(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 357 674**

(51) Int. Cl.:
G10L 15/20 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **06742938 .1**
(96) Fecha de presentación : **16.05.2006**
(97) Número de publicación de la solicitud: **2022042**
(97) Fecha de publicación de la solicitud: **11.02.2009**

(54) Título: **Compensación de la variabilidad intersesión para extracción automática de información a partir de la voz.**

(45) Fecha de publicación de la mención BOPI:
28.04.2011

(45) Fecha de la publicación del folleto de la patente:
28.04.2011

(73) Titular/es: **LOQUENDO S.p.A.**
Via Arrigo Olivetti 6
10100 Torino, IT

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

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

DESCRIPCIÓN

Compensación de la variabilidad intersesión para extracción automática de información a partir de la voz.

Campo técnico de la invención

[0001] La presente invención se refiere en general a la extracción automática de información a partir de la voz, como el reconocimiento automático de hablante y habla, y en particular a un procedimiento y un sistema para compensar la variabilidad intersesión de características acústicas debida a entornos y canales de comunicación variables intersesión.

Técnica anterior

[0002] Como es sabido, un sistema de reconocimiento de hablante es un dispositivo capaz de extraer, almacenar y comparar características biométricas de la voz humana, y de realizar, además de una función de reconocimiento, también un procedimiento de aprendizaje, que permite el almacenamiento de características biométricas de la voz de un hablante en modelos apropiados, denominados comúnmente impresiones de voz. El procedimiento de aprendizaje ha de llevarse a cabo para todos los hablantes implicados y es preliminar a etapas de reconocimiento posteriores, durante las cuales los parámetros extraídos de una muestra de voz desconocida se comparan con los de las impresiones de voz para producir el resultado del reconocimiento.

[0003] Dos aplicaciones específicas de un sistema de reconocimiento de hablante son la verificación del hablante y la identificación del hablante. En el caso de la verificación del hablante, el propósito del reconocimiento es confirmar o rechazar una declaración de identidad asociada a la pronunciación de una frase o palabra. El sistema debe, es decir, responder a la pregunta: "¿Es el hablante la persona que dice que es?". En el caso de la identificación del hablante, el propósito del reconocimiento es identificar, de un conjunto finito de hablantes de cuyas impresiones de voz se dispone, al cual corresponde una voz desconocida. El propósito del sistema es en este caso responder a la pregunta: "¿A quién pertenece la voz?".

[0004] Una clasificación adicional de los sistemas de reconocimiento de hablante considera el contenido léxico utilizable por el sistema de reconocimiento: reconocimiento de hablante dependiente del texto o reconocimiento de hablante independiente del texto. El caso dependiente del texto requiere que el contenido léxico usado para la verificación o identificación correspondería al que se pronuncia para la creación de la impresión de voz: esta situación es típica en los sistemas de autentificación por voz, en los que la palabra o frase pronunciada asume, a todos los propósitos o efectos, la connotación de una contraseña de voz. El caso independiente del texto, en cambio, no establece ninguna restricción entre el contenido léxico de aprendizaje y el de reconocimiento.

[0005] Los modelos ocultos de Markov (HMMs) son una tecnología clásica usada para reconocimiento de habla y de hablante. En general, un modelo de este tipo está constituido por un cierto número de estados conectados por arcos de transición. Asociada a una transición está una probabilidad de pasar del estado de origen al estado de destino. Además, cada estado puede emitir símbolos procedentes de un alfabeto finito según una distribución de probabilidad dada. A cada estado está asociada una densidad de probabilidad, densidad de probabilidad que se define sobre un vector de características acústicas extraído de la voz a cuantías de tiempo fijas (por ejemplo, cada 10 ms), siendo generado dicho vector por un módulo de análisis acústico (terminal de entrada acústico), y se denomina generalmente vector de observación o de características. Los símbolos emitidos, basándose en la densidad de probabilidad asociada al estado, son, por lo tanto, los infinitos vectores de características posibles. Esta densidad de probabilidad está dada por una mezcla de gaussianos en el espacio multidimensional de los vectores de características. Ejemplos de características ampliamente usadas para reconocimiento de hablante son los coeficientes cepstrales en las frecuencias de Mel (MFCC), y las características de las derivadas de primer orden respecto al tiempo se suman a las características básicas.

[0006] En el caso de la aplicación de modelos ocultos de Markov al reconocimiento de hablante, además de los modelos HMM descritos previamente, con varios estados, se recurre frecuentemente a los llamados modelos de mezclas de gaussianos (GMMs). Un GMM es un modelo de Markov con un único estado y con un arco de transición hacia sí mismo. Generalmente, la densidad de probabilidad de GMMs está constituida por una mezcla de distribuciones gaussianas multivariantes con cardinalidad del orden de algunos miles de gaussianos. Las distribuciones gaussianas multivariantes se usan comúnmente para modelar los vectores de características de entrada multidimensionales. En el caso de reconocimiento de hablante independiente del texto, los GMMs

representan la categoría de modelos más ampliamente usados en la técnica anterior.

[0007] El reconocimiento de hablante se realiza creando, durante una etapa de aprendizaje, modelos adaptados a la voz de los hablantes implicados y evaluando la probabilidad que generan basándose en vectores de características extraídos de una muestra de voz desconocida, durante una etapa de reconocimiento posterior. Los modelos adaptados a hablantes individuales, que pueden ser HMMs o GMMs, se denominan comúnmente impresiones de voz. Una descripción de técnicas de aprendizaje de impresiones de voz que se aplica a GMMs y su uso para reconocimiento de hablante se proporciona en el documento de Reynolds, D.A. y col., Speaker verification using adapted Gaussian mixture models, Digital Signal Processing 10 (2000), págs. 19-41.

[0008] Una de las principales causas de degradaciones de rendimiento relevantes en el reconocimiento automático de habla y hablante es el desajuste acústico que ocurre entre condiciones de aprendizaje y de reconocimiento. En particular, en el reconocimiento de hablante, los errores se deben no sólo a la similitud entre impresiones de voz de diferentes hablantes, sino también a la variabilidad intrínseca de diferentes expresiones del mismo hablante. Por otra parte, el rendimiento se ve muy afectado cuando un modelo, entrenado en ciertas condiciones, se usa para reconocer una voz de hablante recogida por medio de diferentes micrófonos, canales y entornos. Todas estas condiciones de desajuste se denominan en general variabilidad intersesión.

[0009] Se han hecho varias propuestas para contrastar los efectos de la variabilidad intersesión tanto en los dominios de las características como de los modelos.

[0010] Una técnica popular usada para mejorar el rendimiento de un sistema de reconocimiento de hablante compensando las características acústicas es el Mapeo de Características, del cual puede encontrarse una descripción en el documento de D. Reynolds, Channel Robust Speaker Versification via Feature Mapping, en las Actas del ICASSP 2003, págs. II-53-6, 2003. En particular, el Mapeo de Características usa la información a priori de un conjunto de modelos dependientes del canal, entrenados en condiciones conocidas, para mapear los vectores de características hacia un espacio de características independientes del canal. Dada una expresión de entrada, primero se detecta el modelo dependiente del canal más probable y luego se mapea cada vector de características en la expresión al espacio independiente del canal basándose en el gaussiano seleccionado en el GMM dependiente del canal. El inconveniente de este enfoque es que requiere datos de aprendizaje etiquetados para crear los modelos dependientes del canal relacionados con las condiciones que se quiere compensar.

[0011] Por lo tanto, recientemente se han propuesto técnicas basadas en modelos que pueden compensar las variaciones del hablante y del canal sin requerir identificación explícita y etiquetado de diferentes condiciones. Estas técnicas comparten unos antecedentes comunes, concretamente la variabilidad de modelado de las expresiones del hablante que las restringen a un espacio propio de dimensiones bajas. Gracias a la dimensión reducida del espacio propio restringido, las técnicas basadas en modelos permiten una robusta compensación intersesión aun cuando se disponga únicamente de pocos datos dependientes del hablante.

