de tono $p(t_j)$ está por encima de un umbral predeterminado.

4. El método de la reivindicación 2 en el que la determinación de la altura de tono de ejecución incluye un cálculo del tipo:

$$p_{run}(t)=c.p_{run}(t-1)+(1-c.p(t)),$$

en donde c es una constante entre 0 y 1, y $p(t)$ es un valor actual de la altura de tono en el instante *t*.

5. El método de la reivindicación 1 en el que la clasificación del hablante incluye determinar la edad y/o género del hablante.

6. El método de la reivindicación 5 en el que la determinación de la edad y/o género del hablante incluye determinar si la altura de tono de ejecución está situada en una gama, en la que la gama depende de la edad y/o género del hablante.

7. El método de la reivindicación 5 en el que la determinación de la edad y/o género del hablante incluye determinar a partir de la altura de tono si el hablante es hombre, mujer o niño.

8. El método de la reivindicación 1 en el que uno o más modelos acústicos incluyen una frecuencia máxima f_{max} y una frecuencia mínima f_{min} de una batería de filtros usada en la realización del análisis de reconocimiento de voz.

9. El método de la reivindicación 8 en el que los valores de f_{max} y f_{min} son elegidos basándose en un género y/o una edad del hablante determinada durante la clasificación del hablante basada en la altura de tono de ejecución.

10. El método de la reivindicación 8 en el que los valores de f_{max} y f_{min} se eligen basándose en si el hablante es hombre, mujer o niño durante la clasificación del hablante basada en la altura de tono de ejecución.

11. El método de la reivindicación 8 en el que los valores de f_{max} y f_{min} se ajustan dinámicamente en cualquier momento durante el reconocimiento.

12. El método de la reivindicación 1 que además comprende almacenar la clasificación del hablante y/o uno o más parámetros de modelos acústicos basándose en la clasificación del hablante, y asociando la clasificación del hablante y/o uno o más parámetros del modelo acústico basándose en la clasificación del hablante con un determinado hablante.

13. El método de la reivindicación 11 que además comprende usar la clasificación del hablante almacenada y/o uno o más parámetros del modelo acústico basándose en la clasificación del hablante durante un análisis posterior de reconocimiento de voz del hablante.

14. Soporte lógico de ordenador que tiene un código de programa que, cuando es ejecutado por un ordenador, hace que el ordenador realice un método de acuerdo con cualquiera de las reivindicaciones 1 a 13.

15. Un sistema de reconocimiento de voz que comprende:

una interfaz (211) adaptada para obtener una señal de voz;

uno o más procesadores (20) acoplados a la interfaz; y

una memoria (202) acoplada a la interfaz y al procesador, teniendo la memoria incorporado en ella un conjunto de instrucciones leíbles por el procesador, configurada para realizar un método de reconocimiento de voz, incluyendo las instrucciones leíbles por el procesador:

una instrucción para obtener una señal de voz de una pronunciación de un hablante;

una instrucción para determinar una altura de tono de ejecución a partir de la señal de voz de la pronunciación;

una instrucción para clasificar el hablante basada en la altura de tono de ejecución;

una instrucción para ajustar uno o más parámetros del modelo acústico basada en una clasificación del hablante; y

una instrucción para realizar un análisis de reconocimiento de voz de la pronunciación basada en los parámetros del modelo acústico.

