(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 540 995**

(51) Int. CI.:

G10L 15/02 (2006.01)
G10L 15/06 (2013.01)
G10L 15/16 (2006.01)
G10L 15/12 (2006.01)
G10L 15/20 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **24.08.2011 E 11757206 (5)**

(97) Fecha y número de publicación de la concesión europea: **01.04.2015 EP 2609587**

(54) Título: **Sistema y método para reconocer un comando de voz de usuario en un entorno con ruido**

(30) Prioridad:

24.08.2010 CH 13542010

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
15.07.2015

(73) Titular/es:

VEOVOX SA (100.0%)
Chemin des Roches 10, CP 508
1009 Pully, CH

(72) Inventor/es:

DINES, JOHN;
CARMONA, JORGE;
MASSON, OLIVIER y
ARADILLA, GUILLERMO

(74) Agente/Representante:

CURELL AGUILÁ, Mireia

DESCRIPCIÓN

Sistema y método para reconocer un comando de voz de usuario en un entorno con ruido.

Campo de la invención

La presente invención se refiere a un método y a un sistema para introducir y reconocer comandos de voz de usuario en entornos con ruido.

Descripción de la técnica relacionada

Se conocen sistemas de Reconocimiento Automático de Voz (ASR) para reconocer palabras pronunciadas de señales de audio.

El ASR se usa, por ejemplo, en centros de atención telefónica. Cuando un usuario necesita cierta información, por ejemplo la hora de salida de un tren desde una estación de ferrocarril determinada, el mismo puede solicitar oralmente la información deseada a un centro de atención telefónica utilizando un teléfono o teléfono móvil. El usuario habla y solicita información, y un sistema reconoce las palabras pronunciadas para recuperar datos. Otro ejemplo de ASR es una guía telefónica automática. Un usuario puede llamar a un número y solicitar oralmente el número de teléfono de una persona que viva en una ciudad.

En estos dos ejemplos, la voz del usuario se reconoce y se convierte en una señal eléctrica usada por un sistema para recuperar la información deseada. En ambos casos, el canal de comunicaciones entre el usuario y el sistema es conocido. Puede ser un par trenzado en el caso de un teléfono fijo o el canal del operador en el caso de un teléfono móvil. Además, el ruido del canal se puede modelar con modelos conocidos.

Por otra parte, dichas aplicaciones implican típicamente un número elevado de palabras posibles que pueden ser usadas por un hablante. Por ejemplo, el número de posibles estaciones de ferrocarril diferentes, el número de las ciudades de un país o el número de los nombres de las personas que viven en una cuidad dada es habitualmente muy elevado. No obstante, el diccionario de palabras a reconocer no requiere una actualización frecuente, ya que las mismas no cambian muy a menudo; habitualmente, cada palabra del diccionario permanece sin variaciones en este diccionario durante semanas, meses o incluso años.

Además de aquellas aplicaciones conocidas para las cuales se han desarrollado la mayoría de ASR, en ocasiones son necesarios otros ASR en entornos con ruido, tales como (sin carácter limitativo) bares, restaurantes, discotecas, hoteles, hospitales, industria del entretenimiento, tiendas de comestibles, etcétera, para coger, reconocer y transmitir pedidos por voz. Por ejemplo, resultaría útil disponer, en un bar o restaurante, de un ASR con el cual un camarero que coge pedidos de los clientes sentados en una mesa pudiera repetir cada pedido y pronunciarlo a un micrófono de un dispositivo móvil. A continuación, la señal de voz recibida desde un punto de acceso se podría convertir en comandos de texto mediante un servidor de reconocimiento de voz que ejecuta un algoritmo de reconocimiento de voz. El punto de acceso puede pertenecer a una red de área local (LAN), a la cual se conectan otros diversos equipos, tales como el servidor.

En dichos entornos, el reconocimiento de voz resulta difícil ya que la relación de señal/ruido puede ser insuficiente. Por otra parte el ruido del entorno no es conocido y el mismo puede cambiar en función de la gente que haya en el bar o restaurante. Las palabras posibles pronunciadas por el hablante pueden ser, por ejemplo, las palabras contenidas en el menú del día. En ese caso, el número de palabras está limitado normalmente a por ejemplo, unos cuantos cientos de palabras como mucho. Por otra parte estas palabras pueden cambiar cada día – en el caso del menú del día – o, por ejemplo, cada semana o dos veces por mes. Por lo tanto, los requisitos de un sistema de reconocimiento de voz en un entorno del tipo mencionado o para una aplicación del tipo mencionado son muy diferentes con respecto a los requisitos en los que se basan la mayoría de sistemas de ASR disponibles comercialmente.

Es por lo tanto una finalidad de la presente invención desarrollar un método y un aparato de ASR nuevos que estén adaptados más adecuadamente a esos requisitos tan específicos e inhabituales (relación deficiente de señal/ruido, diccionario limitado, diccionario que varía rápidamente, precisión de reconocimiento muy alta, tiempo de respuesta inmediato, idioma de usuario, independencia con respecto a la pronunciación y el acento, robustez frente al entorno y el hablante).

Se conocen diferentes sistemas de ASR. La Figura 1 muestra un ASR por concordancia de plantillas, que es uno de los primeros sistemas de ASR. El mismo usa ejemplos de unidades de habla como base para reconocimiento. Cada ejemplo actúa como una plantilla o secuencia de plantillas 14 para una unidad de habla o secuencia de prueba específica 12 que va a ser reconocida. Puede haber múltiples plantillas 14 para cada unidad 12 con el fin de incrementar la robustez del sistema. Habitualmente, estas unidades se representan en forma de secuencias de características espectrales a corto plazo, tales como Coeficientes Cepstrales en frecuencia Mel (MFCCs). El Cepstrum en Frecuencia Mel (MFC) es una representación particular del espectro de potencia a corto plazo de un

sonido utilizado en el procesado del mismo. Se basa en una transformada discreta de coseno de un espectro de potencia logarítmico representado en la escala Mel de frecuencia.

Un decodificador 10 lleva a cabo una comparación entre observaciones acústicas de plantillas 140, 142, 144 y secuencias de prueba 12. La similitud acústica se deduce habitualmente a partir de medidas de la distancia entre la característica acústica de la plantilla $O^{plantilla}$ y la característica acústica de la secuencia de prueba O^{prueba}. En el ASR convencional por concordancia de plantillas, se usa un parámetro de distancia para medir la similitud de vectores acústicos. Los parámetros de distancia adecuados se basan normalmente en la distorsión espectral. La medida de la distancia puede ser, por ejemplo, euclídea:

$$D_E = \left\| O^{plantilla} - O^{prueba} \right\|$$

o Mahalanobis:

$$D_M = \sqrt{(O^{plantilla} - O^{prueba})^T S^{-1} (O^{plantilla} - O^{prueba})}$$

en donde S representa la matriz de covarianza de los vectores acústicos.

La suposición de fondo del ASR basado en la concordancia de plantillas es que las ejecuciones de sonidos sean suficientemente similares, de tal manera que una comparación entre observaciones acústicas de plantillas correctas 14 y secuencias de prueba 12 proporcione una coincidencia relativamente buena en comparación con el cálculo de plantillas incorrectas. En el caso de un usuario dado y/o de condiciones de grabación estables, puesto que cada usuario proporciona normalmente sus propias plantillas las cuales contienen su propia pronunciación específica, el ASR basado en la concordancia de plantillas es independiente de pronunciaciones e idiomas.

En el caso de una pluralidad de usuarios y/o de condiciones de grabación diferentes, lo mencionado anteriormente en realidad no se cumple normalmente debido a posibles variaciones en la pronunciación para la misma palabra. Estas variaciones pueden ser generadas por diferencias de pronunciación entre hablantes y/o discordancias entre condiciones de grabación.

La ventaja del ASR por concordancia de plantillas es su sencillez de implementación. Por otra parte, no requiere la especificación de un diccionario de pronunciación, ya que éste está implícito en la plantilla 14. Las desventajas incluyen la sensibilidad antes mencionada a diferencias en grabaciones de plantillas/pronunciaciones de prueba en el caso de una pluralidad de usuarios y/o de condiciones de grabación diferentes. Por otra parte, el algoritmo de reconocimiento puede ser costoso desde el punto de vista computacional cuando se utiliza un número elevado de plantillas.

La decodificación en el ASR con concordancia de plantillas se lleva a cabo frecuentemente utilizando un decodificador de Alineamiento Dinámico Temporal (DTW) 10. El mismo implementa un algoritmo de programación dinámico que efectúa una búsqueda sobre la secuencia de prueba, donde se consideran todas las secuencias de plantillas posibles. Durante esta búsqueda se permite cierta deformación temporal de plantillas, y de aquí el aspecto de “alineamiento temporal” de este algoritmo. La plantilla reconocida q^* es

$$q^* = \arg\min_q DTW\left(O^{plantilla(q)}, O^{prueba}\right)$$

donde $DTW(\cdot,\cdot)$ es la distancia de alineamiento dinámico temporal entre dos secuencias, la secuencia de plantillas 14 $O^{plantilla(q)}$ de la biblioteca de plantillas y la secuencia de prueba 12 O^{prueba}.

Se ilustra un ejemplo del procedimiento de búsqueda de DTW en la Figura 2, en la que se calcula una puntuación local 16 en cada nodo del gráfico entre una secuencia de prueba 12 y la plantilla correspondiente 14. La puntuación de DTW 16 es una acumulación de las puntuaciones locales recorridas en el gráfico, donde el trayecto recorrido 20 minimiza la puntuación acumulada total; se corresponde por tanto con una distancia entre la palabra pronunciada y la plantilla de referencia.