[0012] En general, todas las técnicas de espacio propio basadas en modelos construyen supervectores a partir de modelos acústicos. Un supervector se obtiene agregando los parámetros de todos los gaussianos de un HMM/GMM en una sola lista. Típicamente, sólo se incluyen los parámetros gaussianos medios en los supervectores. Considerando, por ejemplo, un GMM de 512 gaussianos, que modela 13 MFCC + 13 características derivadas respecto al tiempo, se genera un supervector de 512x26=13312 características.

[0013] Después se realiza la compensación de hablante o de canal aplicando la siguiente ecuación:

$$\hat{\mu} = \mu + Ux \qquad (1)$$

donde μ y $\hat{\mu}$ son respectivamente supervectores sin compensar y compensado, Ux es un desplazamiento de compensación, U es una matriz de transformación de rango bajo desde el subespacio de variabilidad intersesión restringido hasta el subespacio de supervectores, y x es una representación de dimensiones bajas de la variabilidad intersesión en el subespacio de variabilidad intersesión restringido.

[0014] En los documentos US6.327.565, US6.141.644, US6.915.260 y el documento de S. Lucey y T. Chen, Improved Speaker Verification Through Probabilistic Subspace Adaptation, Actas del EUROSPEECH-2003, págs. 2021-2024, 2003, la matriz de subespacio *U* para compensación de hablante se construye recopilando un gran número de modelos dependientes del hablante de diferentes hablantes y aplicando una transformación lineal que reduce los supervectores de dimensiones altas en vectores base. El Análisis de Componentes Principales (PCA) se usa normalmente para construir la matriz de transformación U como una concatenación de los *K* vectores propios

que corresponden a los K valores propios más grandes. Los vectores propios seleccionados se conocen comúnmente como hablantes propios o voces propias porque cada modelo dependiente del hablante puede representarse casi como una combinación lineal de vectores base en el dominio de los supervectores.

[0015] Un enfoque similar para compensación de canal en reconocimiento de hablante se propone en el documento de P. Kenny, M. Mihoubi, y P. Dumouchel, New MAP Estimators for Speaker Recognition, Actas del EUROSPEECH-2003, págs. 2964-2967, 2003. En particular, está técnica, denominada en la publicación MAP de canales propios, construye el espacio propio restringido a partir de un gran número de supervectores que representan la variabilidad entre hablantes. Para estimar los canales propios, se necesitan varios modelos de hablantes, de una gran colección de hablantes y un conjunto de aprendizaje que comprende varias grabaciones de cada uno de estos hablantes.

[0016] En el documento de R. Vogt, B. Baker, S. Sridharan (2005): Modelling session variability in text-independent speaker verification, en las Actas de INTERSPEECH-2005, 3117-3120, la compensación de la variabilidad intersesión se realiza usando la ecuación previa. En este caso, la matriz de transformación U es entrenada por un algoritmo de maximización de expectativas (EM) para representar los tipos de variaciones entre hablantes esperadas entre sesiones. Con este fin, el subespacio es entrenado sobre una base de datos que contiene un gran número de hablantes, cada uno con varias sesiones grabadas independientemente. Por otra parte, se propone un procedimiento iterativo para estimar el supervector de hablante simple (μ en la ecuación). En la etapa de verificación cada modelo objetivo se compensa en una expresión de prueba dada i:

$$\hat{\mu}_i(s) = \mu(s) + Ux_i(s) \qquad (2)$$

[0017] La compensación se realiza estimando en primer lugar la representación de dimensiones bajas de la variabilidad intersesión en la grabación *i* en el hablante s, concretamente $x_i(s)$, y luego compensando el supervector del hablante a la grabación i, obteniendo el supervector compensado $\hat{\mu}_i(s)$. En particular, la compensación se realiza calculando el desplazamiento $\mathbf{U}x_i(s)$ en el espacio de supervectores como proyección del vector de variabilidad intersesión $\mathbf{x}_i(s)$ respecto al espacio de supervectores, a través de la matriz de transformación de rango bajo **U**, desde el subespacio de variabilidad intersesión restringido hasta el espacio de supervectores.

Objetivo y resumen de la invención

[0018] El solicitante ha observado que las técnicas basadas en modelos permiten mejor mejora de exactitud en la tarea de reconocimiento de hablante que las técnicas de compensación basadas en características como el Mapeo de Características. Sin embargo, el solicitante ha observado que las técnicas basadas en modelos anteriormente mencionadas operan sólo en el dominio del modelo acústico y, de este modo, están excesivamente ligadas a modelos acústicos y estructuras de reconocimiento específicos. Además, el solicitante también ha observado que como en las técnicas basadas en modelos anteriormente mencionadas la compensación se lleva a cabo modelo por modelo, en aquellas aplicaciones en las que ha de compensarse una gran cantidad de modelos, como las tareas de identificación del hablante, estas técnicas han demostrado ser computacionalmente costosas.

[0019] Por lo tanto, el objetivo de la presente invención es proporcionar una solución que permita que se reduzcan los efectos de la variabilidad del entorno, los micrófonos, los canales, etc., sobre el reconocimiento del hablante, y en particular que sea tan eficiente como las técnicas basadas en características en cuanto a costes computacionales y tan exacta como las técnicas basadas en modelos, y que permita que se desconecten los modelos de reconocimiento acústico y el conocimiento de compensación, permitiendo así que la presente invención sea aplicable a diferentes tareas y diferentes algoritmos de reconocimiento.

[0020] Este objeto se logra mediante la presente invención por el hecho de que se refiere a un procedimiento, un sistema y un producto de programa informático para compensar la variabilidad intersesión para extracción automática de información a partir de la voz, según las reivindicaciones adjuntas.

[0021] La presente invención logra el objeto anteriormente mencionado en dos fases distintas, durante las cuales se realiza el cálculo de factores intersesión y su compensación en el dominio de las características acústicas. En particular, la primera fase, que se realiza por anticipado y desconectada, consiste en la creación de una transformación que define el espacio vectorial restringido en el que ocurre la variabilidad intersesión, mientras que la

segunda etapa, que se repite para cada grabación de voz que ha de ser procesada, se aprovecha de la transformación obtenida en la primera fase para llevar a cabo la compensación de las características acústicas. Más detalladamente, durante la primera fase se construye desconectado un pequeño subespacio capaz de representar la variabilidad entre grabaciones del hablante diferentes en cuanto a factores intersesión basándose en una base de datos relacionada con muchos hablantes y que contiene, para cada hablante, un número significativo de grabaciones de voz adquiridas bajo diferentes condiciones. Después, se consideran las diferencias entre diferentes grabaciones de voz del mismo hablante, y se construye un subespacio de factores intersesión restringido basándose en estas diferencias, usando la técnica conocida de Análisis de Componentes Principales, en la que los factores intersesión representan la variabilidad intersesión entre diferentes grabaciones del mismo hablante, que no son significativas para el reconocimiento del propio hablante. Durante la fase conectada posterior, se estiman factores intersesión para cada grabación de voz desconocida. Los factores intersesión se restan luego de los vectores de características directamente en el dominio de las características acústicas. Las etapas de aprendizaje y reconocimiento de impresiones de voz tienen lugar después como normales, es decir, comenzando desde los vectores de características compensados.

[0022] La presente invención permite que las ventajas y la exactitud de las técnicas de espacio propio basadas en modelos sean transferidas en el dominio de las características acústicas. Por otra parte, compensar características en vez de modelos tiene la ventaja de que las características transformadas pueden usarse como vectores de características para clasificadores de diferente naturaleza y complejidad, y también para diferentes tareas como reconocimiento de idioma o habla.

