(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 440 646**

(51) Int. CI.:

***G10L 17/02*** (2013.01)
***G10L 17/24*** (2013.01)
***G10L 17/00*** (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **10.12.2009 E 09771309 (3)**

(97) Fecha y número de publicación de la concesión europea: **27.11.2013 EP 2364496**

(54) Título: **Detección de falsificación por cortar y pegar por alineamiento temporal dinámico**

(30) Prioridad:

**10.12.2008 WO PCT/EP2008/010478**
**26.06.2009 WO PCT/EP2009/004649**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
**29.01.2014**

(73) Titular/es:

**AGNITIO S.L. (100.0%)**
**C/ Gran Vía 39 - 8ª planta**
**28013 Madrid, ES**

(72) Inventor/es:

**VILLALBA LÓPEZ, JESÚS ANTONIO;**
**ORTEGA GIMÉNEZ, ALFONSO;**
**LLEIDA SOLANO, EDUARDO;**
**VARELA REDONDO, SARA y**
**GARCÍA GOMAR, MARTA**

(74) Agente/Representante:

**MILTENYI, Peter**

## DESCRIPCIÓN

Detección de falsificación por cortar y pegar por alineamiento temporal dinámico.

La descripción de las solicitudes PCT del mismo titular, con las solicitudes Nos. PCT/EP 2008/010478 presentada el 10 de diciembre de 2008, y PCT/EP2009/004649 presentada el 26 de junio de 2009 en la Oficina de Patentes Europea, se incorpora totalmente aquí por referencia. Además, se reivindican las prioridades de estas solicitudes de acuerdo con el Convenio de París.

La presente invención se refiere a un procedimiento, un medio informático y un aparato para comparar expresiones de voz.

La comparación de expresiones de voz puede utilizarse para reconocer a una persona que habla. Por ejemplo, la persona que habla realiza una expresión de voz de una muestra de texto determinada la cual se compara después con una expresión de voz previamente grabada de la misma persona que habla. En el caso de que las dos expresiones de voz coincidan razonablemente bien, la persona que habla es identificada con éxito. Tal identificación de una persona que habla puede utilizarse para validar una persona que desea obtener algún tipo de acceso o que tiene que demostrar la presencia en un lugar determinado, por ejemplo.

WO 98/34216 A2 describe un sistema y un procedimiento para detectar una voz grabada que puede utilizarse independientemente o para proporcionar protección ante un uso fraudulento de una grabación para burlar un sistema de reconocimiento de voz automático. Se han empleado diversas técnicas y sistemas independientemente o bien en combinación para verificar que una muestra de audio detectada es en directo y no está grabada. Las características de voz temporales de una muestra de audio se analizan para determinar si una muestra bajo examen es similar a una muestra previa para indicar que se trata de una grabación. Se examinan características del canal de comunicaciones para determinar si una muestra fue grabada en un canal distinto de un canal de comunicaciones predeterminado. Un clasificador de patrones se entrena para distinguir entre una voz en directo y una grabada. Finalmente, se utiliza una "marca de agua de audio" para determinar si una muestra de audio detectada es una grabación de una comunicación previa por un usuario autorizado.

Cuando se realiza una comparación de expresiones de voz pueden aparecer una serie de problemas. En primer lugar, incluso si para dos expresiones de voz que se utilizan en una comparación, la persona que habla así como la muestra de texto hablado son iguales, típicamente no se produce una coincidencia perfecta entre las dos expresiones de voz dado que la persona que habla puede pronunciar algunas palabras de manera algo distintas o la persona que habla podría pronunciar un texto determinado a una velocidad distinta, por ejemplo. En segundo lugar, la comparación debe ser capaz de detectar todo tipo de falsificación, tal como una falsificación por cortar y pegar. La falsificación por cortar y pegar la puede realizar una persona no autorizada cuando la persona no autorizada ha tenido acceso a texto grabado de la persona que habla y produce la muestra de texto cortando y pegando secuencias de estas expresiones de voz grabadas de esa persona que habla con el fin de producir expresiones de voz falsas de esa muestra de texto que tendría el sonido de la voz de esa persona que habla, en este ejemplo.

Por lo tanto, un problema a resolver por la presente invención es mejorar la comparación de expresiones de voz de manera, que por una parte, una persona que habla pueda ser identificada con gran eficacia y, por otra parte pueda detectarse con fiabilidad una falsificación, tal como una falsificación por cortar y pegar.

De acuerdo con la invención, el problema mencionado anteriormente se resuelve mediante el procedimiento de la reivindicación 1, el medio informático de la reivindicación 6 y el aparato de la reivindicación 7.

En las reivindicaciones dependientes se especifican realizaciones adicionales de la presente invención.

Un procedimiento para comparar expresiones de voz comprende las siguientes etapas:

En primer lugar, se extrae una pluralidad de rasgos de una primera expresión de voz de una muestra de texto determinada, y se extrae una pluralidad de rasgos de una segunda expresión de voz de dicha muestra de texto determinada. Todos los rasgos se extraen en función del tiempo y cada rasgo de la segunda expresión de voz tiene un rasgo correspondiente de la primera expresión de voz con el fin de poder utilizar el rasgo correspondiente para la comparación mencionada anteriormente.