En la Figura 1, por ejemplo, la palabra pronunciada se corresponde con la plantilla 142 puesto que esta plantilla tiene la puntuación mínima (puntuación = 20) en comparación con otras plantillas 140 (puntuación = 40) y 144 (puntuación = 30).

Resulta sencillo extender el ASR con plantillas a un reconocimiento de voz continuo mediante la concatenación de plantillas entre sí durante la programación dinámica.

Normalmente, los sistemas de ASR basados en plantillas comprenden los siguientes componentes:

- Etapa frontal de procesado de audio: convierte entradas de audio en características acústicas (las secuencias de pruebas 12) usadas en el sistema de reconocimiento.

- Biblioteca de plantillas: contiene todas las secuencias de plantillas 14 que pueden ser reconocidas por el sistema.

- Software de decodificación basado en el DTW: hace concordar secuencias de prueba 12 con secuencias de plantillas 14 de la biblioteca de plantillas.

- Gramática (opcional): describe las secuencias de plantillas permisibles 14 que pueden ser reconocidas (la sintaxis).

El ASR basado en plantillas ha sido sustituido en gran medida por técnicas de reconocimiento de voz estadísticas en los sistemas de ASR. Dichos sistemas proporcionan una mejor robustez a variaciones no deseadas del entorno y del hablante y presentan una mejor escalabilidad para requisitos de vocabularios grandes. Por contraposición, dichos sistemas requieren recursos mayores para el entrenamiento de modelos, dependen del idioma y carecen de robustez cuando se tratan acentos particulares.

Dichos sistemas se basan típicamente en el HMM (Modelo Oculto de Markov). Un HMM es un proceso de Markov, es decir, para un proceso de Markov de comando n-ésimo, la verosimilitud de un estado futuro dado, en cualquier momento dado, depende únicamente de los n estados previos, y en donde los estados no son directamente observables. Cada estado genera un evento, es decir, un vector de observación acústico, con una función particular de densidad de probabilidad de emisión de estados que depende únicamente del estado. El vector de observación acústica es observable, el estado no.

El conjunto finito de estados en un HMM/ASR modela unidades acústicas con las cuales se constituye una palabra pronunciada. Típicamente, la unidad acústica fundamental es el fonema; algunos sistemas usan trifonos, u otras unidades.

Aún cuando “fono” y “fonema” se usan frecuentemente de manera intercambiable, los mismos tienen un significado exacto ligeramente diferente. La fonética hace referencia al estudio de sonidos del habla, en particular los patrones de sonidos que caracterizan la estructura fonológica subyacente del habla. Por tanto, la información fonética en una pronunciación hace referencia a las unidades de habla fundamentales que comprenden la pronunciación, fonos, donde los fonos son materializaciones acústicas de las unidades funcionales del idioma conocidas como fonemas. La diferenciación entre fonemas y fonos reside en el hecho de que los primeros son un constructo lingüístico y los segundos un constructo acústico. A los fonos se les hace referencia normalmente como características segmentales del habla.

Se puede definir un conjunto de fonemas para un idioma específico (por ejemplo, Alfabeto Fonético de la Agencia de Proyectos de Investigación Avanzada – ARPABET) o para todos los idiomas (por ejemplo, Alfabeto Fonético Internacional – IPA). Con respecto al ASR, la elección del conjunto de fonos es algo arbitraria (es decir, existen varios conjuntos de fonos usados comúnmente para el inglés americano), aunque el mismo se debe usar consistentemente en la definición de la relación de pronunciación entre fonemas y palabras.

Los vectores de observación acústicos son típicamente los mismos que las características acústicas utilizadas en el ASR basado en plantillas.

El reconocimiento de la voz se lleva a cabo habitualmente utilizando una decodificación de Viterbi. El dispositivo de reconocimiento da salida a la secuencia más probable Q^* de acuerdo con la siguiente relación:

$$Q^* = \arg\max_{Q} p(O^T, Q)$$

$$= \arg\max_{Q=q_1,\ldots,q_T} \prod_{t=1}^{T} p(q_t|q_{t-1}) p(o_t|q_t)$$

donde $O^T = [o_1,\ldots, o_T]$ representa el vector de observación acústica, $Q = q_1,\ldots, q_T$ representa la secuencia de estados $q_1,\ldots,q$ para una hipótesis de reconocimiento, $p(o_t|q_t)$ es la verosimilitud generada por la función de densidad de probabilidad de emisión de estados correspondiente al estado q_t y $p(q_t|q_{t-1})$ es la probabilidad de transición entre estados ocultos q_t y q_{t-1}.

La verosimilitud $p(o_t|q_t)$ se puede generar mediante un número ilimitado de modelos estadísticos. Lo más habitual es utilizar un Modelo de Mezcla de Gaussianas (GMM). En sistemas híbridos de HMM/MLP, se usa un Perceptrón Multicapa (MLP) para generar las verosimilitudes de emisión de estados, utilizando la denominada “verosimilitud escalada”. Un Perceptrón es un tipo de red neuronal artificial constituido por un clasificador binario. Un MLP es un tipo particular de red neuronal artificial que utiliza más capas de neuronas.

En sistemas híbridos de HMM/MLP, el MLP se entrena para estimar la probabilidad a posteriori $p(q_t|o_t)$ de clases de fonos. Un MLP se entrena de una manera discriminativa para reconocer clases diferentes relevantes para una tarea particular (por ejemplo, voz/no voz, reconocimiento de objetos, etcétera). En el reconocimiento de voz, las clases se especifican como el conjunto de fonos/fonemas que cubre los datos de entrenamiento/datos de prueba (habitualmente, aunque no siempre, para un idioma específico). Esta probabilidad a posteriori $p(q_t|o_t)$ está relacionada con la verosimilitud $p(o_t|q_t)$ generada por la función de densidad de probabilidad de emisión de estados para el estado q_t de acuerdo con la relación de Bayes:

$$p(o_t|q_t) = \frac{p(q_t|o_t)p(o_t)}{p(q_t)} \approx \frac{p(q_t|o_t)}{p(q_t)}$$

donde $p(q_t|o_t)$ es la probabilidad a posteriori estimada por el MLP y $p(q_t)$ es la probabilidad a priori de las clases de los datos de entrenamiento. $p(o_t)$ se puede omitir del cálculo de la verosimilitud de emisión de estados ya que es común para todas las clases.

Una extensión del enfoque híbrido de HMM/MLP sustituye la relación directa entre estados de HMM y salidas del MLP por una distribución multinomial en cada uno de los estados de HMM. El denominado KL-HMM ha sido descrito por ejemplo por Guillermo Aradilla, Jithendra Vepa y Hervé Bourlard, “An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features”, en *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP), 2007.

El KL-HMM minimiza la divergencia de Kullback-Leibler, o divergencia de KL, entre vectores de características de entrenamiento y estados multinomiales de HMM que generan vectores de observación acústica. Para distribuciones de probabilidad P y Q de una variable aleatoria discreta, la divergencia KL de Q con respecto a P se define como

$$D_{KL} = \sum_i P(i) log \frac{P(i)}{Q(i)}$$

Normalmente, los sistemas de ASR basados en HMM comprenden los siguientes componentes:

- Modelos de HMM: describen la relación entre unidades acústicas (normalmente fonos) y características acústicas.
- Diccionario de pronunciación: establece correspondencias de secuencias de unidades acústicas con palabras.
- Gramática (opcional): describe la sintaxis del sistema.

Por oposición a los sistemas con plantillas, los enfoques basados en estadísticas pueden incorporar probabilidades en el diccionario y la gramática de una manera regular.

Según la publicación de la técnica anterior, de G. Aradilla et al. “Posterior Features Applied to Speech Recognition Tasks with User-Defined Vocabulary”, en *Proceedings of ICASSP* 2009, se conoce un sistema de reconocimiento de voz que usa características a posteriori obtenidas a partir de un MLP para formar plantillas y pronunciaciones de prueba.

Se conoce además, de acuerdo con la publicación de O. Watanuki “Template Selection Method for Speaker-Independent Word Recognition”, IBM, TDB 08-86, 1 de agosto de 1986, un método de selección de plantillas que usa una matriz de distancias y que fusiona pronunciaciones pertenecientes a dos o más categorías con cierta confusión.

Además, de acuerdo con la publicación de J. S. Sánchez et al. “Prototype selection for the nearest neighbour rule through proximity graphs”, *Pattern Recognition Letters* 18 (1997), se conoce el uso de grafos de Gabriel y de Vecindad Relativa para seleccionar un subconjunto adecuado de prototipos para la regla de Vecino Más Próximo aplicando una edición-condensación de vecindad de grafos.

Se requieren un sistema nuevo y un método nuevo para introducir y reconocer un comando de voz de usuario, los cuales sean más robustos ante el entorno y el hablante y más sencillos de implementar que sistemas de la técnica anterior.

Se requieren también un sistema nuevo y un método nuevo en los que el entrenamiento de modelos consuma menos tiempo para los usuarios.

Son también necesarios un sistema nuevo y un método nuevo que puedan trabajar con múltiples idiomas, pronunciaciones y acentos.

Son necesarios también un sistema nuevo y un método nuevo que saquen provecho de la retroalimentación del usuario.