Breve descripción de los dibujos

[0023] Para una mejor comprensión de la presente invención, a continuación se describirá una realización preferida, que está pensada meramente a modo de ejemplo y no ha de considerarse limitativa, con referencia a los dibujos adjuntos, en los que:

Figura 1. Muestra un diagrama de bloques de adquisición y procesamiento de voz;
Figura 2. Muestra un organigrama detallado de construcción de la matriz de subespacio de variabilidad intersesión;
Figura 3. Muestra un organigrama general de estimación del vector de factor intersesión;
Figura 4. Muestra un organigrama general de compensación de características acústicas;
Figura 5. Muestra un organigrama general de creación de impresión de voz del hablante;
Figura 6. Muestra un organigrama general de verificación del hablante; y
Figura 7. Muestra un organigrama general de identificación del hablante.

Descripción detallada de realizaciones preferidas de la invención

[0024] La siguiente discusión se presenta para permitir que una persona experta en la materia haga y use la invención. Diversas modificaciones de las realizaciones resultarán evidentes inmediatamente para los expertos en la materia, y los principios genéricos de este documento pueden aplicarse a otras realizaciones y aplicaciones sin apartarse del espíritu y alcance de la presente invención. Por lo tanto, la intención de la presente invención no es limitarse a las realizaciones mostradas, sino que es que esté de acuerdo con el más amplio alcance coherente con los principios y características desvelados en este documento y definidos en las reivindicaciones adjuntas.

[0025] Además, la presente invención se implementa por medio de un producto de programa informático que incluye porciones de código de software para implementar, cuando el producto de programa informático está cargado en una memoria del sistema de procesamiento y se ejecuta en el sistema de procesamiento, el procedimiento de compensación de la variabilidad intersesión descrito en lo sucesivo con referencia a las Figuras 2, 3 y 4.

[0026] La Figura 1 muestra un organigrama de adquisición y procesamiento de una señal de voz, generada por un hablante y captada por un transductor del micrófono, para obtener características acústicas que son necesarias durante ambas etapas de la presente invención. En particular, la voz del hablante es captada por un transductor de adquisición (bloque 10), que puede ser un micrófono de un auricular de teléfono fijo o móvil o un micrófono de un sistema de grabación, transductor de adquisición que produce como salida una señal de voz analógica (bloque 20), que luego se digitaliza y codifica, antes o después de la posible transmisión (bloque 30). La señal de voz digital así obtenida (bloque 40) normalmente se graba en un dispositivo de almacenamiento no volátil, como el sistema de almacenamiento secundario de un sistema informático (bloque 50), y es procesada por un terminal de entrada acústico (bloque 60), que produce como salida, en cuantías o intervalos de tiempo fijos, típicamente diez

milisegundos, un vector de características (bloque 70), que es una representación vectorial compacta de la voz. En una realización preferida, cada vector de características está formado por coeficientes cepstrales en las frecuencias de Mel (MFCCs). El orden del banco de filtros y de la transformada discreta del coseno (DCT) usados en la generación de los MFCCs puede ser 13. Además, cada vector de observación también puede incluir convenientemente la derivada de primer orden respecto al tiempo de cada MFCCs, para un total de 13+13=26 características para cada intervalo.

[0027] La Figura 2 muestra un organigrama de la primera etapa de la presente invención, concretamente la creación del subespacio de variabilidad intersesión.

[0028] El rendimiento de la primera etapa requiere la disponibilidad de una base de datos de voz (bloque 100) relacionada con un gran grupo de S hablantes, y que contenga, para cada hablante, un número R de grabaciones de voz adquiridas bajo diferentes condiciones, como para abarcar la variabilidad intersesión que se pretende compensar.

[0029] Basándose en esta base de datos de voz, un terminal de entrada acústico (bloque 110) extrae de cada muestra de voz digital vectores de características basándose en los cuales se crea un GMM para cada hablante y en cada una de las condiciones de adquisición disponibles usando una técnica de adaptación conocida comúnmente como técnica de adaptación de Máximo A Posteriori (MAP) (bloque 120) que es una técnica ampliamente usada para reconocimiento de hablante y que está basada en un modelo general del espacio acústico, denominado comúnmente Modelo de Fondo Universal (UBM) (bloque 130). El UBM es un GMM y constituye la "raíz" de la cual se derivan todos los modelos adaptados (bloque 140) usando la técnica de adaptación MAP. Los modelos adaptados, por lo tanto, mantienen las mismas características del UBM, en cuanto a los parámetros representados y la topología, y, en particular, conserva el mismo número de gaussianos y características acústicas. Considerando R grabaciones para S hablantes, se crearán R x S GMMs adaptados. El procedimiento puede generalizarse fácilmente al caso donde los hablantes tienen un número de grabaciones que son diferentes entre sí.

[0030] Para cada uno de los R x S GMMs adaptados, se crea un supervector correspondiente (bloques 150 y 160) colocando en orden los parámetros del GMM adaptado. En una realización preferida, sólo se concatenan los vectores formados por los valores medios de todos los gaussianos, denominados en lo sucesivo como vector medio, despreciando otros parámetros, como los coeficientes de ponderación y la covarianza. Suponiendo que el GMM está formado por G gaussianos y que cada vector medio tiene una dimensión F (la misma que los vectores de características, 26 en la realización considerada), un supervector estará compuesto de G x F parámetros. Considerando S hablantes y R grabaciones por hablante, se crean R x S supervectores, cada uno formado por G x F parámetros. En una realización preferida, los gaussianos se ordenan y examinan en orden ascendente, y los vectores medios correspondientes se concatenan luego para formar los supervectores. La ordenación de los gaussianos no es significativa, siempre que se mantenga constante para la generación de los supervectores. En una realización preferida, se usan 512 gaussianos multivariantes, cada una relacionada con un fenómeno acústico en el espacio de los 26 parámetros de los vectores de características: cada supervector está compuesto así de 512 x 26 = 13312 parámetros.

[0031] En una realización preferida, los supervectores relacionados con los GMMs del mismo hablante, adquiridos bajo diferentes condiciones, son examinados en pares, para resaltar el efecto, en el espacio acústico de los supervectores, del paso de las condiciones de la sesión del primer supervector del par a las condiciones de la sesión del segundo supervector del mismo hablante. En particular, esta operación se realiza calculando un supervector de diferencia para cada par como una diferencia vectorial entre los dos supervectores del par, y se repite para todos los pares disponibles para un hablante y para todos los hablantes (bloque 170). El número total de supervectores de diferencia que puede obtenerse con S hablantes y R grabaciones por hablante es $S \times R \times (R-1)/2$ (bloque 180).

[0032] Después de construir los supervectores se realiza una operación de reducción de dimensionalidad a través de una transformación lineal que reduce los supervectores originales de altas dimensiones en vectores base. En particular, en una realización preferida los supervectores de diferencia se procesan según una técnica de análisis conocida comúnmente como técnica de Análisis de Componentes Principales (PCA) (bloque 190). En particular, cada supervector de diferencia representa un punto en un espacio supervectorial con dimensiones G x F, y la técnica PCA determina un grupo de Vectores de Componentes Principales (bloque 200) que definen una base completa de vectores propios para el espacio supervectorial de manera que pueden generarse todos los supervectores de diferencia introducidos en el algoritmo PCA (puntos observados). Si los supervectores de diferencia son linealmente independientes entre sí, el número de vectores propios necesarios para reconstruirlos con precisión es igual al número de supervectores de diferencia introducidos. Si no es ese el caso, como sucede cuando

se introducen los supervectores de diferencia de los pares de supervectores, el número de vectores propios requeridos es inferior al número de supervectores de diferencia.