En segundo lugar, se aplica alineamiento temporal dinámico a una o más características que dependen del tiempo de la primera y/o la segunda expresión de voz. Esto puede realizarse por ejemplo, minimizando una o más medidas de distancia o maximizando una medida de similitud. El alineamiento temporal dinámico se describe, por ejemplo, en la solicitud PCT del mismo titular con el número de solicitud mencionado anteriormente PCT/EP 2009/004649. Una medida de la distancia es una medida de la diferencia de una característica que depende del tiempo de la primera

expresión de voz y una característica que depende del tiempo correspondiente de la segunda expresión de voz. Una característica que depende del tiempo de una expresión de voz corresponde a una combinación de dos o más rasgos de diferentes tipos de rasgos. Aplicar alineamiento temporal dinámico a una característica que depende del tiempo de la primera o la segunda expresión de voz puede tener el efecto de que dicha característica se extienda o se comprima en determinadas zonas a lo largo del eje de tiempo. Debido a esta variación o flexibilidad, respectivamente, aplicar alineamiento temporal dinámico puede hacer que una característica que depende del tiempo de la primera o la segunda expresión de voz sea más similar a la característica que depende del tiempo que representa rasgos de la segunda o la primera expresión de voz, respectivamente. Por ejemplo, tratando conjuntamente dos o más rasgos en el proceso de alineamiento temporal dinámico, es decir, aplicando el mismo alineamiento temporal dinámico a los dos o más rasgos al mismo tiempo los rasgos pueden combinarse en una característica. Aquí, para el alineamiento temporal dinámico se utiliza una función de la distancia que tiene en cuenta dos o más rasgos al mismo tiempo. El alineamiento temporal dinámico puede realizarse de este modo sobre una combinación de rasgos.

En tercer lugar, se calcula una medida de la distancia total en la que la medida de la distancia total es una medida de la diferencia entre una primera expresión de voz de la muestra de texto determinada y la segunda expresión de voz de dicha muestra de texto determinada. La medida de la distancia total se calcula en base a uno o más pares de las características que dependen del tiempo mencionadas anteriormente, donde un par de características que dependen del tiempo está compuesta por una característica que depende del tiempo de la primera o la segunda expresión de voz y de una característica que depende del tiempo por alineamiento temporal dinámico de la segunda o primera expresión de voz, respectivamente, o donde un par de características que dependen del tiempo está compuesta por una característica que depende del tiempo (202) por alineamiento temporal dinámico de la primera expresión de voz y de una característica que depende del tiempo (202) por alineamiento temporal dinámico de la segunda expresión de voz. En otras palabras, las características que dependen del tiempo de un par se comparan entre sí y estas comparaciones, en las que el número de comparaciones es igual que el número de pares, se reflejan en el cálculo de la medida de la distancia total.

Con el esquema anterior se encontró, en particular, que la falsificación por cortar y pegar puede identificarse y separarse claramente de otras expresiones de voz (por ejemplo, normales generadas por humanos). Los cambios temporales bruscos de valores de rasgos en expresiones de voz generadas por cortar y pegar producen diferencias bien reconocibles en las distancias indicadas anteriormente pero al mismo tiempo dan una buena tasa de aceptación para expresiones de voz no generadas por una falsificación por cortar y pegar.

Utilizar una pluralidad de rasgos en la comparación de expresiones de voz resulta útil, en particular, en situaciones en las que un rasgo no varía significativamente con el tiempo en un determinado intervalo de tiempo, pero otro rasgo varía significativamente con el tiempo en dicho intervalo de tiempo. En el caso de que se tengan en cuenta varios rasgos para la comparación de las expresiones de voz, puede garantizarse mejor que existe una variación significativa en el tiempo en toda la longitud de una expresión de voz lo cual puede ser útil cuando se aplica alineamiento temporal dinámico que funciona mejor para una variación significativa de rasgos simples o combinados con el tiempo.

Además, tener en cuenta una serie de rasgos también puede ser útil cuando se calcula la medida de la distancia total ya que una medida de la distancia total que se calcula en base a varios rasgos puede permitir una comparación de dos expresiones de voz de manera que, en casos en los que la misma persona que habla da correctamente ambas expresiones de voz, se separan mejor de casos en los que la segunda expresión de voz es el resultado de una falsificación por cortar y pegar. De nuevo, una variación de características continua en función del tiempo puede ser útil para detectar similitudes o bien diferencias que podrían resultar de una falsificación por cortar y pegar (donde podrían esperarse cambios abruptos en algunos de los rasgos).

Además, tener en cuenta una pluralidad de rasgos para la comparación de expresiones de voz permite realizar la comparación en forma de varias sub-comparaciones lo que, de nuevo, puede aumentar la fiabilidad de la comparación dado que pueden detectarse similitudes y diferencias en un determinado intervalo de tiempo solamente en el caso de algunos de los rasgos pero no necesariamente en el caso de cualquier rasgo individual.

