Desarrollo de un asistente domótico emocional inteligente

C. Sanz-Moreno, S. Lutfi, R. Barra-Chicote, J.M. Lucas, J.M. Montero

Univ. Politécnica de Madrid {csmoreno,syaheerah,barra,juanmak,juancho}@die.upm.es

***Abstract* — La abundancia de aparatos electrónicos multiplica el número de interfaces que el usuario doméstico debe conocer y manejar. Para mejorar esta situación se ha desarrollado un asistente domótico que permite controlar por voz múltiples equipos eléctricos y electrónicos. A fin de incrementar su aceptación, se ha creado una cabeza robótica que personifica al asistente y que tiene capacidad de reaccionar emocionalmente a las situaciones para conseguir mayor empatía con el usuario**

I. Introducción

En los últimos años el número de dispositivos electrónicos ha aumentado considerablemente en nuestros hogares, aunque las dificultades de algunos colectivos para adaptarse a la tecnología no se ha reducido significativamente, lo cual demuestra la necesidad de crear aplicaciones que nos asistan de manera inteligente. En este contexto se ha trabajado sobre líneas de investigación que contribuyan a acercar personas y máquinas: las interfaces, más amigables; la aplicación, un asistente domótico que proporcione verdadera utilidad; y la emotividad, que el usuario sea capaz de interaccionar con un sistema que se adapte en vez de mostrarse como una máquina monótona y desesperante.

II. Interfaz

Actualmente cualquier aparato electrónico que se pueda adquirir viene con un manual de instrucciones para aprender a utilizarlo. Esto es un claro reflejo de que la interfaz, si bien puede estar simplificada, es más cercana a la máquina que a la persona y por tanto requiere un proceso de adaptación.

Ante esto surge la idea de una interfaz próxima al usuario con el objetivo de que no tenga que adaptarse a la máquina ya que el sistema es capaz de interpretar el lenguaje natural de las personas [1]. En este sentido el reconocimiento y la síntesis de voz juegan un papel muy importante. De hecho, una vez que se aprende a hablar, el lenguaje verbal es la principal forma de comunicación de las personas de modo que, si se crea una base de datos de tamaño y libertad de reconocimiento suficientes, el usuario no tendrá que aprender ningún comando ya que lo que solicitará lo hará de manera intuitiva. Por ejemplo, si el usuario desea encender una luz, podrá decir: “enciende la luz” o “pon la luz” o frases similares que no son distintas a lo que le diría a otra persona.

No obstante, se ha comprobado que un sistema en el que exista una única forma de comunicación presenta dificultades a la hora de adquirir información suficiente del entorno para tomar decisiones. Ante esto surge la idea del empleo de interfaces multimodales. Este tipo de interfaces se basan en la integración de distintos tipos de fuentes de información. En este caso, además de utilizar la voz asociada a síntesis, a reconocimiento y a identificación de usuarios, se utiliza una interfaz visual basada en la librería de tratamiento de imágenes OpenCV para la entrada de datos y la pantalla como salida. También se emplea una interfaz física consistente en una cara y un brazo robóticos. El brazo es utilizado para interaccionar con el usuario en determinadas aplicaciones como los juegos, mientras que la cara es un símbolo del sistema cuya función es que las personas que tienen dificultades, o incluso a las que les desagrada el uso de un ordenador, puedan dirigirse a algo físico con un cierto parecido a las personas.

III. Aplicaciones

En esta línea de trabajo el objetivo es conseguir una *killer aplication* que muestre a los usuarios el efecto que tienen los sistemas electrónicos en el incremento del rendimiento de cualquier actividad. Para perseguir este objetivo se ha implementado la funcionalidad de asistente domótico [2]. Este tipo de asistentes se pueden asemejar a los electrodomésticos y, por tanto, parece viable conseguir situar el grado de aceptación de un sistema de esta clase al nivel de una lavadora o un frigorífico, es decir, aparatos que están presentes en la inmensa mayoría de las casas y que son utilizados por todos.

