(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 535 858**

(51) Int. CI.:

G10L 17/10 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **24.08.2007 E 07114958 (7)**

(97) Fecha y número de publicación de la concesión europea: **18.03.2015 EP 2028647**

(54) Título: **Procedimiento y dispositivo para la clasificación de interlocutores**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
18.05.2015

(73) Titular/es:

DEUTSCHE TELEKOM AG (100.0%)
FRIEDRICH-EBERT-ALLEE 140
53113 BONN, DE

(72) Inventor/es:

RUNGE, FRED;
BURKHARDT, FELIX;
STEGMANN, JOACHIM y
MÜLLER, CHRISTIAN

(74) Agente/Representante:

MORGADES MANONELLES, Juan Antonio

DESCRIPCIÓN

Procedimiento y dispositivo para la clasificación de interlocutores.

Sector de la invención:

La tecnología de voz, es decir, el tratamiento mediante máquinas de la voz natural, ha ganado significación de modo creciente en las décadas pasadas. Pertenecen a las aplicaciones ya establecidas el reconocimiento automático de voz (reconocimiento automático de voz, ASR) y el sintetizador de voz (text-to-speech, TTS). El estado de la técnica de los sistemas de investigación se encuentra en la situación de llevar a cabo un profundo tratamiento estilístico de los textos de voz. Verbmobil [1] constituye, por ejemplo, un sistema para la traducción de frases expresadas espontáneamente en tres idiomas (alemán, inglés y japonés) en el área de convención de plazos.

En la comunicación entre las personas, el sonido de la voz transporta, no obstante, no solamente la significación de una frase, sino además informaciones que se llaman paralingüísticas, que facilitan, por ejemplo, informaciones sobre características del interlocutor. Corresponde a nuestra experiencia diaria que caracterizamos a las personas con las que nos relacionamos por teléfono en base a su voz, y podemos adecuar nuestro proceso de comunicación de conversación de modo correspondiente.

El desarrollo de sistemas que adapten el comportamiento (diálogo) a las necesidades del usuario es el objeto del modelado de usuario. Consigue cada vez más significación, puesto que la utilización de los sistemas de ordenadores ha salido del ámbito de la mesa de trabajo y han encontrado utilización en muchas áreas de la vida. Las exigencias específicas a los sistemas varían con las diferentes situaciones en las que se utilizan. Por ejemplo, un sistema navegador móvil para peatones debe tener en cuenta, por ejemplo, las circunstancias en que el usuario se encuentra, posiblemente en un cruce de calles con elevado nivel sonoro, en la parte interna de la ciudad y debe dirigir una parte muy importante de su atención a su entorno, mientras que en otro caso, se puede encontrar en un tranquilo banco de un parque y se puede dedicar de manera completa al diálogo con el sistema (ver [2]). El área de investigación de la clasificación de interlocutores define, además, el captar las informaciones sobre el interlocutor, que son necesarias para constituir un modelo de usuario adecuado, directamente sobre la base de las informaciones conseguidas por la voz. En este documento se describirá, por ejemplo, el reconocimiento de la edad y el sexo del interlocutor. No obstante, los principios que se establecen pueden ser utilizados en otras características biológicas y mentales del interlocutor tales como, por ejemplo, dimensiones corporales ([3], [4]) o la situación emocional y afectiva ([5]).

Otras investigaciones dadas a conocer hasta el momento [5] describen un experimento para la identificación de características en la voz, en base a las cuales se puede deducir la carga cognitiva del interlocutor. La situación en la que se basa es la siguiente: un sistema de ayuda móvil debe acompañar a un viajero en su recorrido por un gran aeropuerto. Se puede esperar que el viajero tenga una elevada carga cognitiva porque debe tener en cuenta el número de la puerta de embarque y la hora de embarque, y por el hecho de que actúa sobre él mismo una gran cantidad de informaciones del aeropuerto. Además, se encuentra posiblemente sometido a presión por el tiempo, puesto que en un corto periodo de tiempo hasta la salida, no solamente debe encontrar la puerta de embarque sino que, en su recorrido, desea todavía adquirir un regalo. El sistema debe reconocer esta carga y tenerla en cuenta para la generación de indicaciones de recorrido. El experimento simula la situación en el aeropuerto a través de un complejo experimento de doble función, en el que las personas sometidas a prueba son sometidas de modo artificial a carga cognitiva y a presión por el tiempo, y proponen preguntas al sistema de ayuda tales como: “Debo cambiar las ropas a mi pequeño, ¿Cómo llego al lugar de cuidados infantiles más próximo?”. Como resultado del experimento se pudo preparar una lista de características en base a las cuales se manifiesta la carga en la voz: velocidad de articulación más lenta, más pausas de voz y más largas y, en especial, aparición elevada de las llamadas disfluencias (por ejemplo, autocorrecciones, repeticiones, interrupciones de la frase, o frases erróneas).

El objeto principal de investigación de [6] es el desarrollo de un procedimiento para la utilización de las informaciones paralingüísticas contenidas en la voz para deducir la edad y el sexo del interlocutor. Las condiciones técnicas de entorno se facilitan por un sistema de diálogo móvil de voz natural, que se ocupa, mediante la integración de dicho procedimiento, en la constitución de un modelo de usuario no intrusivo (ver, por ejemplo, también [7]; [8]). A continuación, este sistema se designará como sistema AGENDER.

En Müller, C. (2006), Clasificación de interlocutores sensible a contexto, de dos etapas sobre el ejemplo de edad y sexo [Two-layered Context-Sensitive Speaker Classification on the Example of Age and Gender]. Akademische Verlagsges ellschaft Aka, Berlín se describió un procedimiento para el reconocimiento automático de la edad y sexo del interlocutor. En aquel caso, se utilizaron, no obstante, exclusivamente, características que conducían a la investigación de la edad por la voz de forma directa: “Jitter y Shimmer”, así como la armonía como características de la calidad de la voz (que disminuye con la edad), pausas de voz (que, con cargas cognitivas más reducidas aumentan con el aumento de la edad), velocidad de articulación (que disminuye por iguales causas), y frecuencia base de la voz (“Pitch”) que manifiesta, en primer lugar, una característica de diferenciación entre interlocutores femeninos y masculinos pero que, no obstante, es también relevante para el reconocimiento de la edad. No

obstante, una combinación de estas características con las llamadas cepstrales de corto plazo (representación del espectro, que se utiliza en el reconocimiento de voz y reconocimiento de interlocutor), no se investigó en este caso.

En la obra de Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Müller, C., Huber, R., Andrassy, B., Bauer, J., y Littel, B. (2007), Comparación de Cuatro Enfoques para el Reconocimiento de Edad y Sexo para Aplicaciones Telefónicas, Actas de la 32 Conferencia Internacional sobre Acústica, Habla y Proceso de Señales (ICASSP 2007), Honolulu, Hawaii, se encuentra una descripción simplificada de Müller (2006). Además, se describen en dicho trabajo otros procedimientos para el reconocimiento de voz. El sistema más satisfactorio en este estudio comparativo, se basa en un procedimiento en el que se forman reconocedores fónicos para las diferentes clases de edad. La decisión de a qué clase pertenece una determinada muestra de voz (probabilidad), tenía lugar sobre la base de los valores de confianza de este reconocedor fónico. El procedimiento, no obstante, no se ha acreditado en la práctica puesto que, para la formación, eran necesarios datos de voz anotados manualmente, que solamente se pueden conseguir con elevados costes financieros. En la obra de Metze y otros, (2007), se describió, además por primera vez, un procedimiento en base a las anteriormente mencionadas cepstrales de corto plazo. De todos modos, en este caso, tampoco se tuvieron en cuenta las características explícitas procedentes de la investigación de la edad de la voz.

Utilizaciones:

La clasificación de interlocutores tal como, por ejemplo, el reconocimiento de edad del interlocutor y sexo, se puede considerar como procedimiento no intrusivo para la captación de un modelo de interlocutor. El Mobile ShopAssist [9] es, por ejemplo, una aplicación Pocket-PC (ordenador de bolsillo), que sirve además para demostrar la utilización de habla natural en un entorno típico de compras. Un tema central de esta aplicación es la interacción móvil y multimodal, que puede consistir en forma de gestos, voz, escritura y una combinación de los mismos. El Personal Navigator [10] (Navegador personal) es una aplicación muy similar en cuanto a la técnica de interacción: Los usuarios pueden efectuar consultas de ruta mediante una combinación de voz y gestos o pedir informaciones sobre edificios que se encuentran en las proximidades.

En base al modelo de usuario que se ha puesto a disposición por el AGENDER, el auxiliar de compras puede realizar una elección específica de productos, en el caso de cámaras digitales, por ejemplo, en caso de que el interlocutor ha sido reconocido como femenino, puede presentar, en principio, un modelo que el fabricante ha desarrollado especialmente para señoras. El sistema de navegación puede adecuar la elección de rutas alternativas: por ejemplo, cuando se ha reconocido que el interlocutor es un niño, en el caso de un guía para turistas se puede escoger un recorrido por el centro de la ciudad con elementos dignos de contemplación especialmente para niños, además, de manera que el recorrido presente el menor número posible de cruces peligrosos.

Otra área de utilización adicional para la clasificación de interlocutores son los servicios basados en el teléfono (ver figura 2). Un centro de llamadas tal como, por ejemplo, una línea de pedidos o de servicio inmediato (“hotline”) representa costes elevados para el explotador, por lo que hay un interés especialmente elevado en la industria de las telecomunicaciones, como ofertantes de plataformas, en soluciones para el aumento de la eficiencia. Un componente central de la técnica del centro de llamadas es la llamada Distribución Automática de Llamadas (ACD) (“Automatic Call Distribution”), que es un sistema soportado por ordenador que recibe llamadas y las distribuye a colaboradores o grupos de colaboradores individualizados. En este caso, en un procedimiento designado como Conditional Routing (“Enrutado Condicional”), se conmutan las llamadas en base a normas previamente fijadas. En los sistemas ACD habituales actualmente, estas normas se refieren principalmente a la proporción de llamadas y la carga. En tiempos recientes se han desarrollado, no obstante, aplicaciones para enrutar llamadas en base a emociones reconocidas por un sistema. La base para ello está constituida por los reconocedores de emociones tal como se han desarrollado, por ejemplo, en el proyecto Verbmobil antes explicado (ver [11]). También existe la posibilidad de integración de la tecnología AGENDER en un sistema ACD.