[0033] Otra propiedad importante de la técnica PCA aprovechada por el procedimiento propuesto es que los vectores propios se ordenan en cuanto a importancia decreciente, como una función del valor propio asociado con el vector propio, es decir, el primer vector propio más importante se asocia con el valor propio más alto, el segundo vector propio más importante se asocia con el segundo valor propio, etcétera. El término "importancia" es cuantificable en términos de cuánta parte de la varianza del espacio vectorial inicial se describe por un pequeño número de vectores propios elegidos de aquellos con los valores propios más altos (bloque 210). La técnica PCA garantiza que el incremento de varianza captado por los vectores propios disminuye con su orden, y por lo tanto es posible representar aproximadamente los puntos del espacio vectorial completo inicial en un espacio vectorial de dimensiones reducidas (bloque 220) descrito por un pequeño número de vectores propios, elegidos de aquellos con valores propios más altos, con la seguridad de representar los componentes principales.

[0034] En el procedimiento propuesto, los vectores propios obtenidos de los supervectores de diferencia con la técnica PCA permiten que la variabilidad introducida por las variaciones de la sesión sea descrita en el subespacio restringido de los vectores propios. Para representar los componentes principales afectados por la variabilidad intersesión, sólo los vectores propios con los *N* valores propios más altos, con $N < 100 << (G \times F)$ se consideran para construir la base del subespacio de variabilidad intersesión. Los *N* vectores propios elegidos se agrupan en columnas para formar la matriz de transformación U, con *N* columnas y (*G* x *F*) filas. La matriz de transformación U define el espacio restringido de variabilidad intersesión (bloques 230 y 240).

[0035] La segunda etapa de la presente invención prevé la aplicación de la ecuación de compensación (2) descrita previamente al UBM, suponiendo que la distorsión del espacio acústico en la grabación *i* y caracterizada por el vector $\mathbf{x_i}$ en el subespacio de variabilidad intersesión restringido se puede estimar comenzando desde el UBM. La ecuación de compensación (2) puede rescribirse eliminando la referencia al hablante (como los supervectores consideran el UBM) y haciendo explícito el índice m de cada gaussiano que forma un supervector:

$$\hat{\mu}_{i,m} = \mu_m + U_m x_i \qquad (3)$$

donde μ_m y $\hat{\mu}_{i,m}$ son respectivamente subvectores de los supervectores sin compensar y compensado y asociados con el gaussiano m-ésimo del UBM, $\mathbf{U_m}$ es una submatriz de F filas y *N* columnas de la matriz de transformación U y asociada con el gaussiano m-ésimo, y $\mathbf{x_i}$ es el vector de compensación para la grabación *i*, también denominado vector del factor intersesión, en subespacio restringido.

[0036] Para estimar los vectores del factor intersesión $\mathbf{x_i}$, la presente invención aprovecha una técnica denominada comúnmente Adaptación del Subespacio Probabilístico (PSA), para una descripción detallada del cual puede hacerse referencia a la publicación anteriormente mencionada *Improved Speaker Verification Through Probabilistic Subspace Adaptation.*

[0037] La Figura 3 muestra un organigrama general de una estimación del vector del factor intersesión. Una muestra de voz digital (bloque 300) es introducida en un front-end acústico (bloque 310) y los vectores de características producidos como salida por el terminal de entrada acústico son sometidos a la Adaptación del Subespacio Probabilístico (bloque 320), lo cual requiere el conocimiento de la matriz de transformación **U** (bloque 330) y del UBM (bloque 340), y proporciona el vector del factor intersesión correspondiente $\mathbf{x_i}$ (bloque 350).

[0038] Después se obtiene la compensación de las características acústicas proyectando los vectores del factor intersesión $\mathbf{x_i}$ del subespacio de variabilidad intersesión restringido de vuelta al espacio de modelo acústico ampliado. En particular, cada proyección $\mathbf{U}_m\mathbf{x}_i$ genera un vector de compensación de características con una dimensión igual a la de los vectores de características. Las contribuciones de compensación de características con respecto a los diversos gaussianos del UBM son ponderadas con la probabilidad de ocupación $\gamma_m(t)$ de los gaussianos, dado el vector de características. Se calcula una contribución de compensación para cada cuantía de tiempo t y se resta de cada vector de características original $O_i(t)$, que corresponde a la grabación *i*. Después se obtienen los vectores de características compensados $\hat{O}_i(t)$ mediante la siguiente ecuación:

$$\hat{O}_i(t) = O_i(t) - \sum_m \gamma_m(t) U_m x_i \qquad (4)$$

donde $\sum_m \gamma_m(t) U_m x_i$ representa un vector de características de compensación de variabilidad intersesión que ha de ser restado de cada vector de características original $O_i(t)$ para obtener los vectores de características compensados $\hat{O}_i(t)$.

[0039] En la experiencia práctica, la compensación puede llevarse a cabo incluyendo simplemente un número limitado de términos en la suma, en particular aquellos asociados con los gaussianos que presentan la probabilidad de ocupación más alta en cada tiempo t.

[0040] La Figura 4 muestra el organigrama de compensación de características acústicas. Una muestra de voz digital en una grabación *i* (bloque 400) se introduce en un terminal de entrada acústico (bloque 410) y los vectores de características producidos como salida por el terminal de entrada acústico se usan para calcular los vectores de compensación de características (bloque 420), cálculo que requiere el conocimiento de las proyecciones $\mathbf{U_x}$ (bloque 430) y del UBM (bloque 440). Los vectores de compensación de características se restan luego de los vectores de características producidos como salida por el terminal de entrada acústico (bloque 450), obteniendo así los vectores de características compensados correspondientes (bloque 460).

[0041] En el caso del reconocimiento de hablante, la presente invención se aplica tanto durante la creación de la impresión de voz del hablante como la verificación/identificación del hablante. Sin embargo, se puede lograr buenos resultados de reconocimiento aplicando la presente invención simplemente a la verificación del hablante, sin normalizar los vectores de características durante el aprendizaje.

[0042] Las ventajas de la presente invención son evidentes a partir de lo anterior.

[0043] Además, se pone de relieve que como la presente invención opera en el dominio de las características acústicas, puede usarse en contextos y aplicaciones distintas de las descritas previamente.

[0044] En el campo del reconocimiento de hablante mediante GMM, es posible diferenciar el UBM usado para la compensación de características acústicas del que se usa para la modelización de los hablantes. Por ejemplo, podría usarse un UBM con un pequeño número de gaussianos (por ejemplo, 512) para compensación mediante factores intersesión y modelos más detallados para modelizar los hablantes (por ejemplo, 2048 gaussianos).

[0045] Siempre dentro del contexto del reconocimiento de hablante, es posible usar el procedimiento descrito para adaptar los parámetros introducidos a otros tipos de clasificadores, como modelos HMM o Máquinas de Vectores de Soporte (SVM).

[0046] El procedimiento descrito también puede encontrar aplicación en el contexto del reconocimiento de idioma, donde la compensación de la variabilidad intersesión es tan importante como en el caso del reconocimiento de hablante. También en este caso, el procedimiento puede usarse en el procesamiento previo para eliminar la variabilidad intersesión de los vectores de características usados para reconocimiento de idioma.

[0047] Como ejemplo, la Figura 5 muestra un organigrama básico de creación de impresión de voz del hablante, donde una muestra de voz digital (bloque 500) se introduce en un terminal de entrada acústico (bloque 510), y los vectores de características producidos como salida por el terminal de entrada acústico se usan para compensar la variabilidad intersesión (bloque 520) basándose en la matriz de transformación U (bloque 530) y un primer UBM (por ejemplo, con 512 gaussianos) (bloque 540), como se describió previamente. Los vectores de características compensados se usan luego para la creación de impresión de voz del hablante (bloque 550) basándose en un segundo UBM (por ejemplo, con 2048 gaussianos) (bloque 560), obteniendo así la impresión de voz del hablante (bloque 570). En una realización diferente, el primer y el segundo UBMs pueden ser el mismo.