Para realizar la función de asistente domótico se ha creado una arquitectura escalable basada en tareas. Cada tarea es una acción o conjunto de acciones necesarias para conseguir un objetivo concreto. Por ejemplo, una tarea es jugar a un juego de mesa, otra es controlar un equipo de música, etc.

TABLA I: RELACIÓN DE TECNOLOGÍAS UTILIZADAS PARA APLICACIONES DOMÓTICA

Tecnología	Utilidad	Elementos Necesarios
X10	Control remoto de luces	Controlador Marmitek CM11 Dispositivo Marmitek LM15
Bluetooth	Control remoto de robot aspiradora Roomba	Módulo Bluetooth para PC Módulo Rootooth para Roomba Robot aspiradora Roomba
Infrarrojos	Control remoto de aparatos electrónicos mediante infrarrojos.	Módulo iRTrans

En esta arquitectura son las propias tareas quienes, conociendo la información del entorno, son capaces de generar una respuesta acorde a sus objetivos específicos de tarea. De esta forma el modelo de comportamiento se comporta como un *hub* que implementa una sencilla lógica para resolver conflictos, en el caso de que distintas tareas realicen acciones contradictorias, y priorizar la respuesta en función de algún tipo de criterio. Para la comprensión automática de habla se emplea un módulo de aprendizaje de reglas basado en ejemplos [1], lo cual facilita la incorporación de nuevas tareas sin tener que reescribir el sistema. Las tareas que han sido implementadas en el modelo de prueba [2] son: control de la aspiradora robótica *Roomba* mediante *bluetooth*, control de una lámpara mediante protocolo X10, control de un equipo HiFi mediante infrarrojos y dos juegos de tres en raya (uno físico mediante brazo robótico y otro virtual mediante una pantalla).

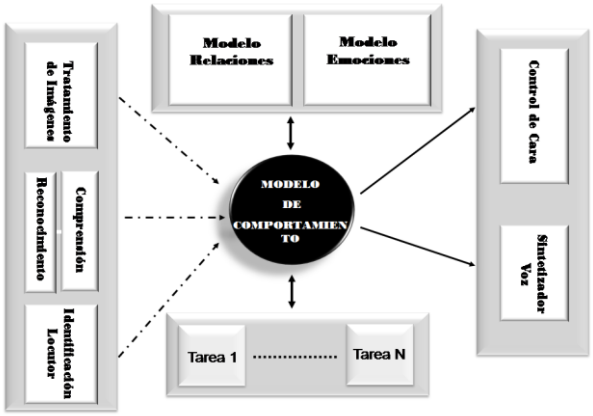

Fig. 1. Representación esquemática de la arquitectura del sistema.

IV. EMOTIVIDAD

Constituye la piedra angular de este sistema. Nace de la idea de elevar las máquinas a la categoría de compañeros domóticos que tendrán más aceptación por parte de los usuarios si se puede interaccionar con ellos de una forma más cercana. Dos son los objetivos; el primero está recogido en el modelo de relaciones y trata sobre la adaptación del sistema a su relación con el usuario; el segundo objetivo es más ambicioso y trata de generar artificialmente emociones de modo que el sistema tenga un comportamiento cercano al humano para realizar tareas, como por ejemplo los juegos, de una manera más natural.

A. Modelo de relaciones.

El modelo de relaciones consiste en la definición de una serie de estados que describen cuál es la relación entre el sistema y el usuario según el historial de interacciones. Aunque el número de estados que se pueden definir es enormemente grande, en el caso implementado se han definido solamente tres estados (desconocido, conocido y amigo). Cada uno de estos estados tiene asociado un tipo de comportamiento distinto. En el sistema implementado esta variación queda perfectamente reflejada en la elección de las frases que el sistema dirija al usuario. Es decir, ante un saludo el sistema respondería: "Hola, ¿Le conozco?" en el caso en el que el usuario fuese desconocido, "Hola, ¿Cómo estás?" en el caso de un usuario conocido y "¿Qué tal?" en el caso de un usuario amigo. Nótese que para realizar este modelo se requiere una forma de identificar al usuario que, como se mencionó anteriormente, en este caso es por voz, empleando un módulo del Grupo de Tecnología del Habla [3] que hemos integrado y adaptado hasta conseguir una tasa de acierto superior al 90% si hubiese hasta 30 locutores.