Breve sumario de la invención

Según un aspecto de la invención, estos objetivos se alcanzan por medio de un sistema automático de reconocimiento de voz para reconocer un comando de voz de usuario en un entorno con ruido, que comprende

- medios de concordancia para hacer concordar elementos recuperados a partir de unidades de habla que forman dicho comando, con plantillas de una biblioteca de plantillas;

- medios de procesado que incluyen un Perceptrón Multicapa para calcular plantillas a posteriori ($P(O^{plantilla(q)})$) almacenadas como dichas plantillas en dicha biblioteca de plantillas;

- medios para recuperar vectores a posteriori ($P(O^{prueba(q)})$) a partir de dichas unidades de habla, usándose dichos vectores a posteriori en calidad de dichos elementos;

- medios de cálculo para seleccionar automáticamente plantillas a posteriori almacenadas en dicha biblioteca de plantillas;

- caracterizado por que dichos medios de cálculo están adaptados para usar un enfoque de grafos, tal como un enfoque de Gabriel o el enfoque de vecinos relativos para la selección de dichas plantillas a posteriori a partir de plantillas de entrenamiento, comprendiendo dicha selección:

- determinación de vecinos de Gabriel/relativos mediante el cálculo de una matriz de distancias entre la totalidad de dichas plantillas de entrenamiento,

- visitar cada plantilla de entrenamiento,

- marcar una plantilla de entrenamiento si la totalidad de sus vecinos es de la misma clase que la plantilla de entrenamiento actual,

- eliminar todas las plantillas de entrenamiento marcadas.

Un vector a posteriori es un vector de probabilidades a posteriori producidas por el MLP. Se produce un vector a posteriori para cada vector de características acústicas que se genera (por ejemplo, MFCCs). En una forma de realización, se genera un vector de características acústicas cada 10 ms. Las probabilidades a posteriori se corresponden con clases de fonos. En el caso del idioma inglés, en una forma de realización hay 42 fonos: en este caso, el vector a posteriori tendría una dimensionalidad 42; la suma de probabilidades a posteriori en el vector sería 1,0; cada probabilidad tendría que ser >= 0 y <= 1.

Una plantilla a posteriori es una colección de vectores a posteriori que describen una única instancia (o plantilla) de un término a reconocer por el sistema. Por ejemplo, si se grabase la plantilla a posteriori para el nombre “1” con una duración de 0,25 s (= 250 ms = 25*10 ms), entonces la plantilla a posteriori comprendería 25 vectores a posteriori (cada uno de dimensión 42 a partir del ejemplo previo), de manera que la plantilla a posteriori tendría una dimensión 25*42.

Un vector a posteriori y una plantilla a posteriori se consideran como características a introducir en el clasificador basado en plantillas (por ejemplo, en lugar de MFCC). Las características son una representación de la señal original (en este caso, voz) que pretenden extraer solamente información destacada para la tarea en cuestión. En el caso del reconocimiento de voz, únicamente se está interesado en lo que se ha dicho y no en quién lo dijo, de manera que la “extracción de características” tiene como objetivo enfatizar la parte de la señal referente a lo que se ha dicho al mismo tiempo que se desenfatizan otros aspectos. En teoría, los vectores a posteriori son la característica perfecta ya que contienen exactamente la información deseada (probabilidades a posteriori de clases de fonos) aunque no contienen ninguna otra información. Desafortunadamente, las probabilidades a posteriori estimadas por el MLP no son perfectas, siendo necesario por ello un análisis adicional para obtener un buen rendimiento (por ejemplo, el uso descrito de plantillas a posteriori).

Los medios para recuperar vectores a posteriori de unidades de habla pueden incluir el MLP, aunque se puede usar otro estimador estadístico para obtener estos vectores a posteriori.

Según otro aspecto de la invención, estos objetivos se logran por medio de un método de reconocimiento automático de la voz para reconocer un comando pronunciado por un usuario en un entorno con ruido, comprendiendo dicho método:

- hacer concordar elementos recuperados a partir de unidades de habla que forman dicho comando, con plantillas de una biblioteca de plantillas;

- determinar una secuencia de plantillas que minimiza la distancia entre dichos elementos y dichas plantillas;

- en donde dichas plantillas son plantillas a posteriori ($P(O^{plantilla(q)})$);

- y dichos elementos recuperados a partir de unidades de habla son vectores a posteriori ($P(O^{prueba(q)})$);

- generándose dichas plantillas a posteriori y dichos vectores a posteriori con por lo menos un Perceptrón MultiCapa;

- caracterizado por que incluye una etapa para seleccionar dichas plantillas a posteriori a partir de plantillas de entrenamiento usando un enfoque de grafos, tal como el enfoque de Gabriel, o el enfoque de vecinos relativos, comprendiendo dicha etapa:

- determinación de vecinos de Gabriel/relativos mediante el cálculo de una matriz de distancias entre la totalidad de dichas plantillas de entrenamiento,

- visitar cada plantilla de entrenamiento,

- marcar una plantilla de entrenamiento si la totalidad de sus vecinos es de la misma clase que la plantilla de entrenamiento actual,

- eliminar todas las plantillas de entrenamiento marcadas.

Este método se basa por lo tanto en un nuevo ASR (Reconocimiento Automático de la Voz), el cual combina aspectos del ASR convencional por concordancia de plantillas, con aspectos correspondientes de un ASR híbrido de HMM/MLP. La independencia con respecto a la pronunciación y el idioma viene dada por el ASR con concordancia de plantillas (puesto que, normalmente, cada usuario aporta sus propias plantillas las cuales contienen su propia pronunciación específica) y la robustez frente al entorno y el hablante la confiere el uso del MLP del ASR híbrido de HMM/MLP. Combinando estos dos sistemas es posible disponer de ventajas de ambos. Aún cuando el MLP se entrena normalmente de manera que sea específico del idioma, el mismo puede seguir funcionando bien para otros idiomas diferentes a aquel para el cual se entrenó el MLP.

Vectores a posteriori generados por un MLP se incorporan como la base para plantillas a posteriori, donde las entradas al MLP son vectores de características acústicas convencionales. Como consecuencia, el reconocimiento de plantillas a posteriori es:

$$q^* = \arg\min_q DTW\left(p\left(O^{plantilla(q)}\right), p(O^{prueba})\right)$$

Las ventajas ofrecidas por este sistema son un aumento de la robustez a una variabilidad no deseada junto con una sensibilidad mejorada a aspectos deseados del habla. Más específicamente, el entrenamiento del MLP se centra en una discriminación aumentada entre clases de fonos e ignora otros aspectos de la señal acústica. Posteriormente, estos aspectos son recuperados por el algoritmo de concordancia de plantillas.

Las plantillas acústicas del ASR convencional por concordancia de plantillas son sustituidas en este sistema por plantillas a posteriori. Por otra parte, la medida de distancias locales se modifica con el fin de que resulte más apropiada para características a posteriori. Este sistema comprende, además del sistema de ASR normal basado en plantillas, un perceptrón multicapa (MLP) entrenado, para el cálculo de características a posteriori. El algoritmo de DTW, la biblioteca de plantillas y la totalidad del resto de aspectos permanecen iguales que en el sistema de ASR convencional basado en plantillas.

De forma ventajosa, el MLP se entrenaría para reconocer fonemas en un idioma específico, por ejemplo se puede entrenar para reconocer 50 fonemas del inglés, o entre 30 y 200 fonemas del inglés. En una forma de realización, es posible entrenar múltiples MLPs para idiomas diferentes, por ejemplo un MLP para reconocer fonemas del francés y un MLP para reconocer fonemas del inglés. En una forma de realización, un único MLP multilingüe se puede entrenar para reconocer fonemas en varios idiomas.

De forma ventajosa, la invención permite una rápida adscripción del usuario. De hecho, según la invención, es posible seleccionar las plantillas a posteriori a partir de plantillas de entrenamiento adquiridas durante la sesión de adscripción del usuario, con el fin de acelerar el funcionamiento del sistema, ya que se seleccionan únicamente las plantillas más útiles, para el reconocimiento.

Para medir la similitud entre vectores a posteriori, el sistema usa parámetros basados en la teoría de la información o medidas probabilísticas, por ejemplo la divergencia de Kullback-Leibler.

En una forma de realización, la biblioteca de plantillas a posteriori es una biblioteca de plantillas preexistente que ha sido grabada por otro usuario.

En otra forma de realización, la biblioteca de plantillas a posteriori se crea de manera ventajosa a partir de un diccionario de pronunciación.

Por otra parte, las estrategias de aprendizaje en línea, de acuerdo con la invención, mejoran el rendimiento del sistema mediante el uso de datos recopilados durante la actividad normal del sistema. Por medio de este aprendizaje en línea es posible, por ejemplo, añadir nuevas plantillas a posteriori extrayéndolas directamente de la pronunciación, es decir, de la unidad de habla completa en el lenguaje en el idioma hablado, y también evitar el uso de una plantilla a posteriori inapropiada, es decir, una plantilla redundante o extraída de forma errónea. También es posible asignar prioridades a las plantillas a posteriori.

De acuerdo con la invención, la biblioteca de plantillas a posteriori se puede gestionar de manera automática. A posteriori se puede gestionar de manera automática. En una forma de realización, se puede usar la retroalimentación del usuario.

De forma ventajosa, el ruido del entorno se puede detectar a largo de direcciones diferentes a la dirección de la voz del usuario y se puede aprovechar con el fin de mejorar la relación de señal/ruido usando una *Degenerate Unmixing Estimation Technique (DUET*) modificada.