La comparación de expresiones de voz puede comprender solicitar y recibir la segunda expresión de voz de una persona que habla y comparar la segunda expresión de voz con una primera expresión de voz que ha sido grabada previamente. Además, la medida de la distancia total se emplea con el fin de validar la persona que habla de la segunda expresión de voz o para detectar que la segunda expresión de voz es el resultado de una falsificación.

La pluralidad de rasgos puede comprender uno o más de los siguientes rasgos:

el tono una función del tono tal como el *logPitch* donde *logPitch* es el logaritmo del tono,

el primer formante o una función del primer formante tal como *logF1* donde *logF1* es el logaritmo del primer formante,

el segundo formante o una función del segundo formante tal como *logF2* donde *logF*2 es el logaritmo del segundo formante,

la energía o una función de la energía tal como *logE* donde *logE* es el logaritmo de la energía,

*C1*, donde *C1* es la energía de baja frecuencia dividida por la energía de alta frecuencia o una función de *C1*,

y derivadas temporales de cualquiera de los rasgos anteriores tales como la derivada temporal de *logPitch*, *logF1*, *logF2*, *logE* y *C1*.

Las derivadas de los rasgos mencionados anteriormente se denominan a continuación con una D adicional delante tal como, por ejemplo, *DlogPitch*, *DlogF1*, *y DlogF2*.

Si en un segmento de tiempo no puede determinarse un rasgo entonces este segmento de tiempo se elimina del rasgo.

Las medidas de distancia utilizadas en contexto con alineamiento temporal dinámico y la medida de la distancia total pueden definirse como

una distancia euclidiana

$$d^2 = \sum_k \int_t (r_k(t) - s_k(t))^2 dt$$

una distancia de Mahalanobis

$$d^2 = \sum_k \int_t \frac{(r_k(t) - s_k(t))^2}{\sigma_k^2} dt$$

y/o una distancia coseno

$$d^2 = \sum_k \frac{\vec{r}_k \cdot \vec{s}_k}{\| \vec{r}_k \| \cdot \| \vec{s}_k \|}$$

donde r y s son características que dependen del tiempo con índice k de una pluralidad de características (en el caso de que k sea sólo 1 existiendo solamente una característica a tener en cuenta), y donde s es una característica extraída de la primera expresión de voz y r es una característica extraída de la segunda expresión de voz. La distancia de Mahalanobis incluye, además, un rango de variación σ para cada característica. En el caso de la distancia coseno para calcular la distancia se utilizan, en cambio, vectores de características que dependen del tiempo de segmentos de tiempo. Aquí cada entrada del vector representa un instante diferente para el cual se da el valor de la característica.

Pueden utilizarse en su lugar otras funciones de distancia.

El rango de variación σ, que se utiliza para calcular la distancia de Mahalanobis, puede calcularse teniendo en cuenta características de varias expresiones de voz. El σ es una medida de la variabilidad (por ejemplo desviación estándar) del valor alrededor de su valor medio (a media que transcurre el tiempo). Por ejemplo, σ se calcula teniendo en cuenta una característica de la primera expresión de voz y o la característica correspondiente de la segunda expresión de voz, o σ se calcula teniendo en cuenta las características correspondientes de varias versiones de la primera expresión de voz y/o las características correspondientes de varias versiones de la segunda expresión de voz (por ejemplo, en caso de que la primera expresión de voz haya sido grabada varias veces, o si la segunda expresión de voz se solicita y se recibe varias veces).

Además, el rango de variación σ, que se utiliza para calcular la distancia de Mahalanobis, puede calcularse teniendo en cuenta una única característica de una expresión de voz, tal como por ejemplo la primera expresión de voz. Pueden utilizarse características que dependen del tiempo para calcular el rango de variación σ, ya sea antes o después de que se haya aplicado alineamiento temporal dinámico a dicha característica.

En otros procedimientos, la medida de la distancia total se calcula en base a un único par de características que dependen del tiempo en el que cada característica que depende del tiempo es una característica de un único rasgo. O, la medida de la distancia total se calcula en base a un único par de características que dependen del tiempo en el que cada característica que depende del tiempo es una característica de una combinación de una pluralidad de rasgos. O la medida de la distancia total se calcula en base a una pluralidad de pares de características que dependen del tiempo en el que cada característica que depende del tiempo es una característica de un único rasgo. O, la medida de la distancia total se calcula en base a una pluralidad de pares de características que dependen del tiempo en el que cada característica que depende del tiempo es una característica de un único rasgo o bien característica de una combinación de una pluralidad de rasgos. O, la medida de la distancia total se calcula en base a una pluralidad de pares de características que dependen del tiempo en el que cada característica que depende del tiempo es una característica de una combinación de una pluralidad de rasgos.