B. Modelo de emociones.

TABLA II: ATRIBUTOS DE EVALUACIÓN Y SUS FÓRMULAS EN FUNCIÓN DEL VALOR ACTUAL Y ANTERIOR DE UNA NECESIDAD Y DE SU VALOR CRÍTICO (VC)

Atributo de evaluación	Descripción de su significado
Deseabilidad	Contribución del evento a conseguir el objetivo.
Inesperado	Indica la diferencia entre el evento ocurrido y el anterior.
Relevancia	Importancia del evento
Urgencia	Tiempo estimado hasta alcanzar una situación crítica.
Infrecuente	Probabilidad de que el evento no hubiese ocurrido

$$Deseabilidad = Valor_n - Valor_{n-1}$$

$$inesperado \left| \frac{(Valor_n - Valor_{n-1}) - (Valor_{n-1} - Valor_{n-2})}{2} \right|$$

$$relevancia \begin{cases} 100, & Valor_n < VC \\ 100 * \left(\frac{100 - Valor_n}{100 - VC} \right), & Valor_n \geq VC \end{cases}$$

$$Infrecuencia = 100 * \left(1 - e^{-\left| \frac{Valor_n - Media}{Desviación} \right|} \right)$$

$$urgencia \begin{cases} 100, & Valor_n \leq VC \text{ ó } (Valor_n + Deseabilidad) < VC \\ \left| (100 - Valor_n) * \left(\frac{Valor_n - Valor_{n-1}}{Valor_n - VC} \right) \right|, & Valor_n > VC \end{cases}$$

El modelo de emociones se encarga del análisis de la situación y del cómputo del estado emocional en cada momento. Para la realización del modelo emocional se ha partido de las teorías psicológicas de evaluación o *appraisal* [4], que establecen que en un entorno se generan eventos que son emocionalmente neutros, pero que son enjuiciados por las personas en función de sus predisposiciones y necesidades personales, produciendo como consecuencia una alteración del estado emocional del sujeto.

Según esta teoría, cada tarea del sistema debe ser capaz de analizar todos los eventos que le resulten relevantes y caracterizar su efecto para la consecución del objetivo de la tarea mediante unos atributos de evaluación. Aunque la teoría sicológica propone numerosos atributos, no se ofrecen fórmula o algoritmos que permitan calcularlos en una implementación. Teniendo en cuenta las emociones expresables por la cara robótica y el sintetizador de habla de este sistema (tristeza, alegría, enfado, sorpresa, miedo y neutral), se han seleccionado los atributos de la tabla II y se han creado fórmulas para su cálculo. Los valores de estos atributos sirven de entrada a un modelo de variables de estado del espacio emocional, que modela los cambios del estado emocional debidos a los eventos evaluados y al paso del tiempo.

Sin embargo, la teoría de la evaluación presenta problemas a la hora de ser aplicada en un sistema multitarea. En concreto implica que cada tarea debe realizar un análisis emocional de modo que la emoción final del sistema responde directamente a la consecución de un determinado objetivo específico. Es decir, el sistema se alegra si, por ejemplo, consigue un buen movimiento mientras juega. Esto provoca que sea difícilmente escalable, pues para añadir una nueva tarea debe ajustarse la totalidad del modelo emocional. Frente a dicho problema se aborda la idea de que las emociones surgen de la satisfacción de necesidades más genéricas. Es decir, el sistema no se alegra por hacer un buen movimiento mientras juega, sino que se alegra porque por medio de ese movimiento, se contribuye a satisfacer su necesidad de reconocimiento social o éxito.

Para establecer estos necesidades genéricas (supervivencia, seguridad, integración social, éxito y auto-realización ética) se ha utilizado la teoría de Abraham Maslow que establece que las personas están motivadas por el deseo de satisfacer una serie de necesidades jerarquizadas (pirámide de Maslow). De este modo la felicidad se alcanza cuando todas y cada una de las necesidades se satisfacen. Incluyendo este concepto, el sistema emocional completo se puede apreciar en la figura 2.

