Procesamiento del Lenguaje Natural, Revista nº 40, marzo de 2008, pp. 51-58 recibido 30-01-08, aceptado 03-03-08

Desarrollo de un Robot-Guía con Integración de un Sistema de Diálogo y Expresión de Emociones: Proyecto ROBINT

Development of a Tour-Providing Robot Integrating Dialogue System and Emotional Speech: ROBINT Project

Juan Manuel Lucas Cuesta, Rosario Alcázar Prior, Juan Manuel Montero Martínez, Fernando Fernández Martínez, Roberto Barra-Chicote, Luis Fernando D'Haro Enríquez, Javier Ferreiros López, Ricardo de Córdoba Herralde, Javier Macías Guarasa, Rubén San Segundo Hernández, José Manuel Pardo Muñoz
Grupo de Tecnología del Habla, UPM
Avenida Complutense s/n. 28040. Madrid
juanmak@die.upm.es, ralcazar@die.upm.es, juancho@die.upm.es, efhes@die.upm.es, barra@die.upm.es, lfdharo@die.upm.es, jlf@die.upm.es, cordoba@die.upm.es, macias@die.upm.es, lapiz@die.upm.es, pardo@die.upm.es

Resumen. Este artículo presenta la incorporación de un sistema de diálogo hablado a un robot autónomo, concebido como elemento interactivo en un museo de ciencias capaz de realizar visitas guiadas y establecer diálogos sencillos con los visitantes del mismo. Para hacer más atractivo su funcionamiento, se ha dotado al robot de rasgos (como expresividad gestual o síntesis de voz con emociones) que humanizan sus intervenciones.

El reconocedor de voz es un subsistema independiente del locutor (permite reconocer el habla de cualquier persona), que incorpora medidas de confianza para mejorar las prestaciones del reconocimiento, puesto que se logra un filtrado muy importante de habla parásita. En cuanto al sistema de comprensión, hace uso de un sistema de aprendizaje basado en reglas, lo que le permite inferir información explícita de un conjunto de ejemplos, sin que sea necesario generar previamente una gramática o un conjunto de reglas que guíen al módulo de comprensión. Estos subsistemas se han evaluado previamente en una tarea de control por voz de un equipo HIFI, empleando nuestro robot como elemento de interfaz, obteniendo valores de 95,9% de palabras correctamente reconocidas y 92,8% de conceptos reconocidos.

En cuanto al sistema de conversión de texto a voz, se ha implementado un conjunto de modificaciones segmentales y prosódicas sobre una voz neutra, que conducen a la generación de emociones en la voz sintetizada por el robot, tales como alegría, enfado, tristeza o sorpresa. La fiabilidad de estas emociones se ha medido con varios experimentos perceptuales que arrojan resultados de identificación superiores al 70% para la mayoría de las emociones, (87% en tristeza, 79,1% en sorpresa).
Palabras clave: reconocimiento de habla, medidas de confianza, síntesis de voz con emociones.

Abstract. This paper describes the implementation of a spoken dialogue system on an autonomous robot which presents a high degree of interaction with the visitors in a Science Museum, providing interactive guided tours. Our main purpose was to provide the robot with some features towards the generation of more human-like interaction. These features are gestual expressivity and emotional speech synthesis.

The speech recognition module is a speaker-independent recognizer which makes use of confidence measures, achieving the recognition of utterances spoken by any person, and a high reduction of the impact of noise in speech. The language understanding module makes use of a self-learning rule-based approach, which allows the system to infer information from the available example utterances. Thus, the generation of a formal grammar becomes unnecessary. Both modules have been evaluated on a task which includes dialogues between our robot and a

ISSN 1135-5948

human speaker. This task has been the control of a HI-FI system. The results of this experiment are 95.9% in Word Accuracy, and 92.8% in Concept Accuracy.

We have also implemented a voice synthesizer that makes use of several prosodic and segmental modifications of the synthesized speech. This way, our system generates a speech with several emotions, such as happiness, anger, sadness or surprise. The performance of this module has been measured with several experiments for emotion identification, that show identification rates higher than 70% for most of tested emotions, (87% for sadness, or 79.1% for surprise).

Keywords: speech recognition, confidence measures, emotional speech synthesis.