En el caso en que los rasgos se combinan con el fin de formar una característica que depende del tiempo, puede combinarse 2, 3, 4 ó 5 o cualquier número de rasgos, donde el número de rasgos típicamente es menor de 10. Además, el número de pares utilizados para calcular una medida de la distancia total puede ser 1, 2, 3, 4, 5 ó cualquier número de pares, que típicamente es menor de 10.

En otro procedimiento, se calcula una pluralidad de medidas de distancia total, y la comparación de la primera expresión de voz con la segunda expresión de voz se basa en la pluralidad de medidas de distancia total seleccionando una o más medidas de distancia total de la pluralidad de medidas de distancia total y, además, o alternativamente, combinando por lo menos dos medidas de distancia total o combinaciones de las mismas. Por ejemplo, una ventaja de calcular dos o más medidas de distancia total es que las medidas pueden compararse. Si las medidas de distancia total concuerdan bien entre sí el resultado de cada comparación puede confiarse más que en el caso en que las medidas de distancia total den resultados significativamente diferentes.

Otros aspectos de posibles realizaciones de la invención quedan claros a partir de las figuras 1, 2 y 3:

La figura 1 resume diferentes casos que pueden darse al comparar expresiones de voz,

La figura 2 es un diagrama de flujo de un procedimiento para comprar expresiones de voz, y

La figura 3 es un diagrama de flujo de un procedimiento para probar la corrección de una (segunda) expresión de voz.

La figura 1 muestra una gráfica que resume diferentes situaciones cuando se realiza una comparación de expresiones de voz. Las medidas de distancia que se utilizan para alineamiento temporal dinámico (DTW) pueden llevarse a cabo para características que dependen del tiempo, donde una característica que depende del tiempo es una característica de un único rasgo (columna izquierda) o bien de una combinación de por lo menos dos rasgos (columna derecha). La medida de la distancia total se calcula en base a pares de las características que dependen del tiempo mencionadas anteriormente. El cálculo de una medida de distancia total se basa en un único par de características que dependen del tiempo (línea superior) o bien se basa en una pluralidad de pares de características que dependen del tiempo (línea inferior).

Las columnas y las líneas mencionadas anteriormente se cruzan en cinco campos de intersección 1, 2, 3, 4, 5 que representan cinco casos distintos (números romanos).

El caso I es la situación en la que se calcula la medida de la distancia total en base a un único par de características que dependen del tiempo en el que cada característica que depende del tiempo (utilizada para DTW) es una característica de un único rasgo. Por ejemplo, la medida de la distancia total se basa en un par de características *C1*, donde una característica *C1* se extrae de la primera expresión de voz y la otra característica *C1* se extrae de la segunda expresión de voz.

En el caso II, la medida de la distancia total se calcula en base a un único par de características que dependen del tiempo en el que cada característica que depende del tiempo es una característica de una combinación de una pluralidad de rasgos. Por ejemplo, la medida de la distancia total se calcula en base a un único par de características que dependen del tiempo, donde cada característica que depende del tiempo de ese par es una combinación de *logF1* y *logF2*.

En caso III, la medida de la distancia total se calcula en base a una pluralidad de pares de características que dependen del tiempo, donde cada característica que depende del tiempo es una característica de un único rasgo. Por ejemplo, la medida de la distancia total se calcula en base a tres pares de características que dependen del tiempo donde las características que dependen del tiempo del primer par son características de *logPitch*, donde las características que dependen del tiempo del segundo par son características de *logF1*, y donde las características que dependen del tiempo del tercer par son características de *logF2*.

El caso IV es la situación en la que la medida de la distancia total se calcula en base a una pluralidad de pares de características que dependen del tiempo donde cada característica que depende del tiempo es una característica de un único rasgo o bien característica de una combinación de una pluralidad de rasgos. En otras palabras, el caso IV es una mezcla de la columna de la izquierda con la columna de la derecha de la figura 1. Por ejemplo, la medida de la distancia total se calcula en base a tres pares de características que dependen del tiempo, donde las características que dependen del tiempo del primer par es una característica de *logPitch*, donde las características que dependen del tiempo del segundo par son características de combinaciones de *logF1* y *logF2*, y donde las características para el tercer par son características para *C1*. Este ejemplo particular resultó ser el más efectivo para distinguir claramente entre expresiones de voz generadas por cortar y pegar y expresiones de voz generadas de manera normal, permitiendo incluso una EER (tasa de error igual) cero en una prueba en particular, lo que significa que todas las 120 expresiones de voz pueden ser identificadas correctamente como de cortar y pegar o como normales. Tal como puede apreciarse a partir de este ejemplo, el cálculo de una medida de la distancia total en base a una característica que es un rasgo único y una característica que es una combinación de rasgos resulta ser particularmente ventajoso.