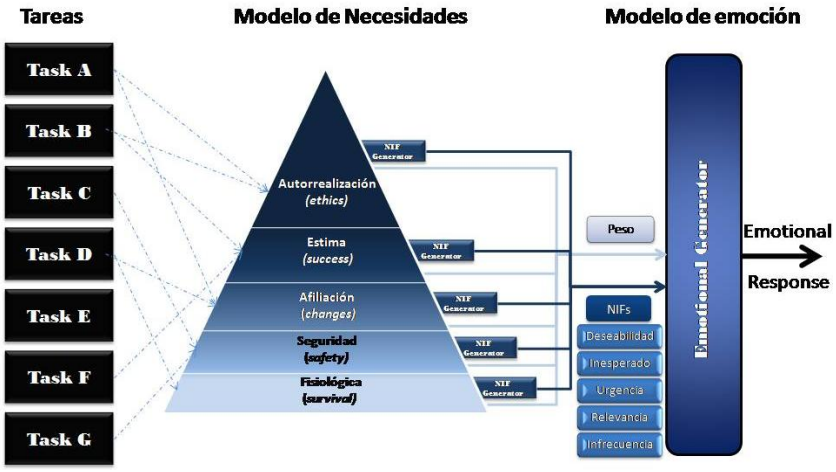

Fig. 2. Arquitectura emocional.

En este modelo las tareas cambian el nivel de satisfacción de las necesidades en función de los eventos detectados. Esta variación es evaluada mediante un vector de evaluación (conjunto de atributos de la tabla I), que se pasa al calculador de emociones junto a un valor asociado a la prioridad del nivel de Maslow. Este módulo determina los nuevos niveles de emociones a partir del estado emocional anterior, su amortiguación en el tiempo y un componente dependiente del vector de evaluación. Véase su funcionamiento con un ejemplo de la tarea de 3 en raya.

- En un primer momento un movimiento se produce y la tarea del juego analizará qué implicaciones tiene para la partida. Las posibles consecuencias afectarán a ciertos niveles. En este sentido si el movimiento conducía a que el sistema ganase la partida sería un evento que afectaría al nivel de estima, mientras que si el evento era considerado como una trampa se relacionará con el nivel de autorrealización. Por último, si el movimiento no tiene ninguna consecuencia para el sistema será asociado con el nivel de changes.

- Las modificaciones de los niveles de Maslow deben ser caracterizadas mediante un vector de peso asociado al nivel de prioridad de Maslow y mediante un vector de evaluación. De esta forma, si el sistema tendía a perder y de repente gana, la modificación del nivel de necesidad de estima podrá ser evaluada como deseable (ha ganado) e inesperada (el sistema solía perder).

- Finalmente, a partir de la información de los distintos vectores de evaluación asociados a cada uno de los niveles de Maslow, el generador emocional actualiza el valor de cada uno de los niveles emocionales del sistema determinando cuál es la emoción dominante, que constituirá la respuesta emocional.

Gracias a esta arquitectura el sistema es fácilmente escalable ya que a la hora de incorporar una nueva tarea lo único que hay que realizar es definir a qué necesidades afecta dicha tarea de modo que la computación emocional permanece invariable. Otra ventaja es que el cálculo emocional se basa en necesidades cuantitativas (de modo que permiten la elaboración de un modelo matemático para la obtención de los atributos de evaluación) y priorizadas (de forma que la contribución a la emoción final de cada nivel está ponderada en función de la jerarquía de Maslow). Esto significa que una emoción de tristeza o miedo provocada por un problema en el nivel básico (como que se agote la batería), será predominante frente a una emoción de alegría de un nivel elevado (como ganar en un juego).