A los servicios basados en el teléfono, corresponden también los llamados Sistemas de Respuesta de Voz Interactivos, que se diferencian de los servicios de los centros de llamadas en que no son llevados a cabo por agentes humanos, sino por sistemas de ordenador con capacidad de comprensión de voz. Además del conocido ejemplo de información de viajes del ferrocarril, se están extiendo de manera progresiva otros sistemas de información de productos y de sistemas de compras. El objetivo al que se dirigen en este caso los esfuerzos de mejora es menos la reducción de costes que el aumento de la satisfacción de los clientes. La aplicación para AGENDER que resulta de ello es similar a la que se ha descrito para los sistemas móviles de diálogo: En base a un modelo de interlocutor se realiza una elección de productos dirigida a un grupo de clientes, y simultáneamente se adapta el proceso de diálogo del sistema, tal como se muestra en el siguiente ejemplo.

Llamada 1: “¿Cuál es la tarifa para teléfonos móviles?”
AGENDER: Reconoce un interlocutor masculino, joven, y facilita esta información al sistema.
Sistema: “La tarifa XY es exactamente la apropiada para ti. Con ella puedes enviar mensualmente 150 SMS gratis”,

Llamada 2: “¿Cuál es la tarifa para teléfonos móviles?”
AGENDER: Reconoce un interlocutor masculino, de cierta edad, y facilita esta información al sistema.

Sistema: “Le recomendamos la tarifa ABC. A parte de una tarifa básica reducida, proporciona la ventaja de un control de costes completo, incluso en el extranjero”.

Utilización de tecnologías de clasificación de interlocutores en sistemas de diálogo de voz:

Para la adaptación de los escenarios indicados, se utilizan sistemas de diálogo que actúan en la interacción de voz y análogamente a la interacción multimodal, con un mínimo de un usuario. Su constitución general corresponde, por ejemplo, a la que es conocida por: http://www.w3.org/TR/2000/WD-voice-intro-20001204/.

En este caso, se investiga a escala internacional la adaptación de sistemas de diálogo de voz a características individuales del usuario (ver [17]).

Para el guiado del usuario en un sistema de diálogo de voz, este está dotado en general, con utilización de tecnologías de síntesis de voz o con emisión de expresiones registradas, con módulos de emisión de voz en los que se utilizan síntesis de voz o la emisión de expresiones registradas. En este caso, se le pueden explicar al usuario varias opciones de entrada, que él puede facilitar como contestación a una expresión de voz facilitada para un determinado diálogo (“Sprachprompt”) (petición de voz).

En los sistemas modernos de diálogo de voz se utilizan para el reconocimiento de las frases habladas los llamados reconocedores naturales de voz, que posibilitan el reconocimiento, no solamente de palabras individuales, sino también de frases completas, pudiéndose extraer varias informaciones de interés de una frase hablada (por ejemplo, hora de partida de vuelo y lugar de salida y destino). Además, los sistemas de diálogo de voz facilitan principalmente una característica adicional, que posibilita hablar en una petición de voz sin que se deba esperar el final del proceso en desarrollo (Barge-In).

En la adaptación de sistemas de diálogo de voz descrita en el documento WO 02/069320 (páginas 23/24), el centro de gravedad no se encuentra en la adaptación de la estructura del diálogo, sino en la adaptación individual de parámetros de reconocimiento de voz a un usuario individual con formas de expresión no habituales.

También, el documento WO 01/50455 A se encuentra en este campo.

Los sistemas que se describen en la literatura disponen en general de módulos para el reconocimiento de sonido DTMF, reconocimiento de voz y/o reconocimiento de interlocutor. Igualmente, se han ideado módulos para el reconocimiento de emociones, reconocimiento de voz (identificación de lenguaje) y clasificación de interlocutores según, por ejemplo, edad y/o sexo. En los sistemas de diálogo multimodales se utilizan, en general, módulos para la interpretación de la introducción de datos mediante lápiz, ratón, teclado, cámara (gestualidad) o diferentes sensores del aparato Final.

Son conocidos sistemas y procedimientos (ver http://www.nuance.com; "Nuance Verifier, Version 3.0, Developers Guide" así como http://www.ietf.org/internetdrafts/draft-ietf-speechsc-mrcpv2-05.txt (página 6)), que posibilitan al usuario después de una verificación de interlocutor satisfactoria (procedimiento de reconocimiento de interlocutor) el acceso, por ejemplo, a un servicio de voz. Además, se determinará mediante voz o mediante un sistema técnico (ver también, documento EP 124901681 y figura 1) un reconocimiento de usuario (9), que quedará asociado, como mínimo, a una expresión biométrica (por ejemplo, expresión de voz) con la que se puede comparar la muestra biométrica del momento. Las informaciones tienen lugar en sistemas de diálogo de voz, en general mediante síntesis de voz o repetición de determinadas expresiones de voz. En sistemas de diálogos multimodales las informaciones tienen lugar adicionalmente mediante otras tecnologías, principalmente visuales (pantalla, indicaciones luminosas, entre otras).

Se indican a continuación procedimientos conocidos y procedimientos básicos del aprendizaje a máquina:

El suplemento AGENDER, conocido con anterioridad para clasificación de interlocutor, da a conocer una combinación de aspectos controlados por datos y aspectos que se basan en conocimiento. Los modelos han sido generados sobre la base de datos que se han originado por numerosos análisis. Pertenecen a las características investigadas, en base a las cuales se deben deducir la edad del interlocutor y el sexo, las características de la voz tales como frecuencia básica media, “Jitter y Shimmer”, así como características del comportamiento del interlocutor tales como velocidad de articulación y número y duración de las pausas de habla.

Las fases del reconocimiento de modelo que se refieren a la extracción de características y a su clasificación, son designadas en el Primer Plano de AGENDER. Con respecto a la clasificación, se han investigado los siguientes procedimientos conocidos del aprendizaje a máquina:

1. Naive Bayes (NB)
2. Modelos de Mezcla Gaussiana (GMM)
3. k-Vecino Más Próximo (KNN)
4. Árboles de decisión C 4.5 (C45)

5. Máquinas con Vector de Soporte (SVM)
6. Redes Neuronales Artificiales (Redes Neurales Artificiales, ANN).

Dado que los procedimientos Máquina con Vector de Soporte y Modelos de Mezcla Gaussiana son de especial significación para la descripción del sistema reivindicado, se explicarán de manera detallada en las secciones siguientes.

Pertenece a las peculiaridades del suplemento de clasificación de interlocutor AGENDER conocido con anterioridad, el tipo de post-proceso, que se designará como Segundo Plano: Varias funciones típicas del proceso de trabajo se solucionan con AGENDER, con ayuda de un único mecanismo, a saber “Dynamischer Bayes'scher Netze” (DBN) (Redes Dinámicas de Bayes). Estas pueden ser utilizadas para modelar explícitamente la inseguridad inherente a la clasificación, en segundo lugar, permitir el flujo descendente de conocimientos en el proceso de decisión tal como, por ejemplo, el hecho de que, según el contexto, los resultados de un determinado clasificador se deben evaluar como más fiables que otros y, en tercer lugar, conseguir una fusión de múltiples resultados de clasificación.

En la tabla 2 se muestra la exactitud de clasificación del sistema conocido anteriormente que, a continuación, se indicará como Sistema de Referencia, en base a una matriz de confusión. Las matrices de confusión constituyen un medio adecuado para demostrar el comportamiento de los clasificadores, puesto que no solamente muestran la exactitud global en forma de tasas positivas reales (“True Positive Rates”), sino que además destacan qué clases fueron escogidas en falso en vez de las verdaderas. Para ello, se indican en la columna de la izquierda, una debajo de otra, las clases correctas, conteniendo las líneas los resultados del clasificador.

De manera correspondiente, la diagonal que en este caso se ha destacado, muestra los correspondientes casos en los que el clasificador ha conseguido una decisión correcta. La información tiene lugar en porcentaje. La definición de las clases de edad es la siguiente: Se designarán como NIÑOS (K) Interlocutores con edades hasta 12 años inclusive (no se diferenciarán los sexos). La clase JÓVENES (JW, JM) comprende interlocutores desde 12 años hasta 19 años inclusive (W= femenino, M= masculino), los interlocutores entre 20 y 64 años inclusive se designan como ADULTOS (más jóvenes) (EW, EM). Desde los 65 años, los interlocutores corresponden a la clase SENIOR (SW, SM).

Máquinas con Vector de Soporte:

Mientras que los procedimientos de clasificación paramétrica parten de un valor de probabilidad conocido y evalúan su parámetro con ayuda de datos de formación (“training”), las máquinas con vector de soporte adoptan la forma de funciones de discriminación lineal y deducen el parámetro del clasificador. Se consideran, por lo tanto, procedimientos no paramétricos (ver [12, página 216]). Una función discriminatoria, que muestra una combinación lineal de componentes de x, se puede describir del modo siguiente.

$$g(x)=w^t x + \omega_o \qquad (1)$$

en la que w, en el vector de ponderación $w^t x$, representa el producto interno del mismo con el vector de característica x, y ω_0 representa un factor de ponderación umbral. La norma de decisión para el caso de dos categorías es la siguiente: Decisión de ω_1 cuando g(x) > 0, y ω_2 cuando g(x) < 0. De este modo se atribuye a x la categoría ω_1, cuando el producto interno supera el valor de umbral $-\omega_0$, y de lo contrario, ω_2. El límite de decisión quedará constituido por g(x) = 0. En un caso de multivariantes, se trata de una superficie de decisión, un hiperplano (ver [12, página 217]).

La figura 3 muestra un clasificador lineal simple con d unidades de entrada (vector de características). La unidad de valor de umbral facilita siempre el valor constante 1.0. Cada uno de los valores de entrada x_i será multiplicado con su ponderación ω_i, encontrándose, por lo tanto, en la unidad de salida ix_i. Esta facilita en este caso +1, en caso de que $\omega^t x + \omega_0 > 0$, y por lo contrario -1.