En una forma de realización, el sistema comprende además una gramática. En una forma de realización, esta gramática comprende una red estática de nodos que definen palabras en el léxico y bordes que definen transiciones permisibles entre palabras.

Según un aspecto de la invención, el decodificador de DTW incluye parámetros para modificar la precisión de cálculo y/o la velocidad.

Breve descripción de los dibujos

La invención se entenderá mejor con la ayuda de la descripción de una forma de realización dada a título de ejemplo e ilustrada por las figuras, en las cuales:

la Fig. 1 muestra una vista de un sistema de ASR de plantillas.

La Fig. 2 muestra un ejemplo del procedimiento de búsqueda de DTW en un sistema de ASR de plantillas.

La Fig. 3 es un diagrama de flujo del sistema de acuerdo con la invención.

La Fig. 4 muestra una configuración típica de transición de plantillas.

La Fig. 5 muestra una configuración de transición de plantillas para plantillas de arranque.

La Fig. 6 muestra esquemáticamente un sistema para capturar y transmitir pedidos por voz en un restaurante.

Descripción detallada de posibles formas de realización de la invención

La invención se basa en un nuevo ASR (sistema de Reconocimiento Automático de Voz), que combina aspectos del ASR convencional por plantillas y el ASR híbrido de HMM/MLP. Vectores a posteriori generados por un MLP (Perceptrón Multicapa) se incorporan como base para plantillas a posteriori. Las entradas al MLP son vectores de características acústicas convencionales.

Este sistema presenta un aumento de la robustez frente a una variabilidad no deseada de la pronunciación, del idioma, del acento y del entorno en comparación con el sistema de ASR por plantillas. Por otra parte, se mejora la sensibilidad o la precisión.

Tal como se ha mencionado, la independencia con respecto a la pronunciación y el idioma viene dada por el enfoque convencional basado en plantillas (ya que cada usuario proporciona normalmente sus propias plantillas, las cuales contienen su propia pronunciación específica), y la robustez al entorno y al hablante la aporta el uso del MLP (de manera similar al sistema de ASR convencional). Combinando estos dos sistemas, el sistema nuevo obtiene las ventajas de ambos. Aún cuando el MLP se entrena normalmente para que sea específico del idioma (aunque esto no es así necesariamente), el mismo puede seguir funcionando adecuadamente para otros idiomas diferentes a aquel para el cual se entrenó el MLP.

El sistema nuevo permite el desarrollo de un sistema de reconocimiento de voz continuo y de vocabulario limitado y controlado, que es robusto frente a entornos con condiciones altamente exigentes, tales como entornos con ruido. El sistema es también robusto frente a la variabilidad de los hablantes gracias a las propiedades del MLP.

En la Figura 6 se ilustra un ejemplo de entorno en el cual se pueden usar el método y el sistema. En este escenario, un camarero 2 que está en un bar o restaurante coge pedidos de clientes 3 que están sentados en una mesa. El camarero repite cada pedido y los pronuncia en un micrófono de un dispositivo móvil 1. En esta forma de realización, la señal de voz grabada se procesa localmente, por ejemplo por medio del procesador del dispositivo móvil 1 ó preferentemente a través de medios de procesado dedicados, con el fin de mejorar la relación de señal/ruido. Este procesado también podría ser realizado por un ordenador o servidor remoto en otra forma de realización, aunque probablemente esto puede introducir un retardo. A continuación, la señal de voz procesada se transmite inalámbricamente por vía aérea a un punto de acceso 7, utilizando un protocolo convencional de comunicaciones inalámbricas, tal como el 802.11, Bluetooth, etcétera. En otra forma de realización, no mostrada, la señal de voz se transmite usando cables. El punto de acceso 7 pertenece a una red de área local 8 (LAN), a la cual están conectados otros diversos equipos, tales como un ordenador personal 5, un servidor 6 con una base de datos 60 para almacenar una biblioteca de plantillas, etcétera. La señal de voz recibida desde el punto de acceso 7 se convierte en pedidos de texto por parte del servidor 6, el cual ejecuta un algoritmo de reconocimiento de voz y un software de gestión para bares o restaurantes. A continuación, los pedidos de texto se visualizan y se procesan para entregar y facturar el pedido solicitado a los clientes 3.

En otra forma de realización, el micrófono para grabar señales de voz no está conectado a un dispositivo móvil 1 sino directamente a un ordenador personal 5: en tal caso, la señal de voz no se transmite inalámbricamente.

El algoritmo de reconocimiento de voz podría ser ejecutado por el dispositivo móvil 1 en caso de que este dispositivo disponga de suficiente poder de procesado; no obstante, esto puede hacer que una actualización de los modelos de voz y de idioma (tales como la lista de comandos a reconocer, y la gramática asociada) resulte más dificultosa. No obstante, esto requiere dispositivos 1 con un mayor poder de procesado, y una sincronización más dificultosa de los modelos dependientes del hablante en caso de que un usuario utilice varios dispositivos diferentes.

El sistema según la invención está compuesto por varios componentes lógicos los cuales pueden estar integrados en el dispositivo 1, por ejemplo en forma de componentes de software y/o hardware, y/o en forma de una extensión del dispositivo, tal como una extensión con un procesador DSP, y/o en forma de una aplicación de software ejecutada por el servidor 6.

Las operaciones en el lado del servidor se llevan a cabo utilizando una etapa frontal de ASR particular para efectuar procesado de señales de voz para un decodificador de DTW.

A continuación se describirá con la Figura 3 un ejemplo de método de acuerdo con la invención. La etapa 30 de la Figura representa la adquisición de audio: comandos de voz pronunciados por el camarero 2 son detectados por un micrófono del dispositivo 1, por ejemplo una matriz de micrófonos, y son procesadas preferentemente utilizando una tecnología de conformación de haz. Esta adquisición de audio se lleva a cabo de manera preferente con una expansión de DSP del dispositivo 1.

La etapa 32 representa la extracción y cálculo de características acústicas: de la señal de voz se extraen coeficientes cepstrales en frecuencia Mel (MFCCs) convencionales. A continuación, estos coeficientes se normalizan con respecto a la media y la varianza. Esta extracción y este cálculo se pueden llevar a cabo en el DSP 1 y/o en el lado de servidor 6. Durante esta etapa, también se lleva a cabo el cálculo de energía de la señal de voz en el DSP 1 y/o en el lado de servidor 6.

La detección de actividad vocal (VAD) se puede llevar a cabo en el lado de servidor 6, posiblemente en dos fases. La primera fase 34 se lleva a cabo utilizando los cálculos de energía de la etapa 32. Esta primera fase 34 es crítica en el cálculo de estimaciones de la media y de la varianza.

El post-procesado de características, por ejemplo normalización de media y varianza, se lleva a cabo en el lado de servidor 6 en la etapa 36.

En una forma de realización, el post-procesado realizado en la etapa 36 comprende también cálculos para mejorar la relación de señal/ruido. En una forma de realización, los comandos de voz pueden ser detectados por múltiples matrices de micrófonos. Por ejemplo se pueden usar cuatro matrices perpendiculares de micrófonos, con el fin de capturar sonidos provenientes desde cuatro direcciones principales: frontal (la dirección de la voz), posterior, izquierda y derecha. De esta manera, se detecta el ruido de fondo de diferentes direcciones. Este aspecto es especialmente importante, por ejemplo, en un contexto de restaurante, donde un camarero puede coger pedidos de clientes que están sentados en una mesa mientras los clientes que están sentados en otra mesa cercana están hablando. La etapa de post-procesado 36 depende de la disposición espacial de múltiples matrices de micrófonos e implica una comparación, en el dominio de la frecuencia, de las cuatro señales entregadas por la matriz de micrófonos (frontal, posterior, izquierda, derecha) usando filtros adaptativos basados en una DUET (*Degenerate Unmixing Estimation Technique)* modificada. Para cada uno de los cuatro canales, estos filtros adaptativos permiten reducir la influencia del ruido en el canal frontal, mediante sustracción espectral de las señales de los otros tres canales que básicamente están detectando ruido.

La etapa 38 representa la estimación del MLP (Perceptrón Multicapa), que da salida a vectores a posteriori basándose en las características acústicas procesadas de entrada. Se da salida a un vector a posteriori para cada vector de característica acústica introducido. La estimación a posteriori y de entrenamiento del MLP se lleva a cabo utilizando una biblioteca de aprendizaje de máquinas, escrita por ejemplo en C++ o C#. La estimación a posterior del MLP está integrada en la etapa frontal del ASR en el lado de servidor 6.

La segunda fase de VAD 42 se lleva a cabo preferentemente en el lado de servidor 6 después de la estimación del MLP, y simplemente fija umbrales para la probabilidad de silencio según se calcula en el MLP. Esta segunda fase es crítica ya que el silencio normalmente no se modela de manera explícita en las plantillas de un sistema de ASR.