[0048] Como ejemplo adicional, la Figura 6 muestra un organigrama básico de una verificación del hablante, donde una muestra de voz digital (bloque 600) se introduce en un terminal de entrada acústico (bloque 610), y los vectores de características producidos como salida por el terminal de entrada acústico se usan para compensar la variabilidad intersesión (bloque 620) basándose en la matriz de transformación U (bloque 630) y un primer UBM (por ejemplo, con 512 gaussianos) (bloque 640), como se describió previamente. Los vectores de características compensados se usan luego para la verificación del hablante (bloque 650) basándose en la impresión de voz del hablante (bloque 660) y un segundo UBM (por ejemplo, con 2048 gaussianos) (bloque 670), obteniendo así una puntuación de probabilidad (bloque 680). En una realización diferente, el primer y el segundo UBMs pueden ser el

mismo.

[0049] Por último, como un ejemplo adicional más, la Figura 7 muestra un organigrama básico de una identificación del hablante, donde una muestra de voz digital (bloque 700) se introduce en un terminal de entrada acústico (bloque 710), y los vectores de características producidos como salida por el terminal de entrada acústico se usan para compensar la variabilidad intersesión (bloque 720) basándose en la matriz de transformación U (bloque 730) y un primer UBM (por ejemplo, con 512 gaussianos) (bloque 740), como se describió previamente. Los vectores de características compensados se usan luego para la identificación del hablante (bloque 750) basándose en impresiones de voz del hablante (bloques 760) y un segundo UBM (por ejemplo, con 2048 gaussianos) (bloque 770), obteniendo así un resultado de identificación (bloque 780). En una realización diferente, el primer y el segundo UBMs pueden ser el mismo.

[0050] Por último, está claro que pueden realizarse numerosas modificaciones y variantes en la presente invención, entrando todas dentro del alcance de la invención, tal como se define en las reivindicaciones adjuntas.

[0051] En particular, como el procedimiento propuesto realiza la compensación al nivel de las características acústicas, también puede usarse en contextos y aplicaciones distintas de las descritas previamente, como reconocimiento de idioma y habla, donde la compensación de la variabilidad del canal es tan importante como en el caso del reconocimiento de hablante. También en estas aplicaciones, la presente invención puede usarse como procesamiento previo para eliminar la variabilidad del canal de los vectores de observación usados para reconocimiento de idioma y habla.

[0052] Además, siempre dentro del contexto del reconocimiento de hablante, es posible usar el procedimiento descrito para adaptar los parámetros que alimentan a otros tipos de clasificadores, como modelos HMM o Máquinas de Vectores de Soporte (SVM).

[0053] Adicionalmente, la variabilidad intersesión puede compensarse basándose en un UBM diferente de un GMM, por ejemplo un HMM. En este caso, cada supervector se forma concatenando vectores medios de todos los gaussianos en todos los estados del HMM.

[0054] Por otra parte, la matriz de transformación U puede calcularse basándose en una técnica de análisis diferente del PCA, por ejemplo PCA de Maximización de Expectativas (EMPCA), Análisis de Componentes Independientes (ICA), Análisis Discriminante Lineal (LDA), Análisis de Factores (FA), y Descomposición de Valores Singulares (SVD), así como el vector de factor intersesión x_i puede calcularse basándose en una técnica de adaptación diferente del PSA, por ejemplo Descomposición Propia de Probabilidad Máxima (MLED).

[0055] Por último, el procedimiento de compensación puede aplicarse iterativamente sobre porciones de una grabación entera, a través de la repetición de los algoritmos descritos para cada porción de la propia grabación. En este caso cada porción de la grabación entera p tendrá un vector de factor intersesión asociado $\mathbf{x}_{ip}$, dicho vector de factor intersesión $\mathbf{x}_{ip}$ ha de considerarse para compensar los vectores de características con respecto a la porción de grabación relacionada.

REIVINDICACIONES

1. Un procedimiento para compensar la variabilidad intersesión para extracción automática de información de una señal de voz de entrada que representa una expresión de un hablante, que comprende:

- procesar la señal de voz de entrada para proporcionar vectores de características formados cada uno por características acústicas extraídas de la señal de voz de entrada en un intervalo de tiempo;
- calcular un vector de características de compensación de la variabilidad intersesión; y
- calcular vectores de características compensados restando el vector de características de compensación de la variabilidad intersesión de los vectores de características extraídos; y **caracterizado porque** el cálculo de un vector de características de compensación de la variabilidad intersesión incluye:
- crear un Modelo de Fondo Universal (UBM) basándose en una base de datos de voz de aprendizaje, el Modelo de Fondo Universal (UBM) incluyendo varios gaussianos y modelizando un espacio acústico que define un espacio de modelo acústico;
- crear una base de datos de grabaciones de voz relacionada con diferentes hablantes, y que contiene, para cada hablante, varias grabaciones de voz adquiridas bajo diferentes condiciones;
- calcular una matriz de subespacio de variabilidad intersesión (U) analizando, para cada hablante, las diferencias entre grabaciones de voz adquiridas bajo diferentes condiciones y contenidas en la base de datos de grabaciones de voz, definiendo la matriz de subespacio de variabilidad intersesión (U) una transformación de un espacio de modelo acústico a un subespacio de variabilidad intersesión que representa la variabilidad intersesión para todos los hablantes;
- calcular una vector de factor intersesión independiente del hablante único basándose en la señal de voz de entrada, realizando una técnica de estimación sobre los vectores de características, extraídos de la señal de voz de entrada, basándose en la matriz de subespacio de variabilidad intersesión (U) y el Modelo de Fondo Universal (UBM), representando el vector de factor intersesión la variabilidad intersesión de la señal de voz de entrada en el subespacio de variabilidad intersesión; y
- calcular el vector de características de compensación de la variabilidad intersesión basándose en la matriz de subespacio de variabilidad intersesión (U), el vector de factor intersesión y el Modelo de Fondo Universal (UBM).

2. El procedimiento de la reivindicación 1, en el que el cálculo del vector de características de compensación de la variabilidad intersesión basándose en la matriz de subespacio de variabilidad intersesión (U), el vector de factor intersesión y el Modelo de Fondo Universal (UBM) incluye:

- calcular contribuciones de compensación de la variabilidad intersesión, una para cada uno de los gaussianos del Modelo de Fondo Universal (UBM) basándose en la matriz de subespacio de variabilidad intersesión (U) y el vector de factor intersesión;
- ponderar las contribuciones de compensación de la variabilidad intersesión con la probabilidad de ocupación de gaussianos respectivos, dado un vector de características.

3. El procedimiento de la reivindicación 2, en el que el cálculo de contribuciones de compensación de la variabilidad intersesión incluye:

- multiplicar el vector de factor intersesión por una submatriz (U_m) de la matriz de subespacio de variabilidad intersesión relacionada con un gaussiano correspondiente del Modelo de Fondo Universal (UBM).

4. El procedimiento de la reivindicación 2 ó 3, en el que cada vector de características compensado se calcula basándose en la siguiente fórmula:

$$\hat{O}_i(t) = O_i(t) - \sum_m \gamma_m(t) U_m x_i$$

donde $\hat{O}_i(t)$ es el vector de características compensado, $O_i(t)$ es el vector de características extraído, x_i es el vector de factor intersesión, i identifica la señal de speech de entrada, m identifica el gaussiano del Modelo de Fondo Universal, **Um** es la submatriz de la matriz de subespacio de variabilidad intersesión U y relacionada con el gaussiano m-ésimo, y $\gamma_m(t)$ es la probabilidad de ocupación del gaussiano m-ésimo en el intervalo de tiempo t.

5. El procedimiento de la reivindicación 1, en el que la técnica de estimación es una Adaptación de Subespacio Probabilístico o una Adaptación de Descomposición Propia de Probabilidad Máxima.