En el caso V, la medida de la distancia total se calcula en base a una pluralidad de pares de características que dependen del tiempo, donde cada característica que depende del tiempo es una característica de una combinación de una pluralidad de rasgos. Por ejemplo, la medida de la distancia total se calcula en base a dos pares de características que dependen del tiempo, donde las características que dependen del tiempo del primer par son una característica de las combinaciones de *logPitch* y *DlogPitch*, y donde las características que dependen del tiempo del segundo par son características de combinaciones de *logF1*, *logF2*, *DlogF1* y *DlogF2*.

A partir de los casos y ejemplos mencionados anteriormente queda claro que existen muchas maneras disponibles para calcular la medida de distancia total. La mejor manera de calcular la medida de distancia total puede depender del tipo de aplicación en la que se emplea la comparación de expresiones de voz. Para una aplicación específica, es posible determinar una configuración que funcione mejor realizando pruebas basadas en muestras de ensayo. Por ejemplo, una primera muestra de ensayo contiene primeras expresiones de voz, una segunda muestra de ensayo contiene correspondientes segundas expresiones de voz, y una tercera muestra de ensayo contiene correspondientes segundas expresiones de voz que han sido producidas uniendo secuencias de expresiones de voz entre sí (con el fin de simular una falsificación por cortar y pegar). Entonces, una primera expresión de voz de la primera muestra de ensayo puede compararse con una correspondiente segunda expresión de voz de la segunda muestra de ensayo, y la misma primera expresión de voz de la primera muestra de ensayo puede compararse con la correspondiente segunda expresión de voz de la tercera prueba de ensayo. Estas comparaciones con segundas expresiones de voz a partir de la segunda muestra de ensayo y la tercera muestra de ensayo pueden repetirse varias veces con el fin de permitir un análisis estadístico de los resultados de la comparación. De esta manera puede probarse lo bien que una medida de la distancia total particular puede separar comparaciones con segundas expresiones de voz a partir de la segunda muestra de ensayo de segundas expresiones de voz a partir de la tercera muestra de ensayo. El poder de separación puede cuantificarse, por ejemplo, calculando la tasa de error igual (EER) o calculando la función de coste en base al cociente de probabilidad logarítmica mínimo (minCllr).

La figura 2 muestra un diagrama de flujo que representa el procedimiento para comparar expresiones de voz. El procedimiento comienza en la etapa 200. En la etapa 201, se extrae una pluralidad de características de la primera expresión de voz y se extrae la correspondiente pluralidad de rasgos de una segunda expresión de voz. Después, en la etapa 202, se aplica un alineamiento temporal dinámico (DTW) a una o más características que dependen del tiempo de la segunda expresión de voz tal que se minimizan, por ejemplo, medidas de distancia correspondientes. Una medida de la distancia es una medida de la diferencia de una característica que depende del tiempo que representa rasgos de las primeras expresiones de voz y una característica que depende del tiempo correspondiente que representa rasgos de la segunda expresión de voz, cuando una característica que depende del tiempo de una expresión de voz es una característica que depende del tiempo de un único rasgo o bien una combinación de varios rasgos.

Para dar un ejemplo, se consideran los dos rasgos F1 y F2. $F1_1$ es el primer rasgo de la primera expresión de voz y $F1_2$ es el primer rasgo de la segunda expresión de voz. $F2_1$ es el segundo rasgo de la primera expresión de voz y $F2_2$ es el segundo rasgo de la segunda expresión de voz. Todos estos rasgos dependen del tiempo. Los rasgos $F1_2$ y $F2_2$ han de someterse a alineamiento temporal dinámico para adaptarse mejor a $F1_1$ y $F2_1$ respectivamente. En el caso de que el rasgo $F1_2$ se someta a alineamiento temporal dinámico a la característica $F1_1$ independientemente

del rasgo $F2_1$ o $F2_2$ (e independientemente de cualquier otro rasgo) entonces cada rasgo se considera que es, en sí mismo, una característica. Los dos rasgos F1 y F2, en otro procedimiento, pueden someterse a alineamiento temporal conjuntamente. Esto significa que la deformación en el eje de tiempo (estiramiento o compresión del rasgo en partes del eje de tiempo) tiene que realizarse igualmente para ambos rasgos F1 y F2. El cálculo de la distancia entre $F1_1$ y $F1_2$ por una parte y $F2_1$ y $F2_2$ por otra parte se utiliza para que el alineamiento temporal dinámico tenga en cuenta ambos pares. Con las fórmulas de distancia mencionadas anteriormente las distancias de ambos rasgos se calculan y, por ejemplo, se suman. Éste es un ejemplo de una combinación de dos rasgos que forman, de ese modo, una característica. De la misma manera, pueden combinarse tres o más rasgos en una característica.

El alineamiento temporal dinámico puede llevarse a cabo varias veces teniendo en cuenta una combinación diferente de rasgos (características) o rasgos individuales, que son características. Cada cálculo del alineamiento temporal dinámico puede dar un alineamiento temporal diferente. Por ejemplo, para un rasgo F1 puede obtenerse un alineamiento temporal diferente entonces que para el rasgo F2 o para la combinación del rasgo F1 con F2. Pueden utilizarse también individualmente uno o más rasgos como una característica en sí misma y también pueden utilizarse en combinación con otro rasgo para formar una característica. Por ejemplo, el rasgo F1 puede utilizarse como una característica y F1 y F2 pueden combinarse para formar una característica.