La detección vocal se realiza preferentemente por análisis de la señal de potencia. Cuando no hay voz, la señal residual se cancela con el fin de eliminar todo el ruido entre periodos de voz. En una forma de realización, la función del detector de actividad vocal también puede ser ejecutada o auxiliada, por ejemplo, por un botón pulsador (físico o de pantalla táctil) del dispositivo 1, similar al botón pulsador de un dispositivo de *walky-talky*, o a través de otro medio (activación de una señal con los pies o los ojos del camarero, la posición del dispositivo móvil, etcétera). Este botón o en general estos medios son seleccionados por el usuario 2 con el fin de activar el dispositivo 1. Si el usuario no selecciona estos medios, aun cuando el mismo hable al dispositivo 1, su voz no es capturada. En una forma de realización, el micrófono está siempre grabando, lo cual permite entonces configurar el sistema para seleccionar exactamente la porción de grabación a enviar para su procesado. Es entonces posible, por ejemplo, seleccionar la posibilidad de enviar el comando de voz, por ejemplo, ½ segundo antes de seleccionar los medios, y, por ejemplo, ½ segundo después de que estos medios se hayan liberado o seleccionado nuevamente. De esta manera, en caso de que el usuario 2 pulse o libere los medios un poco demasiado pronto o demasiado tarde, la porción de grabación que interesa se puede mantener sin pérdidas. El botón pulsador o, en general, los medios, también se pueden usar para seleccionar la gramática que utilizará el ASR (y por lo tanto para reducir el número de palabras a reconocer en un contexto dado) con el fin de aumentar la precisión y reducir el tiempo de respuesta del ASR.

La etapa 44 representa la recuperación de plantillas a posteriori, a partir de la biblioteca de plantillas a posteriori. Cada elemento de la biblioteca de plantillas a posteriori comprende una secuencia de vectores a posteriori y un índice o término de plantillas asociado. Para generar la biblioteca de plantillas a posteriori, se efectúan grabaciones de palabras/expresiones de vocabulario durante una sesión de adscripción y a continuación las mismas se hacen pasar a través de las etapas 30 a 42 para calcular las plantillas a posteriori. La biblioteca de plantillas a posteriori se almacena preferentemente en un servidor central 6 para su uso y gestión comunes desde una pluralidad de dispositivos 1 y usuarios 2.

La flecha discontinua entre las etapas 42 y 44 indica que esta conexión se lleva a cabo durante la generación de esta biblioteca de plantillas a posteriori.

La etapa 40 representa la decodificación de reconocimiento continuo de voz con alineamiento dinámico temporal (DTW). La misma se lleva a cabo preferentemente en el lado de servidor 6. El decodificador acepta como entrada la secuencia de vectores a posteriori que serán reconocidos y procesados por la etapa frontal del ASR durante las etapas 30 a 42, la biblioteca de plantillas a posteriori 44, un diccionario 46 y una gramática opcional 48.

El decodificador da salida a la(s) secuencia(s) de palabras reconocidas, a una información de tiempo (inicio/final) y a medidas de confianza. El decodificador incluye, en una forma de realización, varios parámetros para ajustar la precisión con respecto a la velocidad.

El diccionario 46 define la relación entre plantillas a posteriori (grabaciones concretas de palabras/expresiones) y los nodos en la gramática. Es necesario desambiguar entre múltiples usos de plantillas en la gramática.

Opcionalmente, el decodificador puede utilizar una gramática 48. En una forma de realización, el sistema usa una red estática de nodos que definen palabras en el léxico y bordes que definen transiciones permisibles entre palabras.

El sistema ilustrado en la Figura 3 tiene la capacidad de reconocer comandos de un usuario usando una biblioteca de plantillas a posteriori grabadas durante una sesión de adscripción. El entrenamiento de la biblioteca de plantillas a posteriori se lleva a cabo inicialmente durante una sesión de adscripción en la cual se requiere que el usuario pronuncie las palabras deseadas, por ejemplo las palabras contenidas en el menú del día, por lo menos una vez. Para mejorar la robustez del sistema y tener en cuenta diferentes pronunciaciones o diferentes condiciones de grabación, el usuario pronuncia las palabras deseadas más de una vez. Durante el uso normal del sistema se pueden llevar a cabo otro entrenamiento y una adaptación dependiente del hablante, de las plantillas a posteriori, usando por ejemplo retroalimentaciones y una corrección de entradas de usuarios 2 con medios hápticos en su dispositivo cuando una unidad de habla o una palabra se ha reconocido de manera errónea.

Cuando la lista de términos de los comandos resulta grande, la sesión de adscripción puede llegar a consumir mucho tiempo para el usuario, particularmente si se graban múltiples plantillas por cada término. Por otra parte, esto

da como resultado un funcionamiento más lento del sistema ya que deben considerarse más plantillas durante la decodificación. Por lo tanto, según otro aspecto de la invención, se proporciona un mecanismo para simplificar el proceso de adscripción.

De acuerdo con un aspecto de la invención, es posible seleccionar, de entre un conjunto de plantillas proporcionadas por el usuario, la más útil para el reconocimiento. Esto permite el uso de un número mínimo de plantillas, mejorando el tiempo de respuesta. También es posible determinar si la biblioteca de plantillas proporciona una variabilidad suficiente de entradas posibles del usuario, maximizando los rendimientos con respecto al tiempo consumido durante la sesión de adscripción. Con este fin, se usan enfoques de los vecinos más próximos condensados/editados, para reducir al máximo el número de ejemplos de entrenamiento sin sacrificar el rendimiento del sistema.

El enfoque más sencillo de vecinos más próximos condensados (regla de Hart) simplemente encuentra en su conjunto de datos m que minimiza el error en los datos restantes n – m. Una desventaja de esta regla es que requiere una selección arbitraria de m, lo cual puede dar como resultado demasiados o demasiado pocos ejemplos de entrenamiento.

Por este motivo, la invención se aprovecha de manera ventajosa de enfoques de grafos con el fin de condensar o editar el conjunto de entrenamiento. Estos enfoques tienen como finalidad construir/obtener una aproximación del límite de decisión del clasificador y eliminar muestras de entrenamiento que no contribuyen a este límite. Muchos de estos enfoques no son adecuados para espacios de alta dimensionalidad, y, en el caso del ASR de DTW, son inviables. Según un aspecto de la invención, dichos grafos se consideran de forma ventajosa en términos del parámetro de distancia, que para el DTW es simplemente:

$$D\left(plantilla_q, plantilla_p\right) = DTW\left(p\left(O^{plantilla_q}\right), p\left(O^{plantilla_p}\right)\right)$$

Uno de los enfoques que se puede utilizar es el de Gabriel. En este enfoque, se determina qué puntos A y B son vecinos de tal manera que ningún otro punto X resida en su esfera diametral:

$$D^2(A,X) + D^2(B,X) > D^2(A,B) \quad \forall X \neq A,B$$

Otro posible enfoque es el de grafos de vecinos relativos, donde el conjunto de vecinos relativos A, B debe cumplir la siguiente relación:

$$D(A,B) < max\{D(A,X)D(B,X)\} \quad \forall X \neq A,B$$

A continuación, el conjunto de entrenamiento se puede condensar o editar siguiendo estas etapas:

- determinación de los vecinos de Gabriel/relativos mediante el cálculo de una matriz de distancias entre la totalidad de plantillas de entrenamiento;
- visitar cada plantilla de entrenamiento;
- marcar una plantilla de entrenamiento si la totalidad de sus vecinos es de la misma clase que la plantilla de entrenamiento actual;
- eliminar todas las plantillas de entrenamiento marcadas.

Las restantes plantillas de entrenamiento constituyen el conjunto de entrenamiento editado deseado. Durante la determinación de los vecinos, se puede aplicar heurística para reducir el cálculo necesario. El uso de dichas reglas de edición utilizando la distancia de DTW es solamente una aproximación ya que las relaciones geométricas que determinan vecinos se basan en la medida de la distancia que obedece a la desigualdad triangular. La distancia de DTW no obedece a esta desigualdad, aunque en la práctica funciona bien.

La optimización de la biblioteca de plantillas a posteriori puede ser un proceso costoso desde el punto de vista computacional puesto que la distancia de DTW se debe calcular dos veces para cada par de plantillas a posteriori de la biblioteca de plantillas a posteriori. Es decir, si hay *N* plantillas a posteriori, esto implica $N^2 - N$ cálculos de distancia de DTW. Para reducir el tiempo de cálculo, es posible obtener una aproximación de la distancia de DTW por medio de una interpolación lineal de las plantillas a posteriori. Es decir, si $N = 2$, es decir, hay dos plantillas a posteriori, por ejemplo la $plantilla_q$ y la $plantilla_p$, respectivamente de longitud L_q y L_p, se puede aplicar una interpolación lineal *interp()* a estas dos plantillas, tal que

$interp(plantilla_p, L_q)$

produzca una versión interpolada de la *plantilla*$_p$ de longitud L_q. En este caso, la distancia aproximada de DTW resulta ser:

$$DTW_{aprox}(plantilla_p, plantilla_q) = ||plantilla_p - interp(plantilla_p, L_q)||$$

y la distancia usada para editar la biblioteca de plantillas a posteriori resulta ser:

$$||plantilla_p - interp(plantilla_q, L_p)|| + ||interp(plantilla_p, L_q) - plantilla_q||$$

donde || . || significa la norma L2 o distancia euclídea.

En otra forma de realización, para acelerar la adscripción del usuario, este último puede utilizar una biblioteca de plantillas preexistente que ha sido grabada por otro usuario. El uso de MLP en el sistema aporta al algoritmo de concordancia un grado de robustez de manera que se ve afectado de forma menos significativa por la varianza de características del hablante, particularmente aquellas relacionadas con características de la voz.

Dado un usuario nuevo del sistema, se puede encontrar un usuario ya adscrito que presente la mayor coincidencia con el usuario nuevo. Más específicamente, el usuario nuevo debe grabar una serie limitada de pronunciaciones de validación y se halla la biblioteca de plantillas a posteriori coincidente más parecida, una vez más por medio del algoritmo de DTW. Si la distancia de DTW promedio entre las pronunciaciones de validación y la biblioteca de plantillas a posteriori coincidente más parecida es suficientemente baja, el usuario nuevo puede usar esta biblioteca de plantillas y seguir alcanzando el rendimiento deseado.