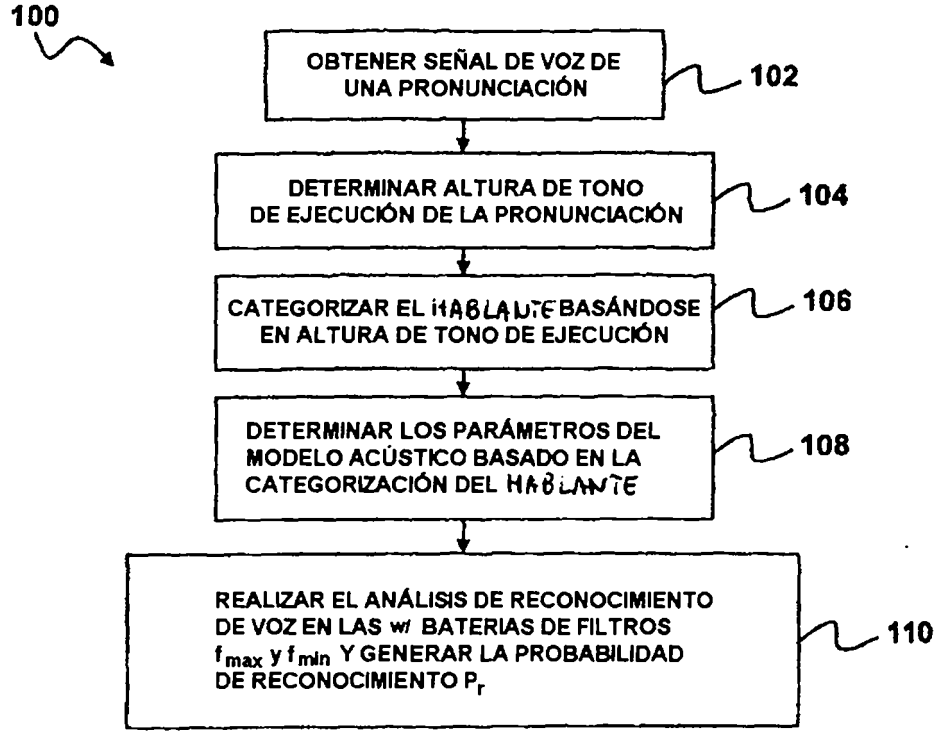
100
OBTENER SEÑAL DE VOZ DE UNA PRONUNCIACIÓN
102
DETERMINAR ALTURA DE TONO DE EJECUCIÓN DE LA PRONUNCIACIÓN
104
CATEGORIZAR EL HABLANTE BASÁNDOSE EN ALTURA DE TONO DE EJECUCIÓN
106
DETERMINAR LOS PARÁMETROS DEL MODELO ACÚSTICO BASADO EN LA CATEGORIZACIÓN DEL HABLANTE
108
REALIZAR EL ANÁLISIS DE RECONOCIMIENTO DE VOZ EN LAS w/ BATERIAS DE FILTROS f_{max} y f_{min} Y GENERAR LA PROBABILIDAD DE RECONOCIMIENTO P_r
110

FIG. 1

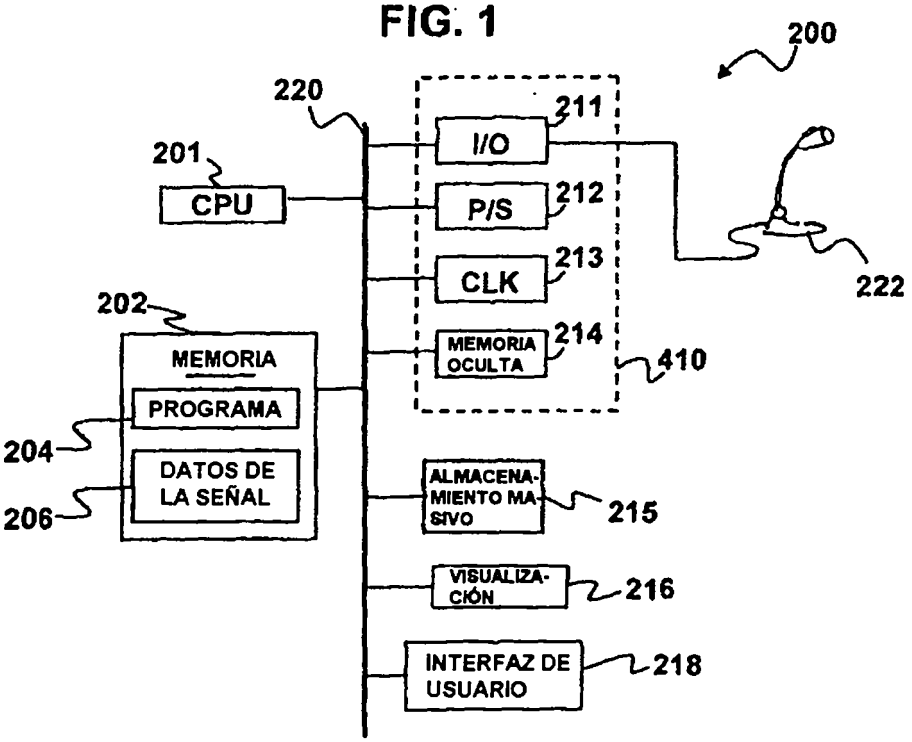
200
220
211
I/O
201
CPU
212
P/S
213
CLK
222
202
MEMORIA
214
MEMORIA OCULTA
410
PROGRAMA
204
DATOS DE LA SEÑAL
206
ALMACENA-MIENTO MA-SIVO
215
VISUALIZA-CIÓN
216
INTERFAZ DE USUARIO
218

FIG. 2