En la etapa 203 se evalúa o se calcula una medida de la distancia total. La medida de la distancia total es una medida de la diferencia entre la primera expresión de voz de la muestra de texto determinada y la segunda expresión de voz de dicha muestra de texto determinada donde la medida de la distancia total se calcula en base a uno o más pares de dichas características que dependen del tiempo. Un par de características que dependen del tiempo está compuesto por una característica que depende del tiempo de la primera expresión de voz y de una característica que depende del tiempo (202) por alineamiento temporal dinámico de la segunda expresión de voz. (El par de características que dependen del tiempo también puede estar compuesto por una característica que depende del tiempo por alineamiento temporal dinámico de la primera expresión de voz y de una característica que depende del tiempo de la segunda expresión de voz, o el par de características que dependen del tiempo puede estar compuesto también por una característica que depende del tiempo por alineamiento temporal dinámico de la primera expresión de voz y de una característica que depende del tiempo por alineamiento temporal dinámico de la segunda expresión de voz). Después, el procedimiento termina en la etapa 299. En lugar de tomar el rasgo/característica de la segunda expresión de voz en la versión por alineamiento temporal dinámico, también puede tenerse en cuenta el (los) de la primera expresión de voz.

Además, los resultados de los cálculos de distancias realizadas durante el alineamiento temporal dinámico pueden utilizarse para determinar la distancia total si es posible.

El alineamiento temporal dinámico puede incluir relaciones lineales entre el eje de tiempo original y el eje de tiempo deformado. La relación puede ser parcialmente lineal o puede ser cualquier función monótonamente creciente.

La figura 3 muestra un diagrama de flujo que representa un procedimiento para validar una persona que habla o detectar una falsificación cuando se comparan expresiones de voz. El procedimiento comienza en la etapa 300. En la etapa 301, se obtiene una primera expresión de voz. Puede haberse grabado previamente una primera expresión de voz (por ejemplo, en una sesión de inscripción o en una solicitud anterior para decir la expresión de voz) y puede obtenerse a partir de, por ejemplo, un almacén de datos o una memoria. En la etapa 302, se requiere una segunda expresión de voz de una persona que habla, y en la etapa 303 se recibe la segunda expresión de voz de dicha persona que habla. Entonces, en la etapa 304, la primera expresión de voz se compara con la segunda expresión de voz. En base a la comparación de la etapa 304, la persona que habla es validada en la etapa 305 si la primera expresión de voz coincide bien con la segunda expresión de voz. De lo contrario, si la primera expresión de voz no coincide bien con la segunda expresión de voz, se concluye que la segunda expresión de voz es el resultado de una falsificación, tal como una falsificación por cortar y pegar, en la etapa 306. Después de la etapa 305 ó 306, el procedimiento termina en la etapa 399.

El procedimiento mencionado anteriormente es solamente un ejemplo de cómo se puede emplearse la comparación de las expresiones de voz para una aplicación. Hay muchas otras posibilidades de aplicaciones posibles, tales como emplear la comparación de las expresiones de voz con el fin de detectar que la persona que habla de la segunda expresión de voz no corresponde con la persona que habla de la primera expresión de voz, por ejemplo.

El procedimiento descrito puede formar parte de la prueba pasiva de falsificación que se describe en la solicitud mencionada anteriormente PCT/EP2008/010478 o PCT/EP2009/004649.

**REIVINDICACIONES**

1. Procedimiento para comparar expresiones de voz, comprendiendo el procedimiento las etapas de:

extraer una pluralidad de rasgos (201) de una primera expresión de voz de una muestra de texto determinada y extraer una pluralidad de rasgos (201) de una segunda expresión de voz de dicha muestra de texto determinada, en el que cada rasgo se extrae en función del tiempo, y en el que cada rasgo de la segunda expresión de voz corresponde a un rasgo de la primera expresión de voz;

aplicar alineamiento temporal dinámico (202) a por lo menos dos características que dependen del tiempo de la primera y/o la segunda expresión de voz minimizando una o más medidas de distancia, en el que una medida de distancia es una medida de la diferencia de una característica que depende del tiempo de la primera expresión de voz y una característica que depende del tiempo correspondiente de la segunda expresión de voz;

calcular una medida de distancia total (203), en el que la medida de la distancia total es una medida de la diferencia entre la primera expresión de voz de la muestra de texto determinada y la segunda expresión de voz de dicha muestra de texto determinada, en el que la medida de la distancia total se calcula (203) en base a una pluralidad de pares de características que dependen del tiempo, y en el que un par de características que dependen del tiempo está compuesto por una característica que depende del tiempo de la primera o la segunda expresión de voz y de una característica que depende del tiempo (202) por alineamiento temporal dinámico respectivamente de la segunda o la primera expresión de voz, o en el que un par de características que dependen del tiempo está compuesto por una característica que depende del tiempo (202) por alineamiento temporal dinámico de la primera expresión de voz y una característica que depende del tiempo (202) por alineamiento temporal dinámico de la segunda expresión de voz; y