En otra forma de realización, un usuario nuevo puede adscribirse en el sistema sin necesidad de grabar plantillas y sin necesidad de usar plantillas de otro usuario. Esto es posible usando un diccionario de pronunciación, el cual establece correspondencias de secuencias de unidades acústicas en una forma canónica, con palabras, con el fin de construir plantillas a posteriori. Las plantillas a posteriori construidas se denominan “plantillas de arranque”.

Las unidades acústicas usadas para generar las plantillas de arranque se pueden obtener a partir de un HMM que usa el mismo parámetro que las plantillas regulares. Para el parámetro de divergencia de KL antes descrito, este es el KL-HMM en el cual cada estado se representa por una distribución multinomial. La distribución multinomial tiene la misma forma que los vectores a posteriori usados en la biblioteca de plantillas y se calcula como la media geométrica de los vectores a posteriori asociados a ese fonema en los datos de entrenamiento. Esto requiere el cálculo de las distribuciones multinomiales para cada salida de clase de fonemas del MLP.

Las plantillas de arranque permiten una configuración de las transiciones permitidas entre estados de la plantilla lo cual proporciona duraciones de estado más flexibles que las plantillas convencionales, tal como se muestra en las Figuras 4 y 5. La Figura 4 ilustra la transición permisible en una “plantilla regular” en la cual la plantilla se representa por medio de una máquina de estados finitos (FSM) con estados $q_1, q_2, ..., q_N$ que representan las N tramas de la plantilla. Por lo tanto, cada estado puede tener una duración de cero (es decir, se omite), uno (sin alineamiento) o dos (estiramiento de la duración). Por contraposición, la plantilla de arranque ilustrada en la Figura 5 en la que cada estado de la FSM se corresponde con estados del KL-HMM puede tener una duración mínima de uno y una duración máxima infinita (debido a la auto-transición).

Cuando se usan juntas plantillas “regulares” y plantillas de “arranque”, para hacer frente a inconsistencias entre puntuaciones generadas por las plantillas “regulares” y de “arranque”, en el decodificador se incorporan varias características adicionales. En primer lugar, las plantillas de arranque tienen una tendencia a producir muchas inserciones ya que tienen una duración breve (típicamente tres estados por cada KL-HMM de un fonema) y proporcionan también una coincidencia con las pronunciaciones de prueba, que es más deficiente que las plantillas regulares. Como compensación, en las plantillas de arranque se introduce una penalización de inserción y la misma se aplica en todas las transiciones en plantillas de arranque. Normalmente, cuando se calcula la distancia en cada trama se utiliza la siguiente ecuación:

$$Q^* = \arg\min \sum_{t=1}^{T} D_{KL}\left(P(o_t), Q(q_t)\right)$$

donde $P(o_t)$ es el vector a posteriori de la secuencia de prueba en el instante de tiempo t para la observación acústica o_t, y $Q(q_t)$ es el vector a posteriori de la biblioteca de plantillas y q_t es el estado (trama) correspondiente de la biblioteca de plantillas con la cual se compara la trama actual o_t de la secuencia de prueba.

Así, en cada plantilla de arranque nueva se inserta una penalización de inserción:

$$Q^* = \arg\min \sum_{t=1}^{T} D_{\text{inserción}}(q_t) + D_{KL}\left(P(o_t), Q(q_t)\right)$$

donde normalmente $D_{inserción}(q_t) = 0$ a no ser que q_t se corresponda con el primer estado (trama) en una plantilla de arranque, en cuyo caso $D_{inserción}(q_t) > 0$.

Al introducir la penalización de inserción se hace frente al problema de inserciones, pero esto ahora favorece notablemente las plantillas regulares en el caso en el que se usen bibliotecas de plantillas compuestas por plantillas tanto regulares como de arranque. Para equilibrarlo, en las puntuaciones de plantillas de arranque se aplica un factor de escala tal que:

$$Q^* = \arg\min \sum_{t=1}^{T} D_{\text{inserción}}(q_t) + D_{\text{escala}}(q_t) + D_{KL}(P(o_t), Q(q_t))$$

donde, para plantillas regulares, $D_{escala}(q_t) = 0$, y para plantillas de arranque, $D_{escala}(q_t) < 0$. Los valores típicos (diferentes de cero) de $D_{inserción}$ y D_{escala} son respectivamente 25 y -0,85.

Además, se puede hacer que el reconocimiento sea más robusto ante los errores en la detección de actividad vocal usando un KL-HMM para el silencio. En el funcionamiento previamente descrito del sistema, se supone que todo silencio se elimina de la grabación usando detección de actividad vocal, de manera que se estén reconociendo únicamente muestras de voz. En realidad, la VAD puede no conseguir eliminar todo el silencio. En este caso, puede resultar útil incluir una “plantilla de silencio” que se genera a partir del KL-HMM de silencio usado para generar plantillas de arranque. Esta plantilla de silencio se puede insertar opcionalmente en el comienzo y/o final de todas las plantillas que se está intentando reconocer y absorberán cualesquiera tramas de silencio que no hayan sido eliminadas por la VAD, funcionando como un silencio de filtro.

El aprendizaje en línea es el proceso por el cual se puede mejorar el rendimiento de un clasificador utilizando datos recopilados durante su funcionamiento normal. Esto se puede realizar de dos maneras:

- una manera no supervisada, utilizando decisiones tomadas por el clasificador para actualizar el modelo, o

- una manera supervisada, utilizando retroalimentación del usuario para actualizar el clasificador.

El aprendizaje en línea se puede considerar como unos medios para actualizar la biblioteca de plantillas a posteriori. En una forma de realización, el aprendizaje en línea se puede usar para mejorar el rendimiento del sistema, particularmente cuando se usa en combinación con las técnicas propuestas de adscripción rápida antes descritas.

Durante el uso del sistema, el usuario proporciona retroalimentación. El mismo puede comprobar si el reconocimiento fue correcto, y confirmar o corregir el comando reconocido por el servidor 6 y visualizada por el dispositivo 1. Usando esta retroalimentación, se pueden introducir plantillas nuevas en el sistema.

En un escenario en línea, es necesario gestionar la activación o adición de plantillas nuevas a la biblioteca. En tales casos, se pueden aplicar los criterios de condensación/edición descritos previamente. La activación o adición de plantillas nuevas se puede llevar a cabo:

- cuando la biblioteca de plantillas a posteriori no contiene suficientes ejemplos; cuando, por ejemplo, se usan plantillas de arranque o plantillas de otros usuarios, es posible activar o añadir a estas plantillas las correspondientes del usuario en cuestión.

- cuando el comando de voz se reconoce de manera incorrecta. La pronunciación de prueba se puede reconocer erróneamente debido a que el nivel de ruido es demasiado alto o debido a que la biblioteca de plantillas a posteriori no contiene una buena coincidencia para la pronunciación de prueba. En este último caso, se puede activar o añadir una plantilla a la biblioteca con el fin de mejorar el rendimiento futuro.

- cuando se debe mejorar la biblioteca de plantillas a posteriori, incluso si la pronunciación de prueba se graba correctamente. Se pueden activar o añadir plantillas nuevas si la pronunciación de prueba contiene plantillas potenciales que son suficientemente diferentes con respecto a aquellas que están actualmente presentes.

La concordancia de plantillas por medio del DTW proporciona las plantillas reconocidas más información de temporización, es decir, los tiempos de inicio/final de cada una de las plantillas reconocidas. En una forma de realización, los límites de tiempo extraídos se pueden usar para extraer plantillas nuevas.

En otra forma de realización, es posible mantener un seguimiento del uso de plantillas para determinar si se ha añadido una plantilla “inadecuada” a la biblioteca. Más específicamente, una plantilla que graba con una precisión de reconocimiento deficiente se puede eliminar si se ha añadido una plantilla “inadecuada” a la biblioteca. Más específicamente, una plantilla que graba con una precisión de reconocimiento deficiente se puede eliminar ya que es

probable que contenga una grabación errónea o altos niveles de ruido. Se puede determinar un conjunto de elementos heurísticos para gestionar dichas circunstancias.

Según la invención, es posible manipular la biblioteca de plantillas a posteriori para:

- crear un usuario nuevo;

- eliminar un usuario existente;

- adscribir un usuario utilizando métodos de adscripción convencional o de adscripción propuesta;

- activar/desactiva o añadir/sustituir plantillas utilizando métodos en línea propuestos o sesiones de (re)adscripción;

- eliminar plantillas.

En una forma de realización, se graban el origen de la plantilla (convencional/de arranque) en comparación con el hablante (usuario/otro hablante) y también condiciones de grabación (adscripción/aprendizaje en línea). De esta manera, durante la actualización de la biblioteca, se pueden asignar prioridades: por ejemplo, el sistema puede sustituir con una prioridad mayor las plantillas generadas durante una adscripción rápida. Se puede asignar una prioridad diferente también sobre la base de condiciones de grabación (adscripción/aprendizaje en línea).

En otra forma de realización, se utilizan estadísticas de uso de las plantillas con la finalidad de realizar una depuración y un aprendizaje en línea. Para cada plantilla de la biblioteca, y opcionalmente para cada elemento de la gramática, se mantienen estadísticas basadas en el número de veces que se ha reconocido o reconocido erróneamente una plantilla o un elemento.