El hiperplano separador H divide el espacio de características en dos medios espacios: La región de decisión R_1 para ω_1 y la región de decisión R_2 para ω_2. Frecuentemente, se utiliza la forma de representación en la que x se encuentran en R_1 en el lado positivo de H, y todas las x en R_2 se encuentran en el lado negativo. La posición de H será determinada mediante el factor de ponderación de umbral w0 y la tendencia por el vector de ponderación w.

El valor positivo o negativo de g(x) constituye una medida algebraica de la distancia de x a H, mediante la ecuación

$$r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}, \qquad (2)$$

en la que w indica la norma euclídistica de w, es decir $\sqrt{\omega^t \omega}$

La ecuación 1 está representada también por

$$g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i. \qquad (3)$$

Para funciones de discriminación lineales, es válido

$$g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=0}^{d} w_i x_i, \qquad (4)$$

en la que se cumple $x_0 = 1$. De esta manera, se puede describir el vector de característica ampliado y mediante la ecuación 5, y de forma análoga, el vector de ponderación ampliado a mediante la ecuación 6.

$$\mathbf{y} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix}. \qquad (5)$$

$$\mathbf{a} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix}. \qquad (6)$$

La función de discriminación g(x) puede ser descrita, por lo tanto, en forma de $a^t y$.

Se supondrá que, en el banco de datos de aprendizaje, se encuentra una serie de muestras $y_1,\ldots y_n$, de los que algunas están marcadas ω_1, y otras con ω_2. Estas muestras deben ser utilizadas para determinar las ponderaciones a en la función de discriminantes lineales $g(x) = a^t y$. Cuando se tiene a, para la cual se clasificarán correctamente todos los datos de aprendizaje, estos se llaman separables linealmente (ver [12, página 223f]). En el caso de dos categorías, todas las muestras y_1 con las etiquetas ω_2 pueden ser cambiadas por y_1. A continuación, se buscará un vector de ponderación a, de manera que se cumpla $a^t y_i > 0$ para todos los datos de aprendizaje. Este vector se designará vector de corte o de separación. Habitualmente, hace máxima la distancia mínima de las muestras con respecto al hiperplano H, de manera que se cumple $a^t y_i \geq b$ para todas las i La constante positiva b se designará borde. Las máquinas con vector de soporte (SVM) se basan en la idea básica de clasificadores lineales con bordes, transfiriendo, no obstante, los datos de aprendizaje previamente en un espacio superdimensional. En este caso, se partirá de la suposición que, con una figura no lineal apropiada $\gamma(\cdot)$ en una dimensión suficientemente elevada, los datos de aprendizaje de dos categorías se pueden separar siempre mediante un hiperplano (ver [12, página 262]).

Las SVM buscan el hiperplano óptimo H, que es aquel con el borde b más elevado. Los vectores de soporte ("support vectors") son aquellas muestras (transformadas) que determinan el borde y, por lo tanto, H (ver, 4). Cuando se han determinado los vectores de soporte, se pueden deducir todos los otros datos de aprendizaje del modelo, sin que varíe la posición y orientación de H [13, página 190]. Constituyen simultáneamente las muestras que son más difíciles de clasificar (ver [12, 5.262j).

El exclusivo ODER (XOR) muestra el problema más sencillo que no se puede solucionar con una función de discriminante lineal. De acuerdo con SVM, las características son representadas en una etapa previa de preparación en una dimensión más elevada, en la que son entonces separables. [12, página 264] utilizan para este ejemplo las funciones de transformación $\Gamma = \{1, \sqrt{2}\,X1, \sqrt{2}\,X2, \sqrt{2}\,X1X2, X1^2, X2^2\}$.

La figura 5 (izquierda) muestra el espacio original de características del problema: Los puntos oscuros, rojos, pertenecen a la categoría ω_1, y los puntos claros, verdes, a la categoría ω_2. Los cuatro puntos de aprendizaje son representados con ayuda de r en un espacio de seis dimensiones, de manera que son separables mediante la función de discriminantes $g(x) = x_1 x_2$. En la figura 5 (derecha), se ha mostrado una proyección bidimensional de dicho espacio. A causa de la fuerte simetría del problema, la totalidad de las cuatro características son vectores de soporte (ver [12, página 264f]).

De este ejemplo resulta evidente que una etapa central en la construcción de una SVM es la elección de la cantidad r apropiada. Depende frecuentemente del conocimiento de dominios del diseñador. De lo contrario, se escogen frecuentemente funciones polinómicas de Gauss u otras funciones elementales. [13, S. 188f] facilitan un ejemplo en el que la cantidad original de atributos es transformada por una factorización de n veces. Para dos atributos y n = 3, sería

$$\mathbf{y} = w_1 x_1^3 + w_2 x_1^2 x_2 + w_3 x_1 x_2^2 + w_4 x_2^3. \quad (7)$$

La dimensionalidad del espacio representado puede ser tan elevada como se desee, pero en la práctica, es limitada mediante recursos técnicos de cálculo. Para una transformación de diez características originales con n=5, se debe determinar el algoritmo de aprendizaje con intermedio de 2000 coeficientes (ver, igual referencia).

Una ventaja de la hipótesis de SVM consiste en que, es menos propensa en general para problemas de “Overfitting” que otros procedimientos. Según [13. página 191], estos se generan siempre cuando los modelos son inestables, es decir, los límites de decisión desplazan menos instancias con la variación. El hiperplano con el borde mayor permanece, no obstante, relativamente estable, dado que solamente varía cuando se añaden vectores de soporte o se anulan. Esto es válido también para un espacio de muchas dimensiones que es solicitado por una transformación no lineal. Los vectores de soporte son representantes globales del banco de datos de aprendizaje en su conjunto. Habitualmente, existe solamente un reducido número de ellos, lo que significa una reducida flexibilidad y, por lo tanto, un menor peligro de “Overfitting” (ver, [13, página 191 f]).

Modelos de mezcla Gaussiana

Los modelos de mezcla Gaussiana (Gaulβ’sche Mixtur-Modelle, GMM) están muy íntimamente relacionados con el clasificador de Bayes. Sirven como modelo probabilístico para densidades multivariadas de probabilidad que pueden representar las densidades deseadas (de Gauss, de Laplace). En la aplicación se calcularán, con ayuda de GMM, las densidades de probabilidad básicas específicas de la clase en base a, las cuales un clasificador de proporción de probabilidades muestra entonces un modelo determinado de una categoría. Para un vector de características x de d dimensiones, la densidad mixta (“mixture-density”) se define del modo siguiente:

$$p(\mathbf{x}|\gamma) = \sum_{j=1}^{M} w_j p_j(\mathbf{x}). \quad (8)$$

La densidad de probabilidad es una combinación lineal de densidades de probabilidad M de Gauss. Los factores de ponderación de mezcla ω_i cumplen además la condición

$$\sum_{j=1}^{M} w_j = 1. \quad (9)$$

En su conjunto, se designan los parámetros del modelo por [$\gamma = \{\omega_j \mu_j, \Sigma_j\}$, en el que j=1,..., M. En base a una reunión de muestras de aprendizaje se determinan los parámetros con ayuda del algoritmo de maximización-expectativa iterativo (algoritmo EM). Este adapta los parámetros de GMM de forma tal que se alcanza una mejora monótona de la probabilidad del modelo de los vectores de características observados. Para las iteraciones k y k+1 se cumple, por ejemplo, $p(x| \gamma_{k+1} > p(x|\gamma_k)$.

Un GMM puede ser considerado como modelo híbrido entre modelos paramétricos y no paramétricos puesto que, a pesar de que los parámetros básicos determinan el comportamiento, un grado más elevado de libertad permite las densidades de probabilidad deseadas.

Los modelos de mezcla Gaussiana constituyen un procedimiento posible para la clasificación basada en trama, en el reconocimiento de interlocutor y similares (ver, por ejemplo, [14], [15], [16]). Las características utilizadas son, en la mayor parte de casos, Coeficientes Cepstral de Frecuencia Mel (MFCC), que se combinan frecuentemente con la primera y segunda desviaciones temporales (Delta-MFCC, DeltaDelta-MFCC). Los MFCC son características establecidas basadas en trama en el reconocimiento de interlocutor. Se designarán como basados en trama porque la señal de voz continua se divide en primer lugar mediante una ventana de exploración en secciones (tramas) de una longitud aproximada de 10 ms, para poder llevar a cabo una transformación de Fourier. La designación “Mel” indica la escala de frecuencias de igual nombre, que se orienta a la percepción humana de la frecuencia básica. La designación “cepstral” se deriva de “cepstrum”, que muestra una derivación de “spectrum” (espectro). Se debe

resaltar con ello que se trata de una característica que de algún modo se deriva del “espectro de los espectros”. Las MFCC se facilitan en forma de un vector de características de 13 dimensiones. Para clasificación, la dimensión nula se deja habitualmente fuera de consideración, puesto que muestra una medida de la amplitud de la señal.

Desarrollos adicionales conocidos

Los desarrollos adicionales actuales del sistema AGENDER reflejan los resultados de un “Benchmark-Workshops” para el reconocimiento de edad y género. El procedimiento más satisfactorio en esta prueba tenía en cuenta el hecho de que una afirmación con respecto al interlocutor (o bien la clase de interlocutor) es más exacta cuando esta es llevada a cabo en base al conocimiento del contenido de la expresión.

En el módulo que se ha explicado, se tiene en cuenta este conocimiento de forma implícita, puesto que para cada una de las clases de interlocutores se forma un reconocedor de fonemas separado y la asignación de clase se lleva a cabo en base al mayor valor de confianza de todos los reconocedores. Este procedimiento tiene, no obstante, el inconveniente que para una definición de edad alterada se debe formar nuevamente el reconocedor de fonemas, lo cual presupone la disponibilidad de suficiente material de formación etiquetado fonéticamente. El sistema AGENDER que se ha desarrollado posteriormente utiliza por el contrario, independientemente de la definición de clases de edad básicas con un solo reconocedor de fonemas.

En la figura 6 se muestra un sistema A. Se utiliza un reconocedor de fonemas único como alineador (“Aligner”), que sirve como frontal para una extracción de características basada en segmentos. De manera adicional, se utilizan las características basadas en una expresión, tal como es conocido.