1. Introducción

La interacción entre seres humanos y máquinas ha pasado de ser un paradigma de investigación a convertirse en la actualidad en una realidad que se da en diferentes niveles. El nivel de interacción más básico, más próximo a la máquina que al hombre, lleva décadas siendo usado (a través de dispositivos como teclados, generando comandos que la máquina debe interpretar). Sin embargo, el campo más interesante es el desarrollo de plataformas que permitan una interacción a niveles más próximos a los que el ser humano emplea de manera intuitiva, tales como el uso de la voz o la expresión corporal.

Si se concibe la interacción persona-máquina como el establecimiento de una comunicación entre un ser humano y un robot, aparecen robots que desempeñan tareas con un elevado número de interacciones con seres humanos diferentes de sus programadores. Así, en (Fong, Nourbakhsh y Dautenhahn, 2003), se definen los *robots sociales* como aquellos robots en los que la interacción persona-máquina adquiere un nivel relevante. En la actualidad, tales robots se encuentran todavía en una fase de investigación, si bien se pueden encontrar ya implantados en determinados contextos, entre los que destacan su empleo como guías en museos (Willeke, Kunz y Nourbakhsh, 2001), (disam, 2008) o para la rehabilitación de niños hospitalizados (Plaisant et al., 2000), (Saldien et al., 2006).

En función de la complejidad del escenario en el que se produce la interacción, (Breazeal, 2003) clasifica los robots sociales en cuatro grupos: *socialmente evocativos*, *robots de interfaz social*, *socialmente receptivos*, y *sociables*. Atendiendo a las características de esta clasificación, nuestro robot puede clasificarse dentro del tipo *socialmente receptivo*, pues ha de permitir la interacción natural con los visitantes del museo, además de responder de manera distinta ante diferentes intervenciones de dichos visitantes.

Se quiere orientar el funcionamiento del robot hacia uno de los grupos mayoritarios de visitantes de un museo, como puede ser el formado por niños en edad escolar. Los motivos que nos impulsan a tener en cuenta este sector son varios. En primer lugar, es un sector de población en el que las intervenciones habladas son más espontáneas. Además, los grupos de escolares suelen hacer este tipo de excursiones de manera obligada, por lo que resulta complicado mantener la atención de los mismos durante toda la visita, en especial si durante la misma se producen presentaciones excesivamente prolongadas (Willeke, Kunz y Nourbakhsh, 2001).

En la actualidad ya existen robots capaces de interactuar con niños. Se trata sobre todo de sistemas de terapia de niños hospitalizados (Plaisant et al., 2000, Saldien et al., 2006, Shibata et al., 2001) o que presentan problemas en su comportamiento, como autismo (Dautenhahn y Werry, 2000). Estos robots suelen tener la forma de animales de compañía, con una serie de sensores y actuadores que permiten que los robots respondan a los estímulos producidos por la actividad de los niños. En cuanto a sistemas con capacidad de narrar una historia, (Silva, Vala y Paiva, 2001) desarrollan un agente virtual, mientras que (Druin et al., 1999), o (Plaisant et al., 2000), analizan un robot con capacidad de contar cuentos, aplicado en un contexto de rehabilitación pediátrica. En nuestro caso, el sistema cuentacuentos contará con un nivel expresivo mayor, gracias a su expresión de emociones, tales como la alegría, la tristeza o el enfado, a través de la voz, de tal manera que dicha emoción pueda ser percibida por los niños a lo largo de las intervenciones del robot.

Se pretende, por tanto, dotar al robot de la capacidad de reconocer el habla de cualquier persona, y de generar habla sintética

expresiva. A mayor nivel, se pretende que el robot pueda narrar historias, modificando la voz emitida de acuerdo al contexto de la narración, o bien en función de las intervenciones de sus interlocutores humanos.

Este artículo se estructura como sigue. La sección 2 presenta la plataforma física que soporta las estructuras de la cara y el brazo, así como el sistema de localización del robot. La sección 3 está dedicada a los bloques que componen el sistema de diálogo, y las pruebas realizadas sobre los mismos. La sección 4 presenta las conclusiones extraídas del trabajo realizado, además de plantear posibles líneas futuras de investigación.