V. Resultados

El resultado obtenido es un sistema domótico [2] capaz de realizar las acciones que le sean solicitadas por voz pero que además expresa emocionalmente: se entristece cuando tiene problemas para comprender a las personas o cuando se despiden de él; que se alegra cuando resulta útil o se le hace una caricia en uno de sus sensores; que puede enfadarse cuando pierde en alguno de los juegos que conoce o si se le insulta; y que incluso puede tener miedo cuando no puede ver debido a la oscuridad o al mal funcionamiento de alguno de sus módulos. Además las emociones están modeladas por niveles individuales de forma que el

sistema es capaz de realizar transiciones de un estado emocional a otro de forma continua. Por ejemplo, si estaba enfadado y se le hace una caricia el nivel de enfado se reduce y el de alegría aumenta pudiendo ser incluso mayor que el de enfado, pero queda una componente de enfado que permite matizar la expresión de emociones.

En este sentido, la síntesis de habla mediante selección de unidades es la forma idónea de reflejar este comportamiento interno ya que en un experimento de identificación de emociones en voz sintética generada por los métodos modelos ocutlos de Markov (HMM)[5] y selección de unidades [6] que hemos adaptado, se obtuvieron unos resultados excelentes en tanto por ciento de emociones identificadas al escuchar frases con texto neutro y habla sintética emocional. El mejor método para sintetizar alegría, enfado y tristeza es el de selección de unidades (por su capacidad de imitar la calidad de voz original), mientras que el basado en el HMM (HTS) es mejor para sorpresa, miedo y asco (por su robustez al modelar la prosodia emotiva).

TABLA III: RESULTADOS DE PRUEBAS SOBRE SÍNTESIS DE HABLA CON EMOCIONES

	Alegría	**Enfado**	**Sorpresa**	**Tristeza**	**Miedo**
Selección de unidades	64	81	40	91	26
HTS	44	67	57	64	37
Habla natural	80	79	82	81	50

Por otro lado, se ha utilizado un modelo matemático adaptativo para computar los eventos. Esto permite que aquellos que sean muy frecuentes tengan una contribución pequeña frente a los que son poco frecuentes. De esta forma si se está acariciando al asistente de manera reiterada, el sistema estará alegre, pero se irá acostumbrando a las caricias; si en ese momento se le insulta, el insulto, al ser mucho menos frecuente e inesperado que las caricias, provocaría que el sistema se enfadase.

VI. Conclusión

Se ha desarrollado un asistente personal adaptativo [2] con capacidad para entender habla con independencia del locutor y basándose en ejemplos y con capacidad de expresar oralmente emociones. La tasa de identificación del habla sintética con emociones es 66% en frases cortas con texto neutro.

Agradecimientos

Este proyecto ha sido financiado y apoyado por el Grupo de Tecnología del Habla (DIE-ETSIT-UPM), la cátedra Indra-Fundación Adecco (ETSIT-UPM) y el proyecto DPI2007-66846-C02-02 (Ministerio de Cª e Innovación). Agradecemos también su apoyo al Grupo de Control Inteligente (ETSII-UPM).

Referencias

[1] J.M. Lucas Cuesta et al, "Desarrollo de un Robot-Guía con Integración de un Sistema de Diálogo y Expresión de Emociones: Proyecto ROBINT" *Procesamiento del Lenguaje Natural*, Vol 40 pp. 51-5, March 2008.

[2] http://www.youtube.com/watch?v=YtLdZ8wkq4E, July 2009.

[3] J. Ferreiros, D.P.W Ellis, "Using Acoustic Condition Clustering to Improve Acoustic Change Detection on Broadcast News" *Proceedings of ICSLP*, 2000.

[4] K.R. Scherer, A. Shorr and T. Johnstone (Ed.), *Appraisal processes in emotion: theory, methods, research*, Canary, NC: Oxford University Press, 2001.

[5] R. Barra-Chicote et al, "Generación de una voz sintética en castellano basada en HSMM para la Evaluacion Albayzin 2008: conversión texto a voz," *V Jornadas en Tecnología del Habla*, pp. 115-118, November 2008.

[6] R.A.J. Clark, K. Richmond and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system", *Speech Communication*, Vol. 49(4) pp. 317-330, 2007.