Para cada clase de edad se entrenará una máquina con vector de soporte como clasificador binario, cuyos resultados se relacionarán entre sí a nivel de calificación “Score-Level”. A continuación, se explicarán de manera detallada los componentes individuales del módulo.

En todos los sistemas que se describen (A, B) se diferenciará entre características basadas en segmentos y características basadas en afirmación. Estas últimas se corresponden a las características utilizadas hasta el momento, es decir, por ejemplo, pitch_mean (frecuencia básica media), jitter_ppq (microvariaciones de la frecuencia según el procedimiento PPQ) o shimmer_aq11 (microvariaciones de la amplitud según el procedimiento APQ11). Se representarán de modo correspondiente por un valor único para una afirmación determinada. En conjunto, se han utilizado hasta el momento diecisiete de estas características siendo, por lo tanto, el vector de características de 17 dimensiones.

Las características basadas en segmentos se diferencian por el hecho de que para cada segmento (fonema) se determina un valor. Se han llevado a cabo estudios para una serie de características distintas basadas en diferentes segmentos: Frecuencias de formación (F1 a F5), las variantes basadas en segmentos de las características anteriores, así como MFCC (con y sin Delta-MFCC). Al contrario que en ambas variantes mencionadas en primer lugar, se puede conseguir con utilización del MFCC una significativa mejora de los resultados, especialmente en combinación con una selección de las características anteriores basadas en la expresión (ante todo “Dingen Pitch”). Este resultado se explica por el hecho de que las MFCC muestran características que están correlacionadas con la configuración del aparato bucal del interlocutor, pero no con las características de la señal de excitación. Estas últimas son representadas por el contrario por las características anteriores, que además de “Pitch”, se basan también en las características Jitter y Shimmer en la frecuencia básica.

El inventario de las características basadas en segmentos comprende MFCC_O hasta MFCC_12, lo que corresponde a un inventario disponible de 50 fonemas de un vector de características de 650 dimensiones. De manera adicional, se añaden según la configuración individual de los modelos, de 5 a 10 características “tradicionales”.

En base a las marcas finales y de principio de un fonema determinado, que son recogidas en la salida del alineador, la expresión es dividida en segmentos (fonemas). Para cada uno de los fonemas se calcula la media de MFCC_O a M FCC_ 12. La base para la utilización de la mediana en vez de la media aritmética consiste en que esta es menos influenciada por “elevaciones”. Se debe contar con estas en el procedimiento determinado por dos causas: En primer lugar, las marcas del inicio y final de un fonema no son exactas y, en segundo lugar, las zonas de borde de un fonema reciben la influencia de la transición del fonema anterior al fonema siguiente. El ejemplo de la tabla 3 muestra la ventaja de la mediana con respecto al valor medio.

Un resultado similar se puede conseguir también con utilización del valor medio cuando éste es calculado en base a una ventana en el centro del segmento. Se han llevado a cabo experimentos con ventanas de diferentes longitudes. Los mejores resultados se alcanzaron con 60% de la longitud del segmento, la diferencia con respecto a la mediana era, no obstante, marginal. Se hace observar en este punto que no es deseable en todos los casos despreciar las zonas de borde para el cálculo del valor. También es muy probable que la transición entre los fonemas contenga informaciones específicas de la clase del interlocutor. La utilización de estas zonas como característica presupone, no obstante, otras exploraciones adicionales. En la tabla 5, se muestra la exactitud de clasificación del sistema A

para idéntica formación y material de prueba. A pesar de que la exactitud promedio no es significativamente más elevada que en el sistema de referencia, la matriz está mejor equilibrada, es decir, la diferencia entre la menor “True-Posilive-Rate” (Tasa Positiva Real) y la siguiente es menor.

A continuación, se describirá el sistema B de la figura 7. Puesto que solamente muy pocas expresiones cubren el inventario de fonemas completo, es produce en el sistema A una proporción mayor de “missing values” (valores que faltan). El sistema B es una variante en la que se puede superar este problema, de manera que para cada segmento se genere un modelo separado. Tal como se muestra en la figura 7 se enlazaron los modelos entre sí en el “Score-Level” (nivel de calificación). La tabla 7 comprende la exactitud de clasificación conjunta del sistema B: La exactitud promedio es más elevada que en el sistema A, no obstante, esta mejora ha sido conseguida a costa de la comparatividad. El conjunto del sistema B consiste en que son necesarios múltiples modelos, lo cual tiene un efecto negativo tanto en el comportamiento del transcurso del tiempo del sistema como también en la verosimilitud del sistema.

Resumen de la invención

Un objetivo esencial de la presente invención es no solamente la mejora de la exactitud de la clasificación de interlocutor con reducción de la tasa de fallos, sino también la preparación de un procedimiento para aumentar la eficiencia del proceso de clasificación.

Se consigue este objetivo mediante un procedimiento y dispositivo que presenta las características de las reivindicaciones independientes.

En las secciones siguientes, se describirán tres procedimientos distintos, que en los estudios más nuevos han sido comparados entre sí (Sistemas A, B y C). Se exceptuará el sistema C, que será descrito en una sección posterior de manera precisa.

Mediante el procedimiento combinado de varias etapas que se describe según la invención, se consigue, en comparación con la utilización separada de la identificación de lenguaje, una reducción sustancial de la tasa de errores en la clasificación de la lengua hablada en los sistemas de diálogo de voz. Es ventajoso que no se requieren recursos adicionales, sino solamente la utilización adicional combinada de los sistemas de reconocimiento de voz existentes en los sistemas de diálogo de voz para conseguir tasas de éxito mejoradas para la consecución del resultado final.

Descripción de las figuras

A continuación, se describirán de manera abreviada las figuras, sin que ello signifique una limitación del ámbito de protección. Se muestra:

La figura 1, El sistema AGENDER con el escenario de utilización “Adaptive mobile Systeme” con el ejemplo de m3i Navegador/personal y m3i ShopAssist

Figura 2, Sistema AGENDER con el escenario de utilización “Callcenter” con el ejemplo de línea de servicio inmediato (“Service Hotline”) y sistema de compra.

Figura 3, Un clasificador lineal simple según [12, página 216]

Figura 4, Límite de decisión, vectores de borde y de soporte de un SVM según [12, 5.262] Y2

Figura 5, Izquierda: espacio de características original del problema XOR. Derecha: proyección de un espacio de características transferido a un espacio de seis dimensiones. Eje-x: 'J x1, Eje-y: 2 xlx2. El límite de decisión es ahora lineal (ver [12, 5.264])

Figura 6, Representación esquemática del sistema A para clasificación de interlocutor

Figura 7, Representación esquemática del sistema B

Figura 8, Representación esquemática de una realización preferente del sistema reivindicado C

Figura 9, Representación esquemática de un sistema D

Figura 10, Representación esquemática de un sistema E con almacenamiento,

Figura 11, Representación esquemática de un sistema F

Figura 12, Representación esquemática de un sistema de diálogo con clasificación de interlocutor implementada

Tabla 2, Muestra la exactitud de clasificación del sistema de referencia

Tabla 3, Muestra el valor medio con respecto a la mediana con una serie de valores con elevaciones

Tabla 5, Muestra la exactitud de clasificación del sistema A

Tabla 7, Muestra la exactitud de clasificación del sistema B

Tabla 9, Muestra la exactitud de clasificación del sistema C

Descripción de las formas de realización

El sistema C mostrado esquemáticamente en la figura 8 presenta dos modelos por clase (por ejemplo, clase de edad o clase de sexo) a base de: Una es la SVM en base a las características de tono basadas en la expresión, según el sistema B. El otro es un modelo que se basa igual que antes sobre las MFCC, pero estas tratadas, no obstante "trama a trama" en vez de valores promedio (o valores mediana) de un segmento. Para ello, se utilizó un GMM en vez de un SVM, un procedimiento que es utilizado en el ámbito de la clasificación de interlocutor de manera satisfactoria con respecto a la preparación de voz basada en tramas (ver anterior). Mediante la elaboración de la señal basada en tramas, se puede prescindir de una parte central ("Frontend") de segmentación, lo que tiene un efecto positivo sobre la complejidad y la velocidad de clasificación del sistema.

Por lo tanto, se reivindica la combinación de, como mínimo, un SVM para cada clase de interlocutor con, como mínimo, un GMM para cada clase de interlocutor, que se basa en MFCC, que es tratado "trama a trama".

Los GMM en base a MFCC (Delta- y Delta-Delta-MFCC) han sido utilizados también anteriormente en sistemas para la clasificación de interlocutor. Existe un inconveniente por el hecho de que los MFCC muestran ciertamente una medida apropiada para el modelado de las características del tracto bocal humano, pero prescinden de las características de la señal de excitación (que contienen ciertamente informaciones específicas del interlocutor). Las características de tono (características que se basan en la frecuencia básica de la voz) son apropiadas, por otra parte, para modelar las características de la señal de excitación. Diferentes derivados estáticos del tono se han utilizado ya anteriormente en sistemas para la clasificación del interlocutor (ver [6]). En el procedimiento que se describe se ha utilizado, no obstante, por primera vez, para la aplicación de la clasificación de interlocutor una combinación de MFCC y características basadas en el tono.

Un parámetro importante en el desarrollo de un GMM para una aplicación determinada es el número de mezclas (también designadas de Gauss). Un número mayor de mezclas modela habitualmente el material de aprendizaje mejor, pero introduce el peligro de los "Overfittings", lo que significa que el modelo muestra una exactitud de clasificación poco satisfactoria en los ejemplos que no se han considerado hasta el momento. Mientras que en el reconocimiento de interlocutores habitualmente se utilizan hasta 1024 mezclas, el problema de clasificación existente requiere una generalización más fuerte y, por lo tanto, un número significativamente más reducido.

Se han realizado experimentos con 16, 32, 64, 96 y 128 mezclas. Los mejores resultados se han conseguido con un número de 96. Los GMM basados en tramas y los SVM basados en expresiones se combinaron al nivel de calificación entre sí. Las ponderaciones se determinaron con ayuda de una búsqueda completa específica de clase (es decir, de siete dimensiones) en el intervalo de 0 a tres, y un incremento de 0,1. Las ponderaciones óptimas para el problema determinado se encuentran en GMM + 0,3 * SVM para todas las clases.