6. El procedimiento de cualquiera de las reivindicaciones precedentes, en el que la determinación de una matriz de subespacio de variabilidad intersesión (U) incluye:

- calcular un modelo gaussiano para cada hablante y para cada grabación de voz en la base de datos de voz, incluyendo cada modelo gaussiano varios gaussianos;
- calcular un supervector (SV) para cada modelo gaussiano; y
- calcular la matriz de subespacio de variabilidad intersesión (U) basándose en los supervectores (SV).

7. El procedimiento de la reivindicación 6, en el que el cálculo de un modelo gaussiano incluye:

- realizar una etapa de adaptación basándose en los vectores de características y el Modelo de Fondo Universal (UBM).

8. El procedimiento de la reivindicación 7, en el que dicha realización de una etapa de adaptación incluye:

- realizar una adaptación de Máximo A Posteriori (MAP) basándose en los vectores de características y el Modelo de Fondo Universal (UBM).

9. El procedimiento de cualquiera de las reivindicaciones 6 a 8, en el que el cálculo de un supervector (SV) incluye:

- formar vectores medios con valores medios de todos los gaussianos del modelo gaussiano; y
- concatenar los vectores medios.

10. El procedimiento de la reivindicación 9, en el que la formación de vectores medios incluye:

- numerar los gaussianos del modelo gaussiano; y
- considerar los gaussianos en orden ascendente.

11. El procedimiento de cualquiera de las reivindicaciones 6 a 10, en el que el cálculo de la matriz de subespacio de variabilidad intersesión (U) basándose en los supervectores (SV) incluye:

- para cada hablante, calcular un supervector de diferencia para cada par de supervectores relacionados con los modelos gaussianos del hablante como una diferencia vectorial entre los dos supervectores del par; y
- realizar una reducción de dimensionalidad sobre los supervectores de diferencia para generar un grupo de vectores propios que definen el espacio de supervectores; y
- calcular la matriz de subespacio de variabilidad intersesión (U) basándose en los vectores propios.

12. El procedimiento de la reivindicación 11, en el que la realización de la reducción de dimensionalidad incluye:

- elegir vectores propios específicos según un criterio dado; y
- calcular la matriz de subespacio de variabilidad intersesión (U) basándose en los vectores propios elegidos.

13. El procedimiento de la reivindicación 12, en el que el cálculo de la matriz de subespacio de variabilidad intersesión (U) basándose en los vectores propios elegidos incluye:

- agrupar los vectores propios elegidos en columnas para formar la matriz de subespacio de variabilidad intersesión (U),

14. El procedimiento de la reivindicación 12 ó 13, en el que cada vector propio está asociado con un valor propio respectivo, y en el que la elección de vectores propios específicos según un criterio dado incluye:

- elegir los vectores propios con los valores propios más altos.

15. Un sistema para extraer automáticamente información de una señal de voz de entrada que representa una expresión de un hablante, que comprende un sistema de compensación de la variabilidad intersesión configurado para implementar el procedimiento de compensación de la variabilidad intersesión según cualquiera de las reivindicaciones 1 a 14 precedentes.

16. Un producto de programa informático que se puede cargar en una memoria de un sistema de procesamiento y que comprende porciones de código de software para implementar, cundo el producto de programa informático se ejecuta en el sistema de procesamiento, el procedimiento para compensar la variabilidad intersesión según cualquiera de las reivindicaciones 1 a 14 precedentes.

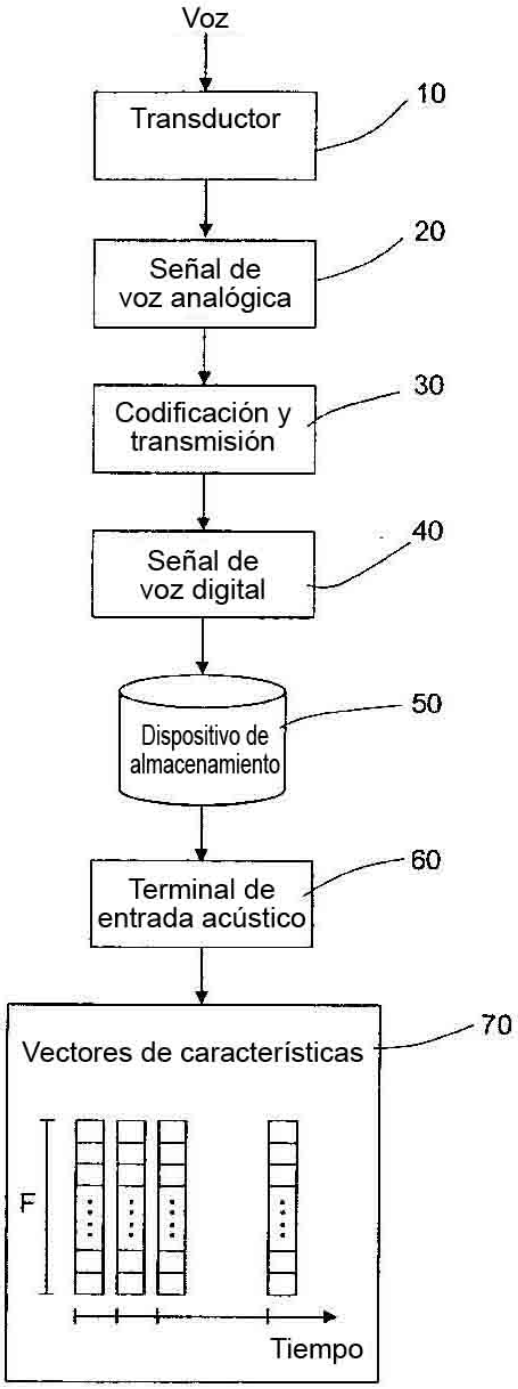
Voz
Transductor
10
Señal de voz analógica
20
Codificación y transmisión
30
Señal de voz digital
40
Dispositivo de almacenamiento
50
Terminal de entrada acústico
60
Vectores de características
70
F
Tiempo