en el que por lo menos una característica que depende del tiempo es una característica de un único rasgo y por lo menos una otra característica que depende del tiempo es una característica de una combinación de una pluralidad de rasgos (4) a los que se aplica el mismo alineamiento temporal dinámico (202),

en el que la primera expresión de voz ha sido grabada previamente, y en el que la segunda expresión de voz se recibe (302) de una persona que habla, a petición (303), y

en el que la medida de la distancia total se emplea para

- detectar que la segunda expresión de voz es el resultado de una falsificación por cortar y pegar (306) si la primera expresión de voz no coincide bien con la segunda expresión de voz, y

- validar a la persona que habla si la primera expresión de voz coincide bien con la segunda expresión de voz.

2. Procedimiento según la reivindicación 1, caracterizado por el hecho de que la pluralidad de rasgos comprende uno o más de los siguientes rasgos:

el tono o una función del mismo tal como el *logPitch*, donde *logPitch* es el logaritmo del tono,

el primer formante o una función del mismo tal como *logF1*, donde *logF1* es el logaritmo del primer formante,

el segundo formante o una función del mismo tal como *logF2*, donde *logF2* es el logaritmo del segundo formante,

la energía o una función de la misma tal como l*ogE*, donde *logE* es el logaritmo de la energía,

*C1* o una función del mismo, donde *C1* es la energía de baja frecuencia dividida por la energía de alta frecuencia,

y derivadas temporales de cualquiera de los rasgos anteriores tales como la derivada temporal de *logPitch*, *logF1*, *logF2*, *logE* y *C1*.

3. Procedimiento según una de las reivindicaciones 1 a 2, **caracterizado** por el hecho de que una medida de la distancia de alineamiento temporal dinámico se define como una distancia euclidiana, una distancia de Mahalanobis o una distancia coseno.

4. Procedimiento según una de las reivindicaciones 1 a 3, **caracterizado** por el hecho de que la medida de la distancia total se define como una distancia euclidiana, una distancia de Mahalanobis o una distancia coseno.

5. Procedimiento según una de las reivindicaciones 1 a 4, **caracterizado** por el hecho de que se calcula una pluralidad de medidas de distancia total (203), y en el que la comparación de la primera expresión de voz con la segunda expresión de voz se basa en la pluralidad de medidas de distancia total seleccionando una o más medidas de distancia total de la pluralidad de medidas de distancia total y/o combinando por lo menos dos medidas de distancia total.

6. Medio informático que comprende instrucciones ejecutables por un ordenador para realizar cualquiera de los procedimientos de las reivindicaciones 1 a 5.

7. Aparato que está configurado para realizar cualquiera de los procedimientos de las reivindicaciones 1 a 5.

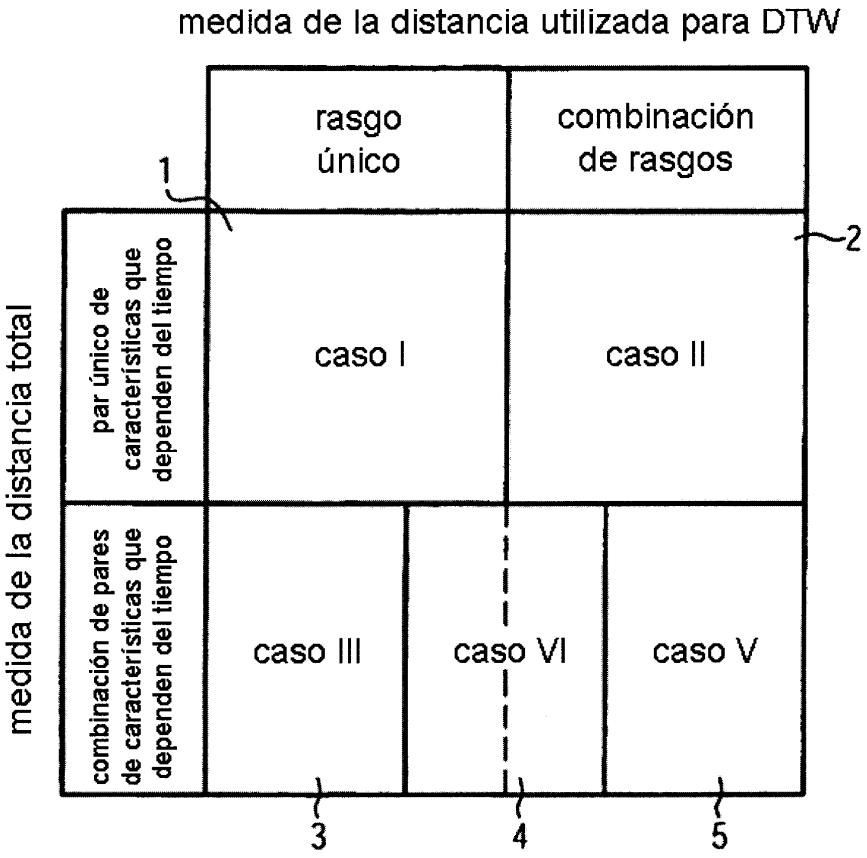
medida de la distancia utilizada para DTW
rasgo único
combinación de rasgos
1
2
medida de la distancia total
par único de características que dependen del tiempo
combinación de pares de características que dependen del tiempo
caso I
caso II
caso III
caso VI
caso V
3
4
5