Tal como se ha mostrado en la tabla 9, la exactitud promedio con un valor de 49,11% es sustancialmente más elevada que en todos los demás sistemas. Simultáneamente, las "True-Positive-Rates" (Tasas Positivas Verdaderas) individuales están relativamente equilibradas.

El procedimiento descrito al principio de esta sección y reivindicado, es una disposición ventajosa de un procedimiento para la clasificación de interlocutor que combina, como mínimo, dos procedimientos de aprendizaje a máquina, con el objetivo de la clasificación de interlocutor. Se han indicado ya ejemplos para procedimientos utilizables del aprendizaje a máquina en la introducción. Se reivindica, por lo tanto, un procedimiento para la clasificación de interlocutor en el que, como mínimo, se combinan dos procedimientos, preferentemente tres, entre sí, tal como se ha mostrado, por ejemplo, en la figura 8 para los procedimientos combinados (figura 9). En este caso, según los procedimientos de aprendizaje individuales que se han combinado entre sí tal como, por ejemplo, se ha indicado ya en la figura 8 para dos procedimientos combinados, se extraen de la señal de voz, características dependientes de las características necesarias para los procedimientos individuales, se procesan y se facilitan al correspondiente procedimiento del aprendizaje a máquina.

Tal como se ha descrito inicialmente, un procedimiento de este tipo puede ser unido a un sistema de diálogo de voz, que está conectado a una red apropiada para el envío de datos de voz (por ejemplo, red fija, red de radio, internet, y otras) (figura 12).

Las características del usuario determinadas según el procedimiento de aprendizaje utilizado, se pueden enlazar ahora con el conocimiento determinado en el sistema de dialogo del aparato final relacionado con el sistema de dialogo (por ejemplo, HLR, CLI, IMEI, reconocimiento SIM, Dirección IP, número de llamada, Dirección SIP y otros), y un reconocimiento previamente determinado y facilitado a través del sistema de diálogo y eventualmente verificado, y pudiendo ser almacenado, por ejemplo, en una base de datos (ver figura 10, almacenamiento I…n). Como mínimo uno de estos reconocimientos, en especial el reconocimiento indicado en último lugar, puede ser adicionalmente verificado mediante la utilización de la verificación de interlocutor (3b de la figura 12) u otros procedimientos de verificación conocidos. De esta manera, para una repetida conexión con el sistema de diálogo, las características almacenadas y asociadas a un usuario de manera cierta, según el proceso de verificación anteriormente indicado, se pueden añadir a las características de voz de la interacción momentánea, de manera que se consigue una mayor cantidad de características y, por lo tanto, una mayor seguridad en la determinación de la clase de interlocutor.

Igualmente, se puede trabajar con la reunión de datos de voz en bruto (figura 11) que para una nueva conexión con el sistema de diálogo se añaden a los nuevos datos recogidos. Estos datos ampliados serán nuevamente almacenados bajo la designación determinada, y facilitados a la extracción de características.

La figura 12 muestra la constitución básica del sistema de diálogo 8. Un aparato final 1 del usuario de un sistema de diálogo de voz constituye mediante la red de conexión 2 (de forma inalámbrica y/o mediante cable, orientado a la conexión y/o basado en paquetes) una conexión.

Mediante un reconocimiento de entrada 3 se llevan a cabo, por una parte, la determinación de un aparato final/reconocimiento de conexión 3a, y por otra, un reconocimiento de interlocutor o bien verificación de interlocutor 3b, preferentemente también para llevar a cabo un diálogo para la verificación del usuario.

A continuación, tiene lugar una clasificación de interlocutor 3c, de acuerdo con el procedimiento reivindicado y, eventualmente, un reconocimiento de voz 3d. Adicionalmente, puede tener lugar una captación de un teclado, y/o lápiz, y/o un ratón 3e, y/o captación de movimiento, preferentemente también para la realización de un diálogo para la verificación del usuario.

Las informaciones introducidas que se han citado se comprobarán en la unidad de evaluación de informaciones introducidas y/o unidad de interpretación 4. Esto tiene lugar habitualmente mediante la unidad de proceso reivindicada que consiste en un procesador y/o un ASIC.

El gestor de interacción/aplicación 7 recibe los datos interpretados de la unidad de interpretación 4 y facilita estas a aplicaciones, que se han indicado y que procesan adicionalmente los resultados, tales como por ejemplo, un sistema de diálogo, banco de datos, sistema de autentificación, etc. A continuación, tiene lugar un envío al sistema de planificación de salida 6 y, finalmente, una salida óptica o acústica 5.

Referencias de literatura

[1] Wolfgang Wahlster, Verbmobil: Bases de la traducción voz-voz, Eliis Horwood Series in Artificial lntelligence, Springer, Berlín -Heidelberg -New York, 2000.

[2] Christian Müller, “Diálogo multimodal en un sistema de navegación para peatones,” en Proceedings of ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments, Kloster lrsee, Germany, 2002, pp.42-44,

[3] Sorin Dusan, “Estimación de la altura del interlocutor, y longitud del tracto bucal a partir de la señal de voz,” in Proceedings of 9th European Conference on Speech Communication and Technology Unterspeech 2005), Lisbon, 10 Portugal,2005.

[4] B.L. Pellom and J.H.L. Hansen, “Análisis de voz en condiciones adversas: Llamada a 911 de la bomba en El Parque Olímpico del Centenario,” in Proceedings of the 40th IEEE Midwest Symposium an Circuits and Systems, Sacramento, CA, 1997,

[5] Christian Müller, Barbara Groβmann-Hutter, Anthony Jameson, Ralf Rummer, and Frank Wittig, “Reconocimiento de la urgencia temporal y carga cognitiva en la base del habla: Estudio Experimental,” in UM2001, UserMadeling. Proceedings of the Eighth international Conference, Mathias Bauer, Piotr Gmytrasiewicz, and Julita Vassileva, Eds., New York- Berlín, 2001, pp. 24- 33, Springer.

[6] Christian Müller, Clasificación de interlocutores de dos etapas, sensible al contexto con el ejemplo de edad y sexo, Ph.D. thesis, Fachbereich 6.2 lnformatik, Universitat des Saarlandes, Deutschland, 2005.

[7] Christian Müller, F. Wittig, and J. Baus, “Utilización de la voz para el reconocimiento de usuarios de edad en responder a sus necesidades especiales,” in Proceedings of the Eighth European Conference n Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, 2003, pp. 1305- 1308.

[8] Christian Müller and Frank Wittig, “La voz como fuente de modelación de usuarios ubicuos,” in Proceedings of the Workshop on User Modeling in Ubiquitaus Computing in conjunction with the Ninth lnternational 25Conference on User Modeling (UM 2003), Pittsburgh, USA, 2003, pp. 46- 50.

[9] Rainer Wasinger, Antonio Krüger, and Oliver Jacobs, “Integración de gestos intra y extra en un asistente multimodal de teléfono móvil,” in Proceedings of the 3rd lnternational Conference on Pervasive Computing (Pervasive), Munich, Germany, 2005, pp. 297-314.

[10] Rainer Wasinger, Christoph Stahl, and Antonio Krüger, “M31 en sistema de navegación y exploración para peatones,” in Proceedings of the Fourth international Symposium on Human Computer lnteraction with Mobile Devices, Pisa, 5 ltaly, 2003, pp. 481-485.

[11] A. Batliner, R. Huber, H. Niemann, E. Niith, J. Spilker, and K. Fischer, Reconocimiento de Emociones, pp. 122-130, Springer, New York-Berlin, 2000.

[12] Richard O. Duda, Peter E. Hart, and David G. Stark, Clasificación de Modelos, Wiley-lnterscience, New York, USA, 2. edition, 2000.

[13] lan H, Witten and Eibe Frank, Data Mining: Herramientas y técnicas prácticas para aprendizaje a máquina con implementaciones de Java, Margan Kaufmann Publishers, San Mateo, CA, USA, 1999.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Verificación de interlocutores utilizando modelos de mezcla Gaussiana adaptados,” Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[15] D.E. Sturim, D.A. Reynolds, R.B. Dunn, and T.F.Quatieri, “Verificación de interlocutores utilizando modelos de mezcla Gaussiana limitada por el texto,” in Proceedings of the lnternational Conference an Acoustics, Speech, and Signal Processing (ICASSP 2002), Orlando, FL, 2002, pp. 1: 677-680,1EEE.

[16] C. Miyajima, Y. Hattori, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Identificación de interlocutores utilizando modelos de mezcla Gaussiana basados en una distribución de probabilidades multiespacial,” in Proceedings ofthe lnternational Confer ence an Acoustics, Speech, and Signal Processing (ICASSP 2001), Salt Lake City, Utah, 2001, IEEE.