En una forma de realización, la biblioteca se actualiza usando medios automáticos o semi-automáticos. El sistema puede mantener un seguimiento de rendimientos y sugerir al usuario que grabe otras plantillas durante una sesión de adscripción adicional. Otras plantillas infrautilizadas o plantillas con un rendimiento de reconocimiento deficiente se pueden eliminar automáticamente de la biblioteca.

El sistema puede entonces recopilar datos, por ejemplo, los datos de adscripción de varios usuarios y datos obtenidos durante uso normal del sistema.

El método y el sistema según la invención se pueden usar, por ejemplo, para coger pedidos u comandos en hoteles, restaurantes, bares, negocios minoristas (comestibles, panaderías, carnicerías, etcétera), hospitales y laboratorios. El método y el sistema según la invención se pueden usar, por ejemplo, también para comandos en consolas de videojuego, comandos de equipos, vehículos, robots, edificios, etcétera. Otra posible aplicación de la invención es la gestión de existencias/inventarios. Ninguna de estas posibles aplicaciones debe considerarse en un sentido limitativo.

En una forma de realización, el sistema según la invención comprende

- un dispositivo de usuario 1 adaptado para permitir que un usuario 1 introduzca comandos de voz;

- unos medios de preprocesado en el dispositivo de usuario 1, adaptados para preprocesar los comandos de voz introducidos;

- unos medios de conexión 7, 8, los cuales pueden ser inalámbricos o por cable, para transmitir señales preprocesadas a un servidor central 6 que controla equipos, robots, vehículos o edificios;

- un software de gestión y control 5 para gestionar/controlar equipos, robots, vehículos o edificios de acuerdo con comandos/pedidos introducidos por el usuario 2 a través de los comandos de voz.

En otra forma de realización, el método según la invención comprende además:

- introducir comandos de voz correspondientes a pedidos de bares, restaurantes u hoteles en un dispositivo de usuario 1;

- preprocesar los comandos de voz en un dispositivo de usuario 1;

- transmitir 7, 8 las señales preprocesadas a un servidor 6;

- convertir las señales preprocesadas en pedidos de texto en el servidor 6;

- visualizar los pedidos de texto;

- comunicar los pedidos a software y/o sistemas tales como, aunque sin carácter limitativo, POS (Puntos de Venta), PMS (Sistema de Gestión de Propiedades), Sistema de Gestión de Existencias/Inventarios, ERP (Planificación de Recursos Empresariales) usados por el bar, restaurante u hotel.

El método según la invención puede comprender además las siguientes etapas:

- introducir comandos de voz en un dispositivo de usuario para controlar equipos, robots, vehículos o edificios 1;
- preprocesar los comandos de voz en el dispositivo de usuario 1;
- transmitir 7, 8 señales preprocesadas a un servidor 6;
- convertir las señales preprocesadas en pedidos de texto en el servidor 6;
- visualizar los pedidos de texto;
- comunicar los pedidos a software y/o sistemas que controlan los equipos, robots, vehículos o edificios.

REIVINDICACIONES

1. Sistema automático de reconocimiento de voz para reconocer un comando de voz de usuario (2) en un entorno con ruido, que comprende:

- unos medios de concordancia para hacer concordar unos elementos recuperados a partir de unas unidades de habla que forman dicho comando, con unas plantillas de una biblioteca de plantillas (44);

- unos medios de procesado (32, 36, 38) que incluyen un Perceptrón Multicapa (38) para calcular plantillas a posteriori $P(O^{plantilla(q)})$ almacenadas como dichas plantillas en dicha biblioteca de plantillas (44);

- unos medios para recuperar unos vectores a posteriori $P(O^{prueba(q)})$ a partir de dichas unidades de habla, siendo dichos vectores a posteriori usados como dichos elementos;

- unos medios de cálculo para seleccionar automáticamente unas plantillas a posteriori almacenadas en dicha biblioteca de plantillas (44);

- caracterizado por que dichos medios de cálculo están adaptados para usar un enfoque de grafos, tal como el enfoque de Gabriel o el enfoque de vecinos relativos para la selección de dichas plantillas a posteriori a partir de unas plantillas de entrenamiento, comprendiendo dicha selección:

- determinar los vecinos de Gabriel/relativos mediante el cálculo de una matriz de distancias entre la totalidad de dichas plantillas de entrenamiento,

- visitar cada plantilla de entrenamiento,

- marcar una plantilla de entrenamiento si la totalidad de sus vecinos es de la misma clase que la plantilla de entrenamiento actual,

- eliminar todas las plantillas de entrenamiento marcadas.

2. Sistema según la reivindicación 1, que comprende asimismo:

un decodificador de DTW (40) para hacer concordar vectores a posteriori con plantillas a posteriori.

3. Sistema según la reivindicación 2, que comprende asimismo

- un detector de actividad vocal (34, 42); y

- un diccionario (46).

4. Sistema según una de las reivindicaciones 1 a 3, en el que dicho Perceptrón Multicapa (38) es multilingüe.

5. Sistema según una de las reivindicaciones 1 a 3, que comprende por lo menos dos Perceptrones Multicapa (38), siendo cada uno de dichos Perceptrones Multicapa (38) usado para un idioma específico.

6. Sistema según una de las reivindicaciones 1 a 5, en el que dicha biblioteca de plantillas (44) es una biblioteca de plantillas preexistente generada a partir de unas plantillas de entrenamiento pronunciadas por otro usuario.

7. Sistema según una de las reivindicaciones 1 a 6, que comprende unos medios para crear dicha biblioteca de plantillas (44) a partir de un diccionario de pronunciación.

8. Sistema según la reivindicación 7, en el que dichos medios comprenden un parámetro de divergencia de KL.

9. Sistema según una de las reivindicaciones 1 a 8, que comprende unos medios para adaptar automáticamente dicha biblioteca de plantillas (44), incluyendo la adaptación la activación/desactivación y/o adición y/o eliminación y/o sustitución de dichas plantillas a posteriori.

10. Sistema según la reivindicación 9, en el que dicha adaptación usa una retroalimentación de entrada de dicho usuario (2) en un dispositivo de usuario (1).

11. Sistema según una de las reivindicaciones 1 a 10, que comprende una gramática (48).

12. Sistema según una de las reivindicaciones 1 a 11, que comprende unos medios de detector de actividad vocal que pueden ser seleccionados y deseleccionados por dicho usuario (2).

13. Sistema según la reivindicación 12, en el que dicha gramática (48) es seleccionada por medio de dichos medios de detector de actividad vocal.

14. Sistema según una de las reivindicaciones 2 a 13, en el que dicho decodificador de DTW (40) incorpora una penalización de inserción, un factor de escala y un silencio de filtro.

15. Sistema según una de las reivindicaciones 10 a 14, que comprende:

- dicho dispositivo de usuario (1) adaptado para permitir que un usuario (2) introduzca dichos comandos de voz;
- unos medios de preprocesado en dicho dispositivo de usuario (1), adaptados para preprocesar dichos comandos de voz introducidos;
- unos medios de conexión (7, 8) para transmitir unas señales preprocesadas a un servidor central (6) en un bar, restaurante u hotel;
- software de gestión de restaurantes, bares u hoteles (5) para gestionar pedidos de bares, restaurantes u hoteles, introducidos por dicho usuario (2) a través de dichos comandos de voz.

16. Método de reconocimiento automático de la voz para reconocer un comando de voz pronunciado por un usuario (2) en un entorno con ruido, comprendiendo dicho método:

- hacer concordar unos elementos recuperados a partir de unidades de habla que forman dicho comando, con unas plantillas de una biblioteca de plantillas (44);
- determinar una secuencia de plantillas que minimiza la distancia entre dichos elementos y dichas plantillas;
- siendo dichas plantillas unas plantillas a posteriori $P(O^{plantilla(q)})$ y siendo dichos elementos recuperados a partir de unidades de habla unos vectores a posteriori $P(O^{prueba(q)})$;
- siendo dichas plantillas a posteriori y dichos vectores a posteriori generados con por lo menos un Perceptrón MultiCapa (38);
- caracterizado por que incluye una etapa para seleccionar dichas plantillas a posteriori a partir de plantillas de entrenamiento usando un enfoque de grafos, tal como el enfoque de Gabriel, o el enfoque de vecinos relativos, comprendiendo dicha etapa:
  - determinar los vecinos de Gabriel/relativos mediante el cálculo de una matriz de distancias entre la totalidad de dichas plantillas de entrenamiento,
  - visitar cada plantilla de entrenamiento,
  - marcar una plantilla de entrenamiento si la totalidad de sus vecinos es de la misma clase que la plantilla de entrenamiento actual,
- eliminar todas las plantillas de entrenamiento marcadas.

17. Método según la reivindicación 16, en el que se usa un decodificador de DTW (40) para hacer concordar vectores a posteriori con plantillas a posteriori.

18. Método según una de las reivindicaciones 16 a 17, que comprende asimismo:

- introducir dichos comandos de voz correspondientes a pedidos de bares, restaurantes u hoteles en un dispositivo de usuario (1);
- preprocesar dichos comandos de voz en dicho dispositivo de usuario (1);
- transmitir (7, 8) unas señales preprocesadas a un servidor (6);
- convertir dichas señales preprocesadas en pedidos de texto en dicho servidor (6);
- visualizar dichos pedidos de texto;
- comunicar dichos pedidos a software y/o sistemas usados por dicho bar, dicho restaurante o dicho hotel.