Fig. 1

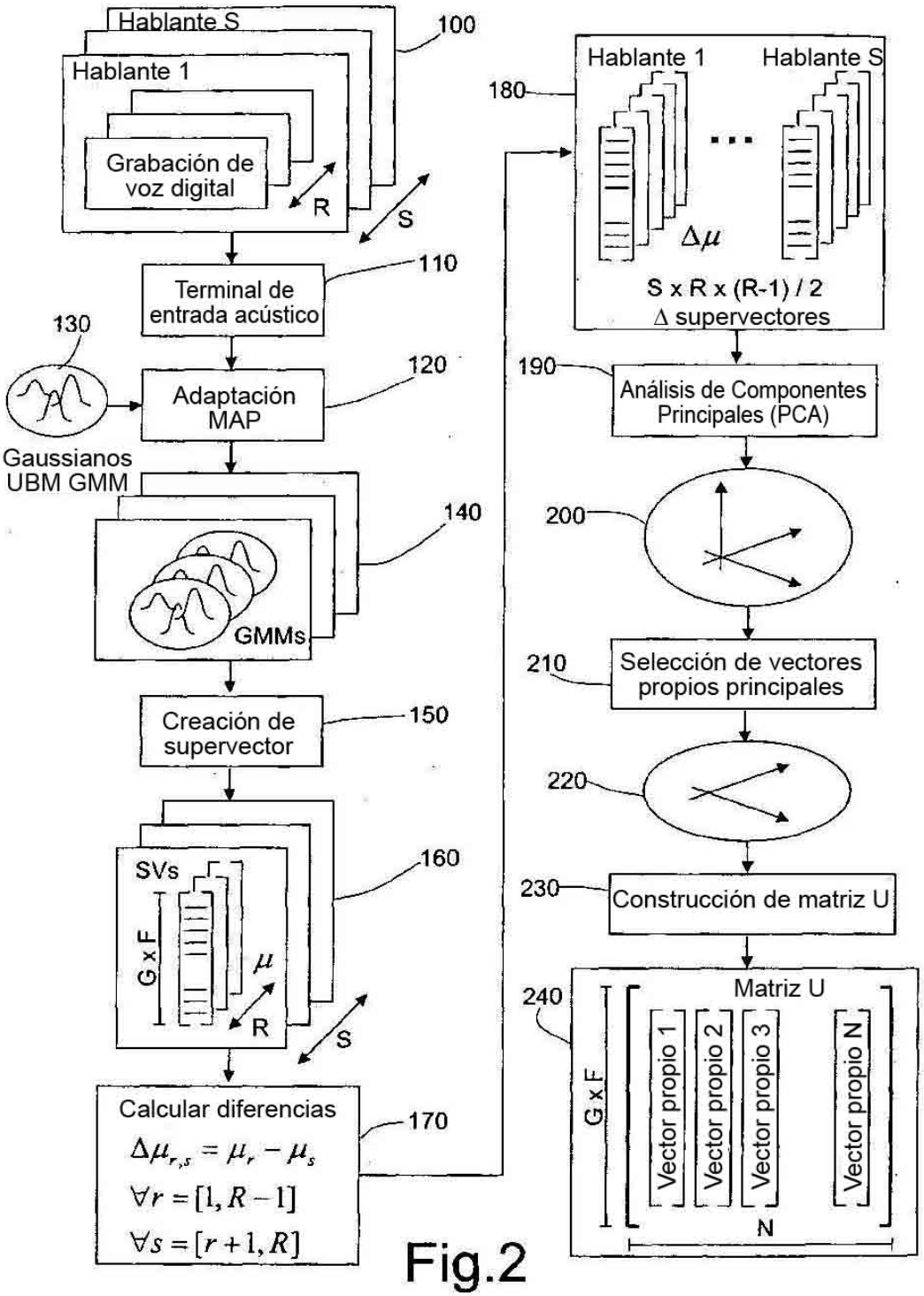
Hablante S
Hablante 1
100
Grabación de voz digital
R
S
Terminal de entrada acústico
110
130
120
Adaptación MAP
Gaussianos UBM GMM
140
GMMs
Creación de supervector
150
SVs
160
G x F
μ
R
S
Calcular diferencias
$\Delta\mu_{r,s} = \mu_r - \mu_s$
$\forall r = [1, R-1]$
$\forall s = [r+1, R]$
170
180
Hablante 1
Hablante S
Δμ
S x R x (R-1) / 2
Δ supervectores
190
Análisis de Componentes Principales (PCA)
200
210
Selección de vectores propios principales
220
230
Construcción de matriz U
240
Matriz U
G x F
Vector propio 1
Vector propio 2
Vector propio 3
Vector propio N
N

Fig.2

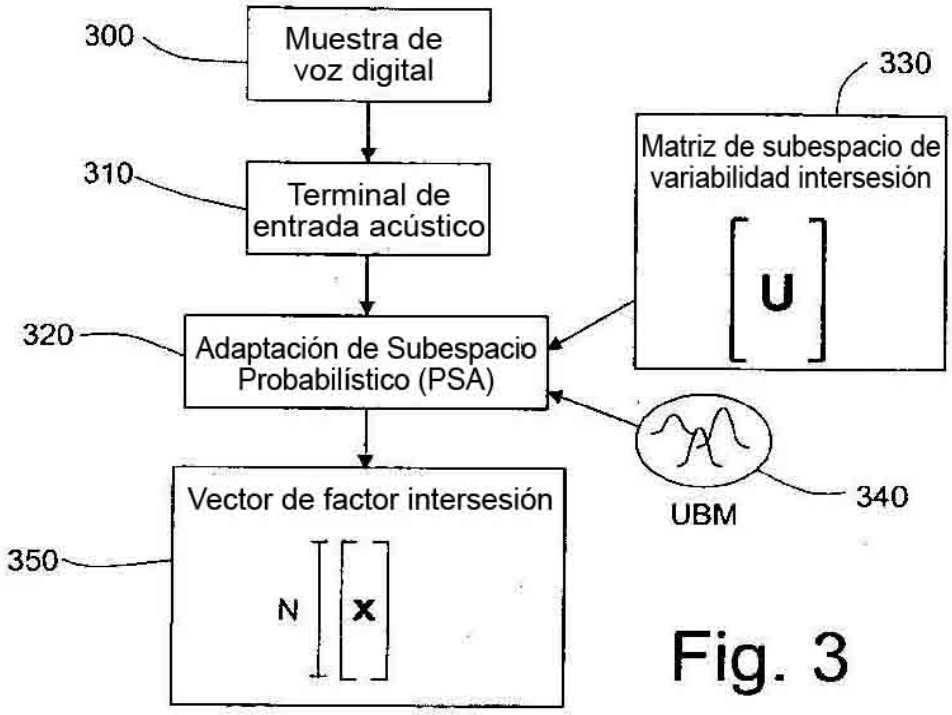
300
Muestra de voz digital
330
Matriz de subespacio de variabilidad intersesión
U
310
Terminal de entrada acústico
320
Adaptación de Subespacio Probabilístico (PSA)
UBM
340
Vector de factor intersesión
350
N
x

Fig. 3

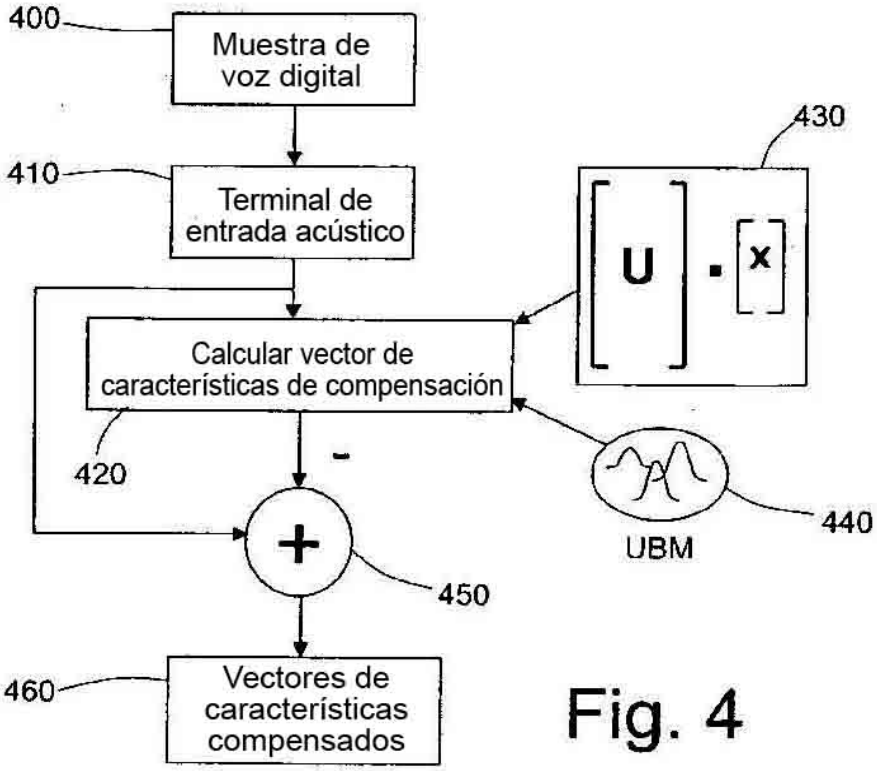
400
Muestra de voz digital
430
410
Terminal de entrada acústico
U
x
Calcular vector de características de compensación
420
-
+
UBM
440
450
460
Vectores de características compensados