2. Arquitectura física y sistema de guiado

El robot consta de una plataforma móvil sobre la cual se ha construido una estructura que da soporte a la cara y el brazo de nuestro robot. El desplazamiento que se puede aplicar a los párpados, labios y brazo puede ser modificado de acuerdo a la emoción que se desee expresar, por ejemplo elevando las cejas para indicar sorpresa, o frunciendo los labios para denotar tristeza.

La estructura lleva dos procesadores empotrados. El primero se encarga de las tareas de guiado, construcción del mapa y movimiento del robot. Para ello, hace uso de una técnica conocida como SLAM (Localización y Mapeo Simultáneos), desarrollada en (Rodríguez-Losada, 2004) y (drodri, 2008), que le permite determinar su posición en tiempo real.

El segundo equipo lleva a cabo parte de las tareas de diálogo. El resultado de la síntesis de voz se obtiene a través de dos altavoces incorporados a la plataforma. Adicionalmente, se emplea un ordenador portátil al que se conecta un micrófono, y en el cual se ejecuta el módulo de reconocimiento de voz. La comunicación entre el equipo portátil y el robot se lleva a cabo mediante sockets a través de un enlace Ethernet de radio.

3. Sistema de diálogo

El objetivo de un sistema de diálogo es establecer una interacción hablada con un interlocutor humano con una finalidad doble: por un lado, interpretar la intervención del usuario para identificar los servicios que éste solicita, y por otro, prestar dichos servicios y ofrecer al usuario información acerca del resultado de los mismos. En nuestro caso, que la visita por el museo se desarrolle de manera satisfactoria, no restringiéndose a la visita, sino incluyendo otras actividades didácticas, tales como juegos o relatos educativos.

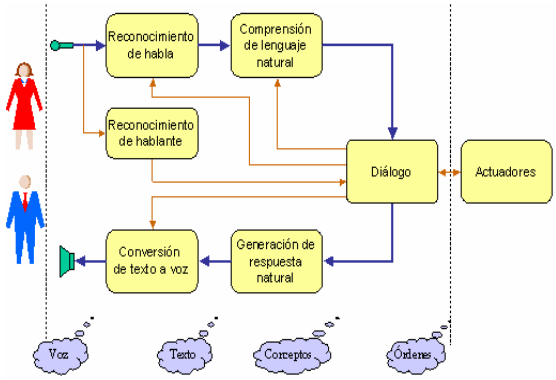

Figura 1: arquitectura de un sistema de diálogo

Los bloques que constituyen un sistema de diálogo son el módulo de *reconocimiento de habla*, que determina la transcripción escrita de la frase enunciada por el hablante, y la evalúa mediante la estimación de una serie de medidas de confianza; el sistema de *comprensión de lenguaje natural*, que extrae los conceptos relevantes del texto anterior; el *gestor de diálogo*, que determina las acciones a realizar a partir de los conceptos extraídos, y genera los conceptos de salida hacia el usuario; el bloque de *generación de respuesta*, que genera un texto comprensible con los conceptos del gestor de diálogo; y el *conversor de texto a voz*, que genera una locución que reproduce el texto que le entrega el generador de respuesta.

3.1. Reconocimiento de habla

El módulo de reconocimiento de habla permite reconocer habla en castellano e inglés, pero en el presente proyecto sólo se empleará el sistema en castellano.

En un primer momento se debe determinar si se dispone de alguna señal acústica válida a la entrada del sistema, es decir, si el micrófono está recibiendo algo diferente al eventual ruido ambiente. En caso afirmativo, se extraen los parámetros significativos de la señal (Huang, Acero y Hon, 2001), mediante el análisis trama a trama de la misma, y el cálculo de los coeficientes perceptuales de predicción lineal (PLP) y la energía de la señal en cada trama, más sus correspondientes derivadas de primer y segundo orden, dando lugar a un vector de 39 parámetros para cada trama.

El reconocedor de habla es de desarrollo propio, basado en modelos ocultos de Markov (HMM) de tres estados por alófono. Se hace uso de un modelo de lenguaje que contribuye a limitar el número de hipótesis entre las que el reconocedor ha de optar en cada instante para determinar cuál es la secuencia de palabras más probable que se está recibiendo. El modelo empleado actualmente se basa en bigramas, es decir, se modela la probabilidad de aparición de cada palabra condicionada a la aparición de la anterior.