[17] Hjalmarsson, A., Sistemas de diálogo hablado adaptativo', Centre for Speech Technofogy, KTH, Jan. 2005

1 Aparato de usuario de un diálogo de voz, interlocutor que llama
2 Red de conexión (inalámbrica y/o con conductores, orientada a la conexión y/o basada en paquetes)
3 Identificación de informaciones de entrada
3a Determinación del identificador de un aparato final/conexión
3b Identificación de interlocutor o verificación de interlocutor, preferentemente, asimismo para llevar a cabo un diálogo para la verificación del usuario
3c Clasificación de interlocutores según la descripción
3d Identificación de interlocutores
3e Captación de teclado y/o botones y/o ratón y/o movimiento, preferentemente, asimismo para llevar a cabo un diálogo para la verificación del usuario
4 Evaluación y/o interpretación de entrada de informaciones
5 Salida de informaciones
6 Planificación de salidas
7 Gestor de interacción/aplicación
8 Sistema de diálogo

Definición de abreviaturas

ANI	Identificación Automática de Número
ANN	Redes Neurales Artificiales- Neuronales Artificiales
APQ	Cociente de Perturbación de Amplitud
ASR	Reconocimiento Automático de voz
C45	Árbol de Decisión C 4.5 (Procedimiento de Aprendizaje a Máquina)
CLI	Identificación de la Línea que Llama
DBN	Red Dinámica de Bayes
EM-Aigorithmus	Algoritmo de Expectativa-Maximización
GMMs	Modelos de Mezcla Gaussiana (Procedimiento de Aprendizaje a Máquina)
Grammatik	Descripción estructurada de posibles informaciones introducidas a evaluar por el usuario (por ejemplo, Voz de conversación, Entradas de texto, Botones, Mímica de Rostro, etc.)
HLR	Registro de Localización Interno
IMEI	Identidad de Equipo Móvil Internacional
KNN	Vecino K más Próximo (Procedimiento de Aprendizaje a Máquina)
MFCC	Coeficiente Mel-Frecuencia-Cepstral
NB	Bayes Natural (Procedimiento de Aprendizaje a Máquina
PRO	Cociente de Perturbación de Tono
SIM	Módulo de Identidad de Abonado
SIP	Protocolo de Iniciación de Sesión
Clasificación de Interlocutores	Determinación de la adecuación, como mínimo, de un interlocutor con respecto a una mayor
Reconocimiento de interlocutores	Autentificación o Identificación de un interlocutor en base a características de
SVM	Máquina con Vector de Soporte
TTS	Texto a Voz

REIVINDICACIONES

1. Procedimiento para la clasificación automática de un interlocutor gracias a un sistema numérico, en el que se aplican, como mínimo, dos procedimientos distintos de clasificación de un interlocutor a datos vocales digitales, efectuando la combinación de sus resultados,
en el que el primer procedimiento procesa características a base de segmento, y el segundo procedimiento procesa características a base de expresiones,
en el que el procedimiento a base de expresiones utiliza, como mínimo, un Aparato con Vector de Soporte (SVM) por clase de interlocutor sobre la base de las características de tono basadas en la expresión,
en el que el procedimiento a base de segmentos utiliza, como mínimo, un modelo de mezcla Gaussiana (GMM) por clase de interlocutor, que se basa en coeficientes de frecuencia Mel-Cepstral (MFCC), tratados trama a trama.

2. Procedimiento, según la reivindicación anterior en el que el número de mezclas está comprendido entre 60 y 170, siendo preferentemente 96.

3. Procedimiento, según una o varias de las reivindicaciones anteriores, en el que una combinación del procedimiento de clasificación de interlocutor se realiza sobre la clasificación de nivel (“Score Level”).

4. Procedimiento, según las reivindicaciones anteriores, en el que en la combinación de SVM y CMM se introducen ponderaciones, y las ponderaciones son determinadas por medio de una búsqueda específica de clase.

5. Procedimiento, según una o varias de las reivindicaciones anteriores, en el que antes de la clasificación de interlocutor, tiene lugar una extracción de características de los datos de voz.

6. Procedimiento, según una o varias de las reivindicaciones anteriores, en el que tiene lugar una combinación de identificativos de un terminal, contactando un sistema para una clasificación de interlocutores, con la finalidad de almacenar información relativa a la clasificación de interlocutor con respecto a los identificadores, que son tenidos en cuenta por una nueva clasificación de interlocutor.

7. Procedimiento, según las reivindicaciones anteriores, en el que se lleva a cabo adicionalmente el almacenamiento de características de clasificación de interlocutor o datos digitales de voz en bruto en relación con el identificador.

8. Dispositivo para la clasificación automática de interlocutores, que comprende:
- un sistema digital, que recibe datos de voz de una persona a través de un interfaz;
- una unidad de proceso, que puede tener acceso a los datos de voz, y que está diseñada y dispuesta de manera tal que, como mínimo, se aplican dos procedimientos distintos de clasificación de interlocutores a los datos de voz, y los resultados son combinados entre sí, en el que el primer procedimiento procesa características basadas en segmento, y el segundo procedimiento procesa características basadas en expresión,

en el que el procedimiento basado en expresión (SVM) utiliza, como mínimo, un aparato con vector de soporte para cada clase de interlocutor en base a características de tono basadas en la expresión,
de manera que el procedimiento basado en segmentos utilizado, utiliza como mínimo un modelo de mezcla Gaussiana (GMM) para cada clase de interlocutor, que se basa en Coeficientes de Frecuencia Mel-Cepstral (MFCC), que son procesados trama a trama.

9. Dispositivo, según la reivindicación anterior, en el que el número de mezclas está comprendido entre 60 y 170, preferentemente 96.

10. Dispositivo, según una o varias de las reivindicaciones anteriores, en el que mediante la unidad de proceso tiene lugar una combinación del procedimiento de clasificación de interlocutores en base a clasificación del nivel (“Score Level”).

11. Aparato, según una de las reivindicaciones anteriores, en el que en la combinación de SVM y CMM se introducen ponderaciones, y las ponderaciones son determinadas con ayuda de búsquedas específicas de clase.

12. Aparato, según una o varias de las reivindicaciones anteriores, en el que mediante la unidad de proceso u otros medios tiene lugar antes de la clasificación de interlocutor, una extracción de características de los datos de voz.

13. Aparato, según una o varias de las reivindicaciones de aparato anteriores, en el que se dispone de medios para llevar a cabo una combinación de la clasificación de interlocutor con características del aparato terminal, que contacta con un sistema para la clasificación de interlocutores para almacenar informaciones de la clasificación de interlocutor en relación con los identificadores, que se toman en cuenta en una nueva clasificación de interlocutores.

14. Aparato, según la reivindicación anterior, en el que se almacena en una memoria, además de las características de clasificación de interlocutor, datos en brutos digitales de voz.

Fig. 1

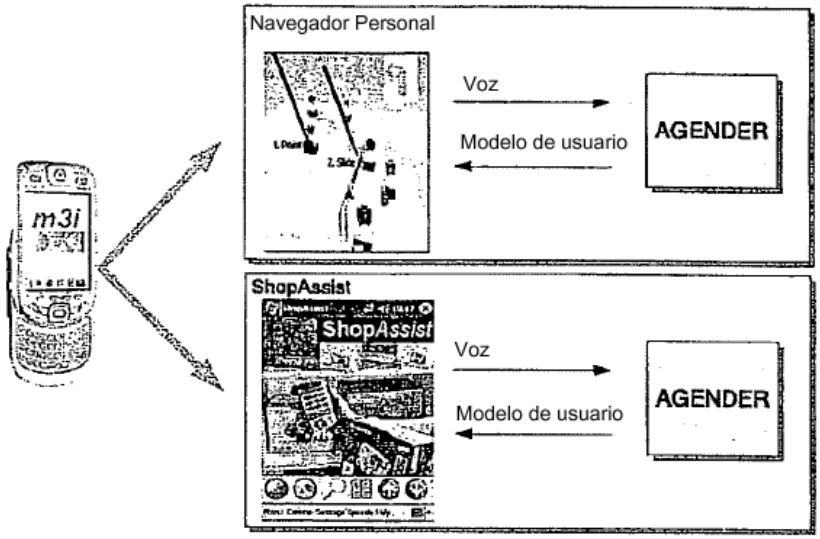

Navegador Personal
Voz
Modelo de usuario
AGENDER
m3i
ShopAssist
ShopAssist
Voz
Modelo de usuario
AGENDER

Fig. 2

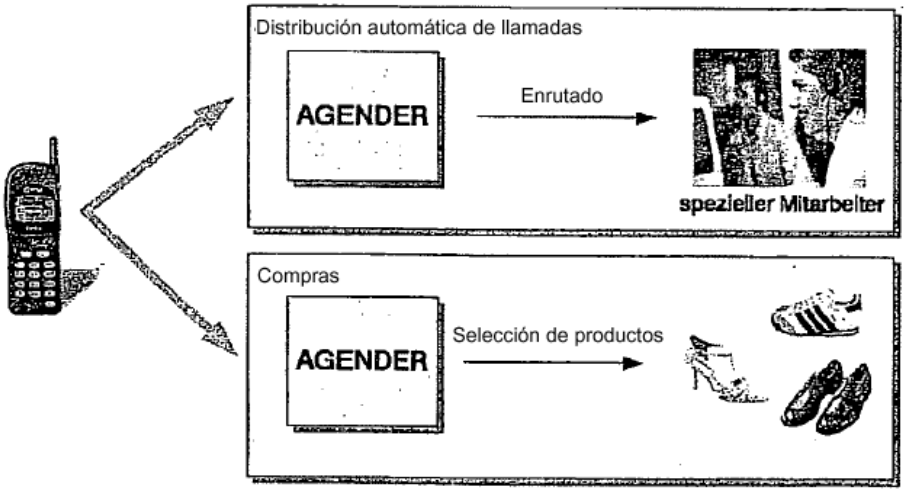

Distribución automática de llamadas
AGENDER
Enrutado
spezieller Mitarbeiter
Compras
AGENDER
Selección de productos