Fig. 4

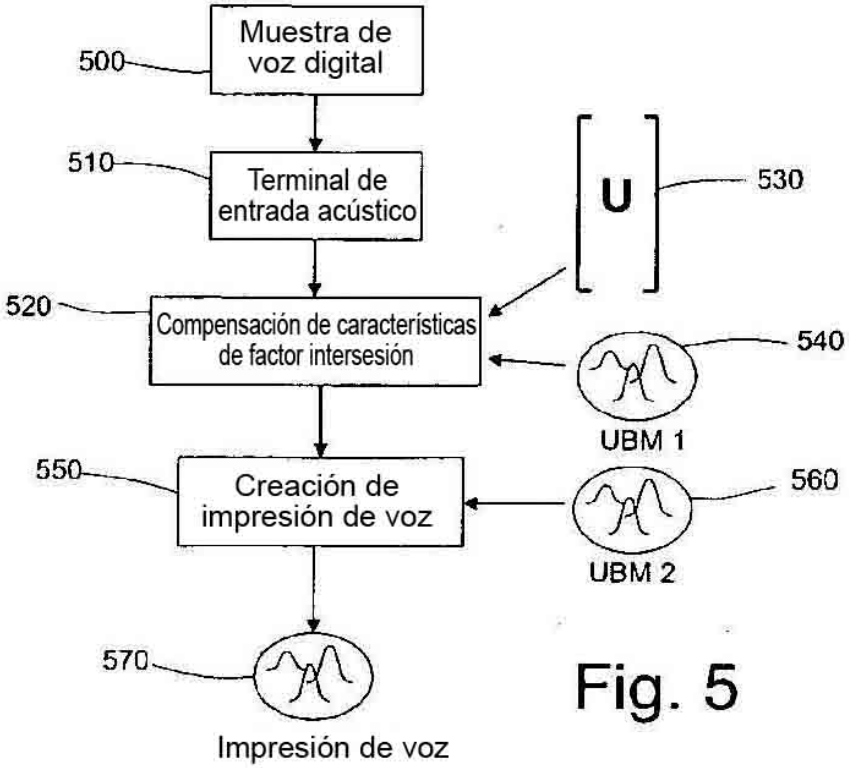
Muestra de voz digital
500
Terminal de entrada acústico
510
U
530
Compensación de características de factor intersesión
520
540
UBM 1
Creación de impresión de voz
550
560
UBM 2
570
Impresión de voz

Fig. 5

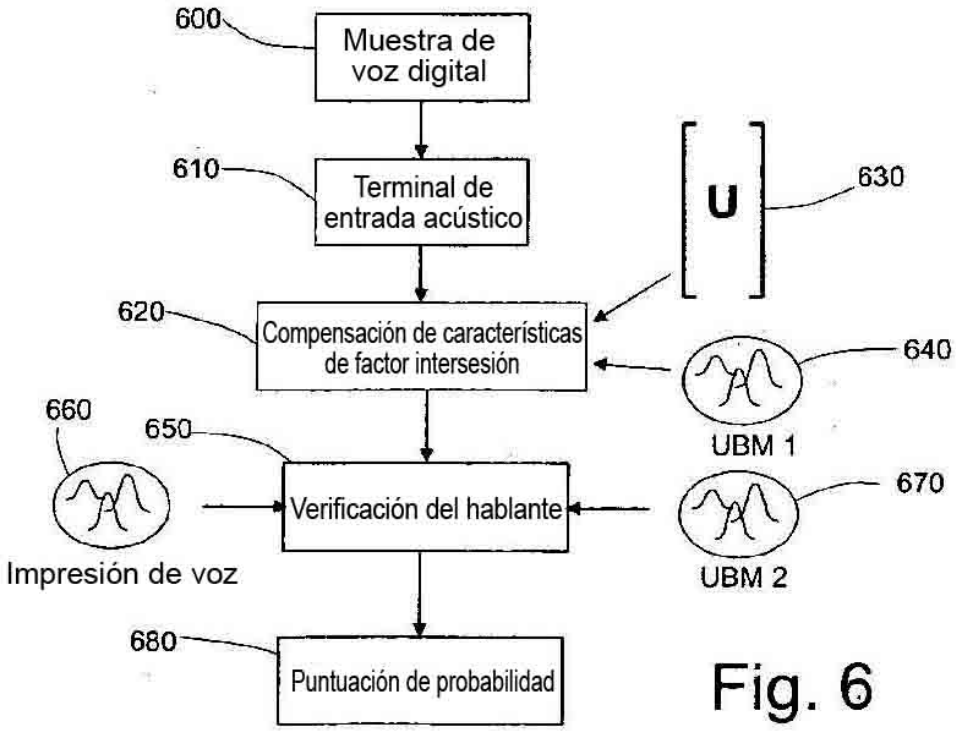
Muestra de voz digital
600
Terminal de entrada acústico
610
U
630
Compensación de características de factor intersesión
620
640
UBM 1
660
650
Verificación del hablante
670
Impresión de voz
UBM 2
680
Puntuación de probabilidad

Fig. 6

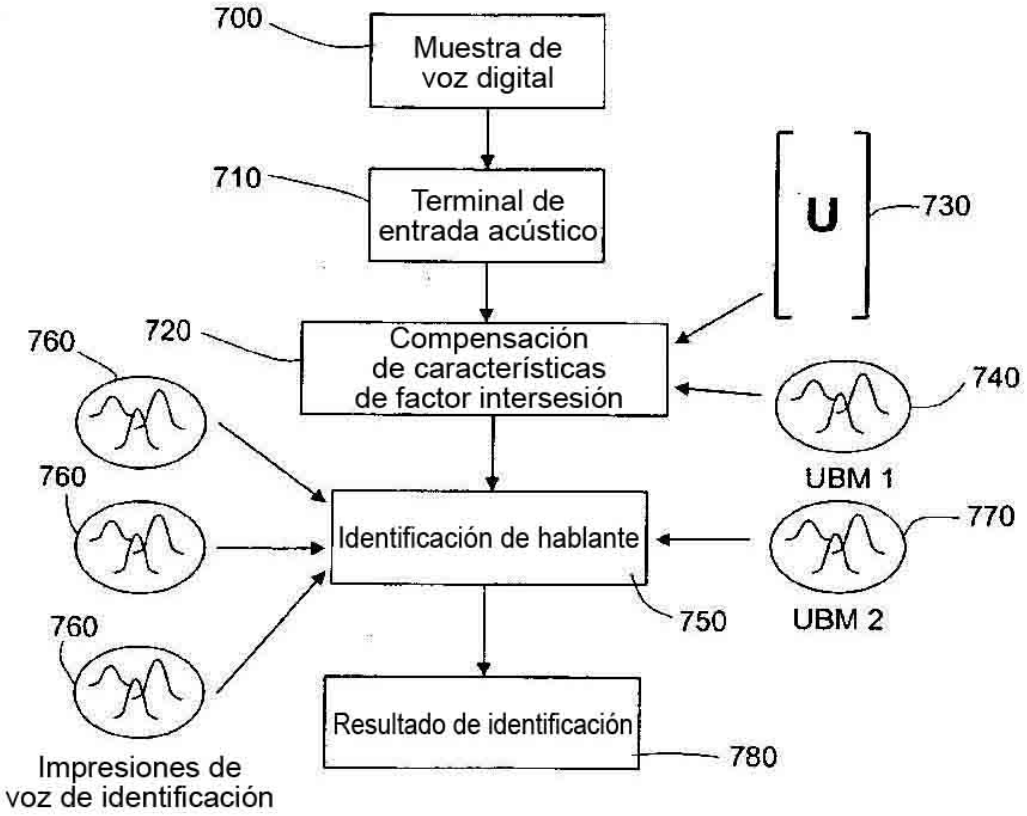
700
Muestra de voz digital
710
Terminal de entrada acústico
U
730
760
720
Compensación de características de factor intersesión
740
UBM 1
760
Identificación de hablante
770
UBM 2
750
760
Resultado de identificación
780
Impresiones de voz de identificación

Fig. 7