FIG. 1

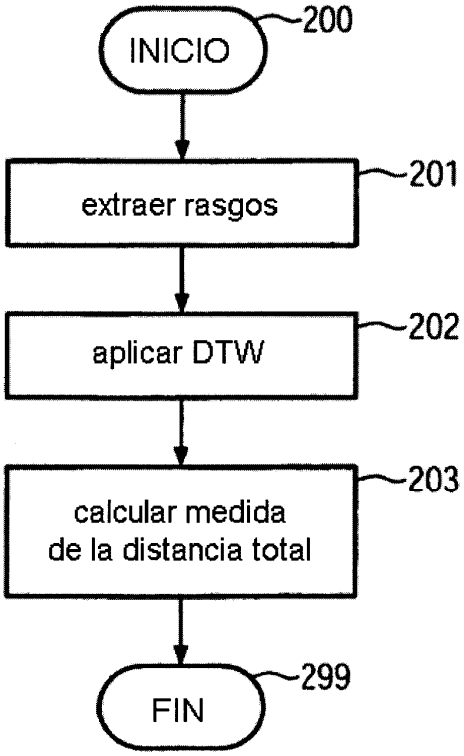
INICIO
200
extraer rasgos
201
aplicar DTW
202
calcular medida
de la distancia total
203
FIN
299

FIG. 2

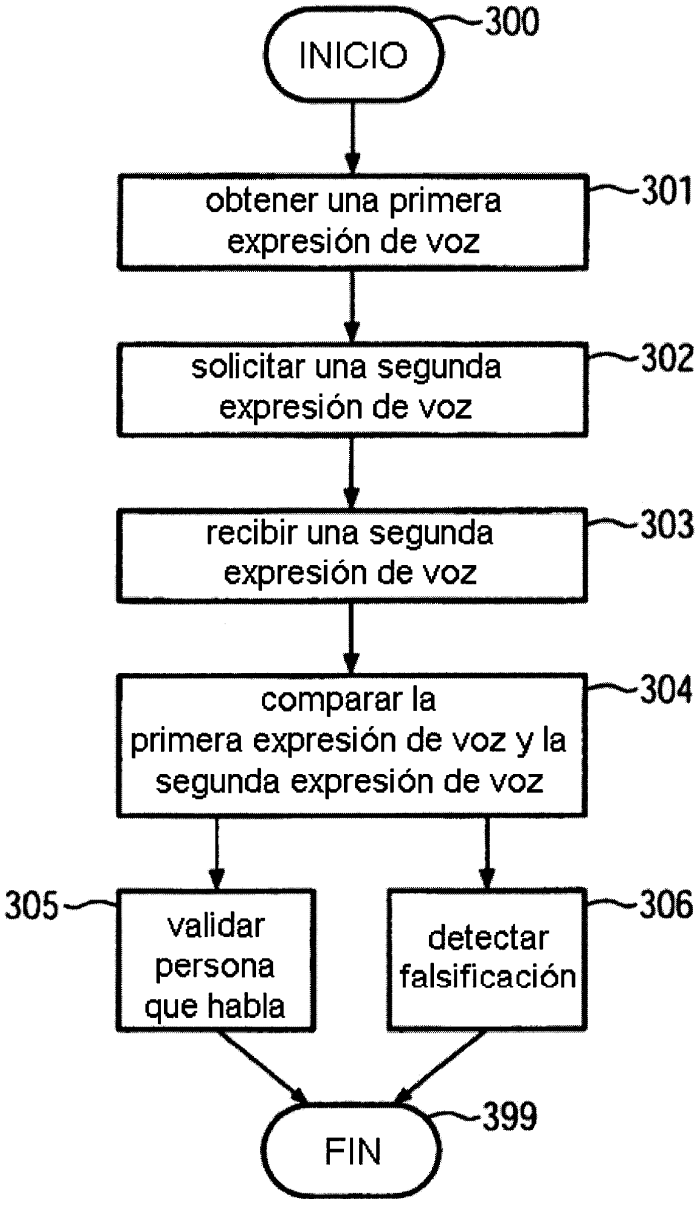
INICIO
300
obtener una primera expresión de voz
301
solicitar una segunda expresión de voz
302
recibir una segunda expresión de voz
303
comparar la primera expresión de voz y la segunda expresión de voz
304
305
validar persona que habla
detectar falsificación
306
FIN
399

FIG. 3