Fig. 3

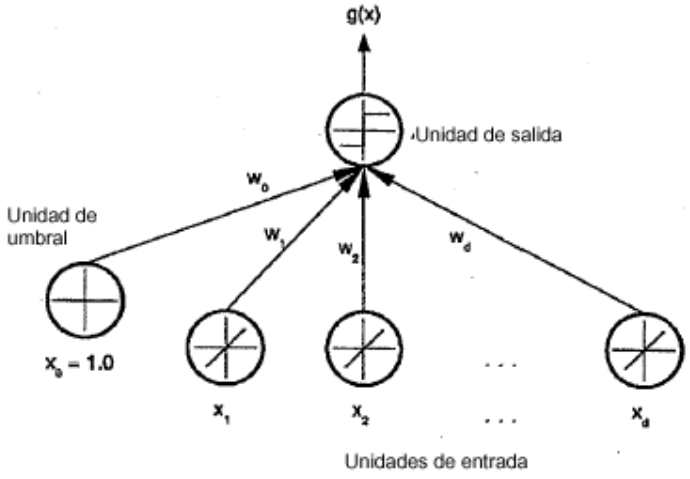


Fig. 4

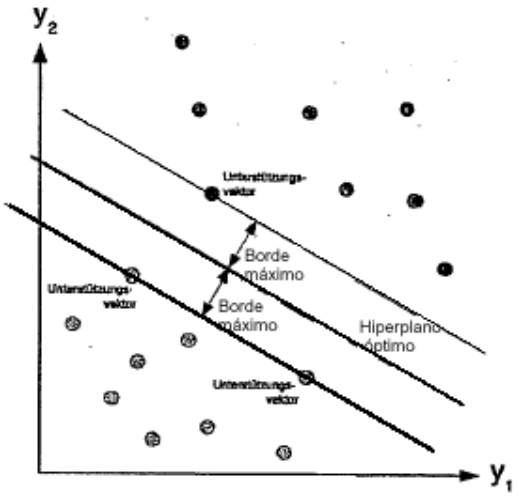


Fig. 5

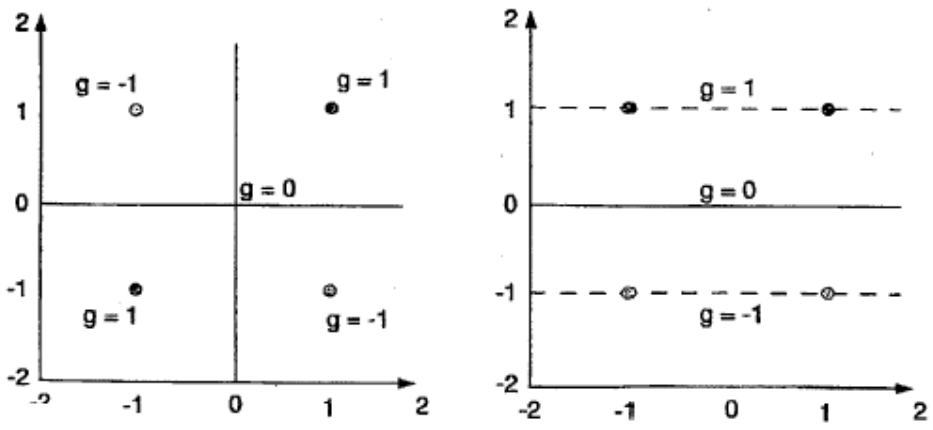


Fig. 6

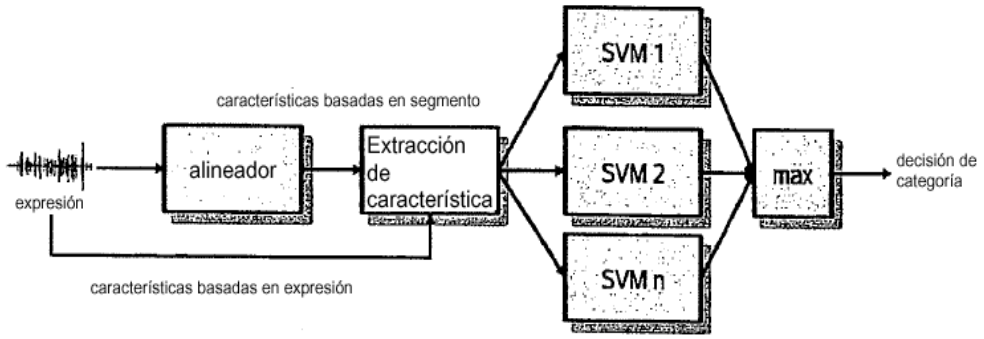


Fig. 7

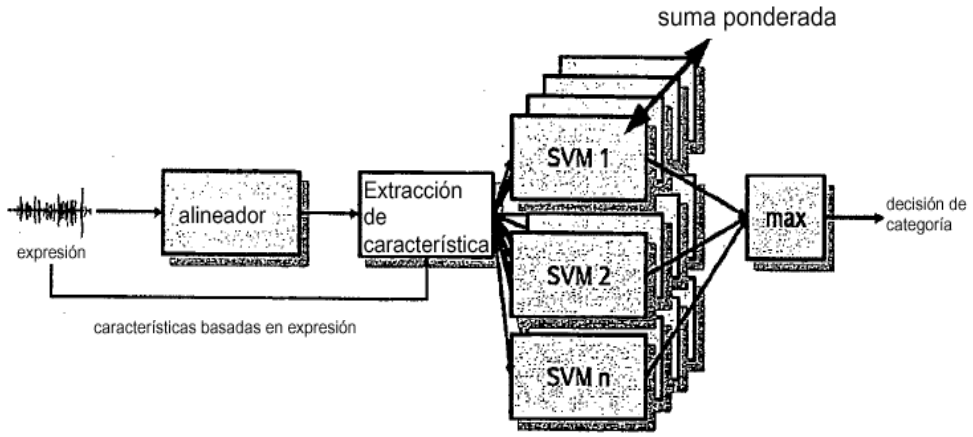


Fig. 8

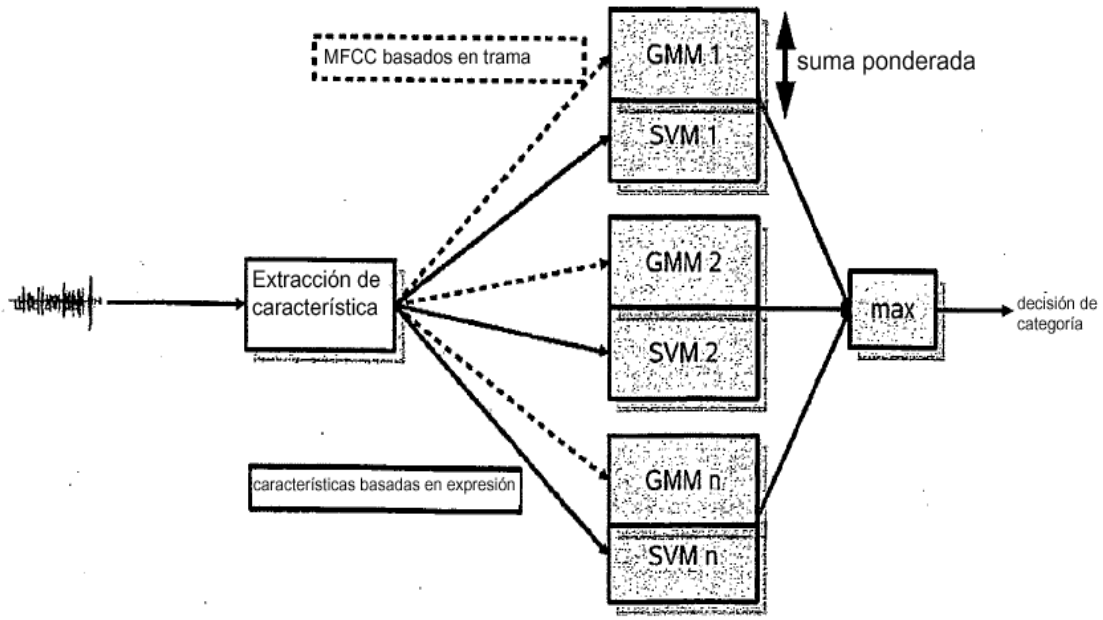


Fig. 9

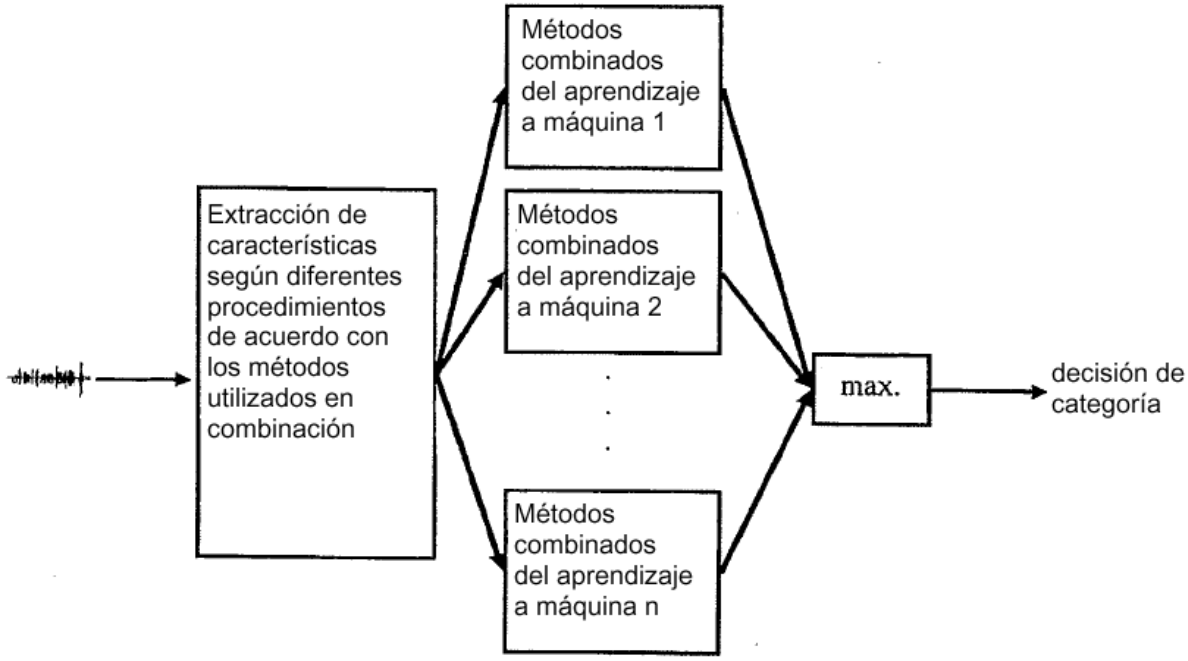

Métodos combinados del aprendizaje a máquina 1
Extracción de características según diferentes procedimientos de acuerdo con los métodos utilizados en combinación
Métodos combinados del aprendizaje a máquina 2
max.
decisión de categoría
Métodos combinados del aprendizaje a máquina n