Un avance importante con respecto al proyecto URBANO ha sido el empleo de micrófonos de habla cercana (close-talk) en la obtención de la señal acústica, que ha permitido, por un lado, una reducción significativa del ruido ambiente (de unos 45 dB a unos 30 dB) y, por otro, de una menor aparición de errores de tipo “false match” (determinar que hay una señal acústica a la entrada cuando sólo hay ruido ambiente), que hacen que el reconocedor asuma que se ha pronunciado alguna palabra, lo que provoca una mayor confusión del sistema.

La evaluación del reconocedor de habla pasa por obtener, como cifras de mérito de las prestaciones (sobre un conjunto de enunciados de prueba) la fracción de palabras reconocidas correctamente, la fracción de palabras erróneas (porcentaje de sustituciones), y las fracciones de palabras insertadas o borradas. La suma de sustituciones, inserciones y borrados se conoce como tasa de error (ER) del reconocedor, cuyo complementario (es decir, 100% - ER) se conoce como Word Accuracy, WA.

Para estimar el WA de nuestro sistema, se ha empleado el robot como interfaz para el control de un sistema domótico sencillo, como puede ser un equipo HI-FI (Fernández et al., 2005), lo cual asegura un vocabulario reducido (en torno a 500 palabras diferentes), con lo que el reconocimiento es más seguro que en vocabularios más amplios, puesto que el sistema ha de tomar una decisión sobre un menor número de hipótesis. Las pruebas realizadas sobre un conjunto de referencia de 1200 frases compuestas por un total de 6185 palabras, arrojan valores de WA del orden del 95,86%.

Si bien el valor anterior resulta de utilidad para un evaluador humano, la tasa de error aporta muy poca información al propio sistema de diálogo. Se han planteado varias fuentes de información entre las diferentes etapas del sistema, pero la más empleada es la basada en *medidas de confianza*, es decir, valores de mérito que informan al propio sistema del grado de bondad que alcanzan sus hipótesis. Siguiendo el trabajo presentado en (Ferreiros et al., 2005), la medida de confianza empleada se basa en la obtención de un grafo de palabras y la evaluación de la *pureza* de cada una de las mismas, entendida como la fracción de hipótesis en el grafo que incluyen una palabra concreta en un instante dado. Mediante el establecimiento de un umbral de confianza, se fija un primer nivel de control de corrección de palabras reconocidas: si una palabra ha sido reconocida con una confianza inferior al umbral, no se tendrá en cuenta en etapas posteriores del sistema de diálogo (como, por ejemplo, en el módulo de comprensión).

Además de la confianza de cada palabra, se calcula el valor de la confianza media para toda la frase. Este valor se obtiene mediante la ponderación de la contribución de cada palabra por el número de tramas que ocupa, valor que da una idea de la duración de dicha palabra. Este cálculo se ha planteado teniendo en cuenta que las palabras más largas suelen incluir información importante (y, por tanto, son de especial relevancia para etapas posteriores del sistema de diálogo). Las pruebas realizadas muestran una mejora significativa en el sistema de comprensión de lenguaje cuando se adopta esta modificación en el sistema de reconocimiento (Ferreiros et al., 2005; Sama et al., 2005).

3.2. Comprensión del lenguaje natural

El módulo de comprensión de lenguaje recibe como entrada la hipótesis que el reconocedor de habla ha determinado como más probablemente enunciada por el locutor, a partir de la cual debe extraer los conceptos clave incluidos en aquélla.

A fin de determinar qué conceptos están contenidos en un enunciado concreto, es necesario establecer diferentes categorías de palabras, es decir, grupos de palabras con características comunes, extraídas de un conjunto de frases de entrenamiento. Además, se ha de indicar que la clasificación de una palabra no depende únicamente de sí misma, sino también del contexto en el que se localiza.

Las diferentes palabras pueden ser categorizadas manualmente por un experto, o

bien realizar una clasificación automática basada en un conjunto de reglas. El primero de los métodos tiene como ventaja la exactitud en la clasificación de cada palabra, mientras que el segundo método permite fijar un número concreto de clases, y es mucho más rápido que el primero, pero es más complicado que la clasificación se realice de acuerdo a la semántica de la lengua, cosa que el primer método permite.