19. Método según una de las reivindicaciones 16 a 18, que comprende asimismo:

- grabar continuamente dicho comando de voz por medio de un sistema de adquisición de audio,

- seleccionar unos medios de detector de actividad vocal,

- deseleccionar dichos medios de detector de actividad vocal,

- procesar dicho comando de voz un tiempo antes de seleccionar dichos medios de detector de actividad vocal y un tiempo después de deseleccionar dichos medios de detector de actividad vocal.

TÉCNICA ANTERIOR

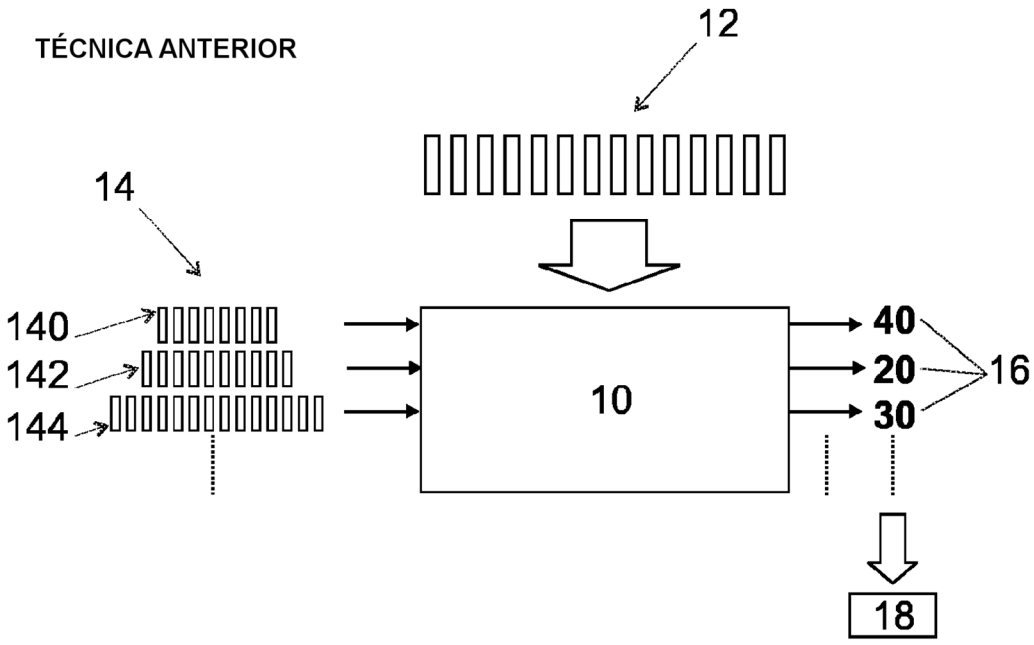

Fig. 1

TÉCNICA ANTERIOR

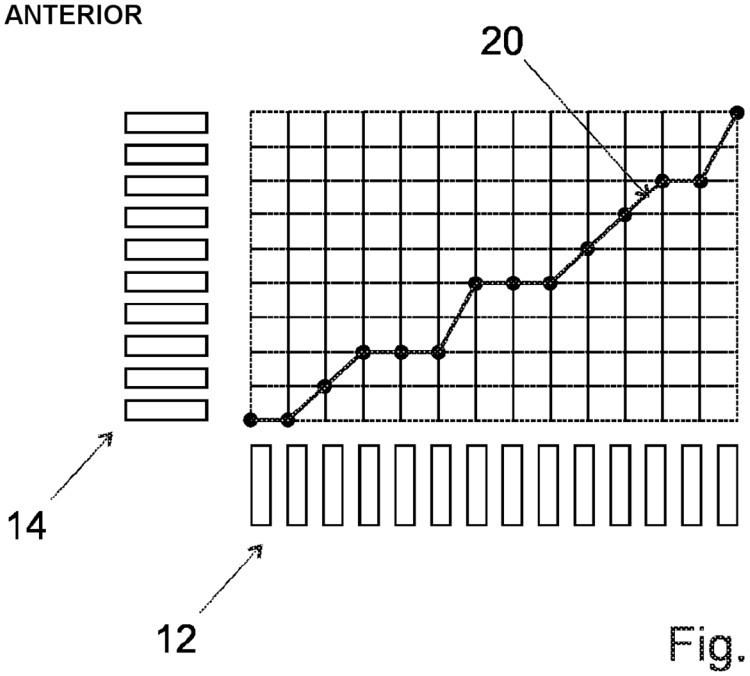

Fig. 2

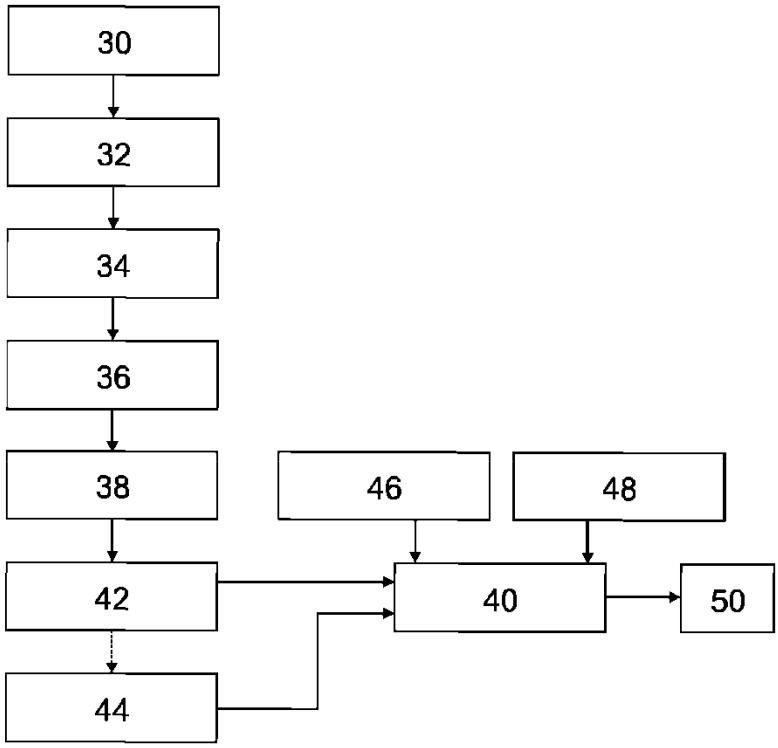

Fig. 3

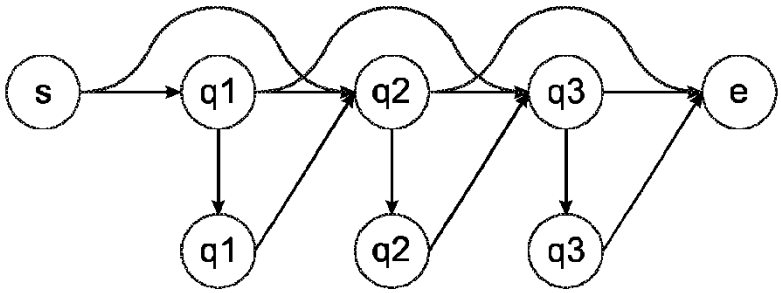

Fig. 4

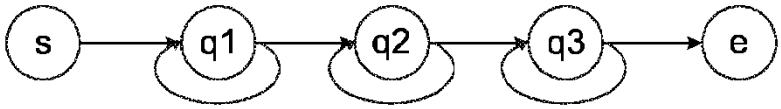

Fig. 5

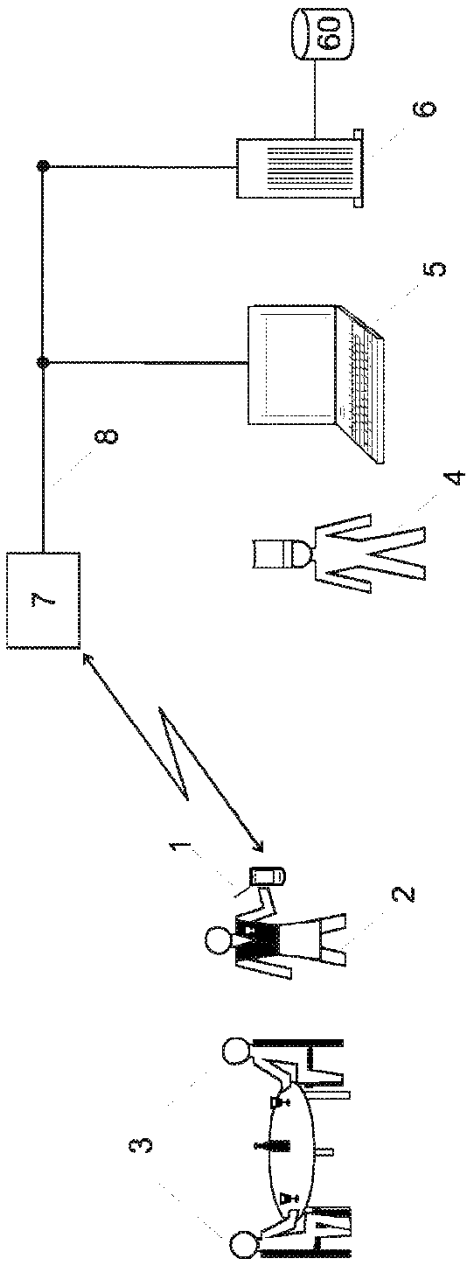

Fig. 6