Fig. 10

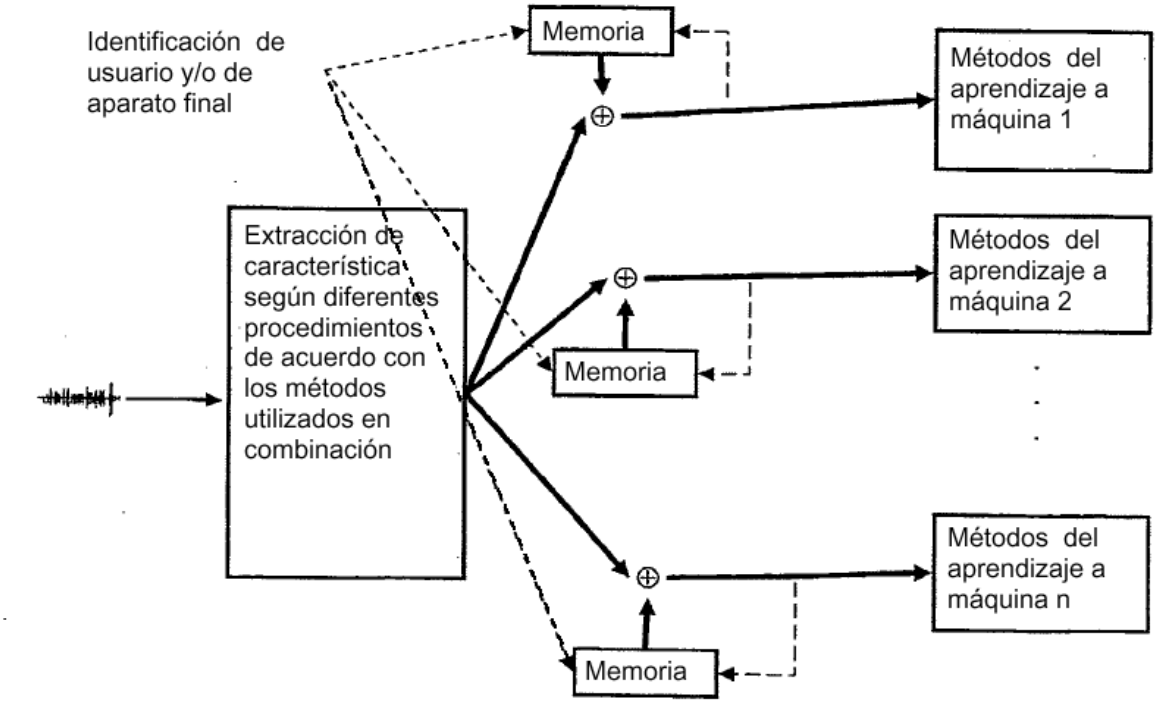

Identificación de usuario y/o de aparato final
Memoria
Métodos del aprendizaje a máquina 1
Extracción de característica según diferentes procedimientos de acuerdo con los métodos utilizados en combinación
Métodos del aprendizaje a máquina 2
Memoria
Métodos del aprendizaje a máquina n
Memoria

Fig. 11

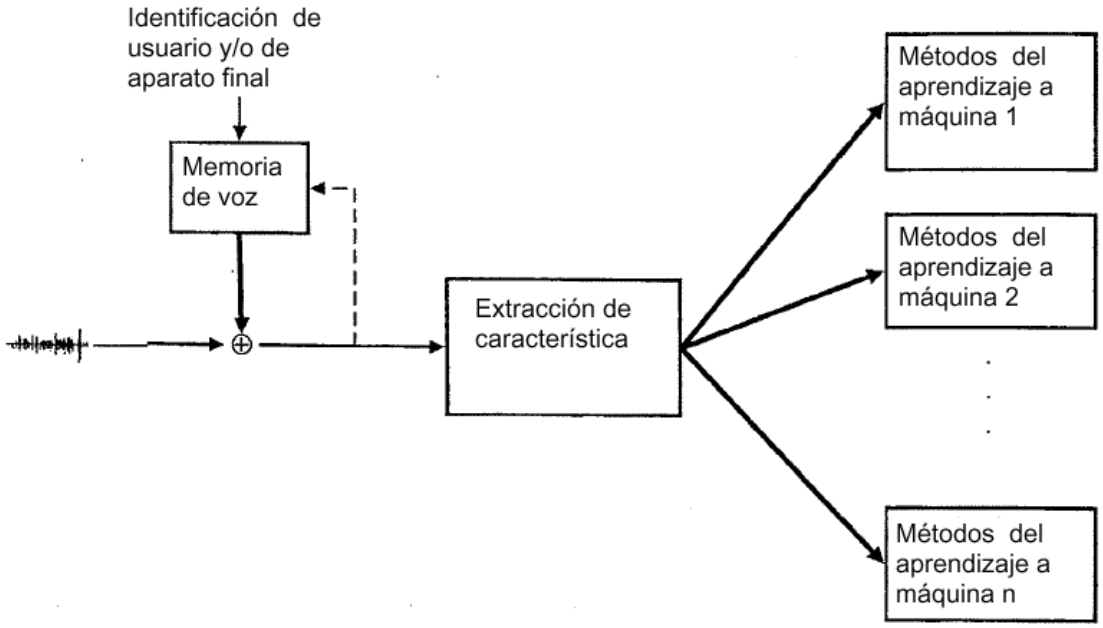
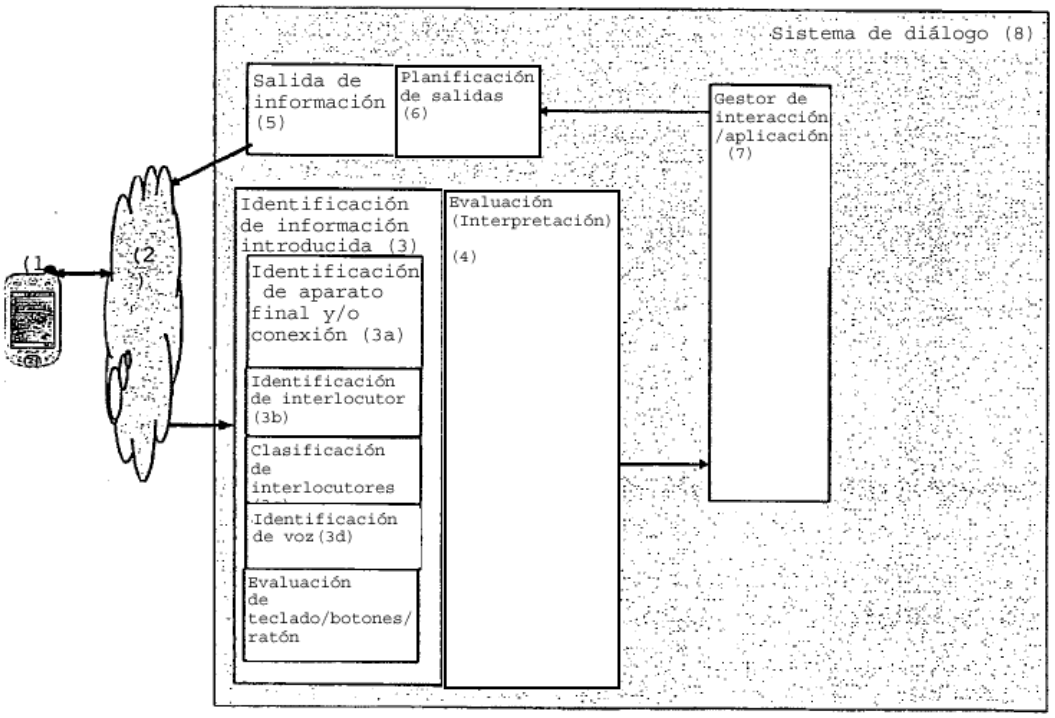

Identificación de usuario y/o de aparato final
Memoria de voz
Extracción de característica
Métodos del aprendizaje a máquina 1
Métodos del aprendizaje a máquina 2
Métodos del aprendizaje a máquina n

Fig. 12


Sistema de diálogo (8)
Salida de información (5)
Planificación de salidas (6)
Gestor de interacción /aplicación (7)
Identificación de información introducida (3)
Evaluación (Interpretación) (4)
Identificación de aparato final y/o conexión (3a)
Identificación de interlocutor (3b)
Clasificación de interlocutores
Identificación de voz (3d)
Evaluación de teclado/botones/ratón
(1
(2

	Exactitud promedio 39%						
	K	JW	JM	EW	EM	SW	SM
K	23	2	49	6	10	0	6
JW	0	31	0	57	4	0	6
JM	10	0	63	0	18	0	6
EW	0	21	0	73	1	1	1
EM	4	0	31	1	42	0	19
SW	0	14	0	61	5	3	15
SM	6	2	23	3	30	0	34

Tab. 2

28.13	20.18	13.5	15.13	12.21	13.76	13.01	14.10	45.67	56.21
media	23.19								
mediana	14.62								

Tab. 3

	Exactitud promedio 39,9%						
	K	JW	JM	EW	EM	SW	SM
K	33.99	38.69	3.64	11.84	2.88	8.5	0.46
JW	15.05	51.18	0.54	21.08	0	11.72	0.43
JM	1.18	1.5	60.64	2.9	23.82	0.43	9.23
EW	12.68	19.58	3.68	37.32	0.74	23.07	2.94
EM	0.3	0	41.6	0.4	36.84	0.3	20.55
SW	10.01	18.25	7.16	26.5	1.47	31.7	4.91
SM	0.53	1.31	38.9	2.37	24.97	4.34	27.6

Tab. 5

	Exactitud promedio 43,7%						
	K	JW	JM	EW	EM	SW	SM
K	42.94	20.03	18.21	7.28	6.07	2.58	2.88
JW	19.89	46.45	2.26	23.66	0.11	6.24	1.4
JM	0.43	0	28.11	3	59.66	1.29	7.51
EW	8.73	16.73	2.48	51.19	0.18	19.49	1.19
EM	0.1	0	1.72	2.83	90.18	1.11	4.05
SW	11.19	16.39	2.94	40.43	0.2	27.67	1.18
SM	0.13	0	11.56	2.37	63.34	3.02	19.58

Tab. 7

	Exactitud promedio 49,11%						
	K	JW	JM	EW	EM	SW	SM
K	41.73	42.64	0.3	7.89	6.37	0.76	0.3
JW	13.12	62.26	0.32	16.56	0	7.74	0
JM	3.65	1.39	54.94	3	28.76	1.07	7.19
EW	6.16	21.72	3.03	43.29	0.92	23.99	1.19
EM	0.61	0.1	18.12	0.3	70.75	0.1	9.82
SW	4.32	21.2	4.32	35.43	0.49	31.99	2.26
SM	4.6	1.97	20.24	1.84	27.6	5.12	38.63

Tab. 9