Una vez se conoce las diferentes categorías a las que puede pertenecer cada palabra, el módulo de comprensión evalúa el enunciado reconocido, obteniendo una serie de conceptos que se pasarán al gestor del diálogo. Como cifras de mérito, se obtendrán medidas de confianza a nivel de concepto, además de la tasa de acierto de conceptos, o Concept Accuracy (CA).

Para evitar ambigüedades en las palabras más comunes del vocabulario, se incluyó en el cálculo de medidas de confianza el concepto de *palabras no confiables*: son aquellas palabras que carecen de una categoría propia, pero que contribuyen a definir la categoría de las palabras a las que acompañan. Dentro de este grupo de palabras se incluyen determinantes, preposiciones o conjunciones. A la hora de estimar la confianza de un conjunto de conceptos, las palabras no confiables se excluirán del cálculo, de tal manera que sólo se tienen en cuenta las palabras categorizadas. Esto asegura una mejor estimación de las medidas de confianza, puesto que se eliminan aquellas palabras que no sólo no incluyen información, sino que además presentan mayor confusión entre sí.

El módulo de comprensión completo, al igual que el reconocedor de habla, se ha evaluado incluyendo el robot como interfaz para el control domótico de un equipo HI-FI. El valor de CA obtenido ha sido de 92,78%.

3.3. Gestor de diálogo

Las tareas que ha de desempeñar el gestor de diálogo son dos. Por un lado, y a partir de los conceptos que el módulo de comprensión ha extraído, debe generar una serie de acciones que el sistema (en nuestro caso, el robot) debe llevar a cabo. Por otra parte, el gestor ha de determinar los conceptos de una eventual respuesta vocal del robot, expresable a través del sistema de conversión de texto a voz.

El gestor de diálogo está basado en marcos. Esta aproximación consiste en mantener un marco con dos tipos de campos, denominados *atributo* y *valor*. En el primero de ellos, el sistema mantiene identificados los conceptos de interés para la tarea que está realizando en ese momento. En el campo de valor, el gestor almacenará las palabras que el módulo de comprensión ha etiquetado como uno de los conceptos presentes en la lista de atributos.

Si el sistema no puede rellenar todos los campos a partir de un único enunciado por parte del locutor, el gestor de diálogo enviará al generador de respuesta uno o varios conceptos que aún no tienen un valor asociado, de tal manera que se solicite al usuario tal información. El generador de respuesta aplicará sobre dichos conceptos las plantillas oportunas para construir un enunciado comprensible por el usuario, y lo pasará al conversor texto-voz para que éste sintetice la frase, estableciendo de esta manera un diálogo con el interlocutor humano. Dicho diálogo continuará hasta que el robot disponga de todos aquellos datos necesarios para que realice la acción deseada.

3.4. Conversor texto a voz

El conversor texto a voz genera un enunciado a partir del texto que le proporciona el generador de respuesta. Para ello, hace uso de un conjunto de parámetros prosódicos, como son el *pitch*, o frecuencia percibida como frecuencia fundamental de vibración de las cuerdas vocales; la *intensidad*, o energía de la señal, y la *duración* temporal de cada sonido.

Uno de los objetivos planteados a la hora de comenzar este proyecto era tratar de humanizar lo más posible el comportamiento del robot. Para eso, uno de los medios imprescindibles consiste en dotarle de una voz más expresiva y capaz de transmitir emociones, que se vea acompañada de los gestos tanto de la cara como del brazo que refuercen la expresión emitida por la voz.

La síntesis de voz con emociones que ofrece una mayor calidad es la consistente en la concatenación de unidades acústicas (generalmente, difonemas) a partir de un corpus amplio constituido por voz grabada de actores expresando diferentes emociones. Sin embargo, hemos optado por realizar la síntesis a partir de la modificación de los formantes de la voz neutra por varios motivos. En primer lugar, porque el modelado matemático de la voz permite aplicar cualquier tipo de

	Emoción identificada					
Emoción simulada	Alegría	Enfado en frío	Sorpresa	Tristeza	Neutra	Otra
Alegría	**53,9%**	9,6%	20,9%		7,8%	7,8%
Enfado en frío	7%	**70,4%**	14,8%	2,6%	3,5%	1,7%
Sorpresa	17,4%	2,6%	**79,1%**			0,9%
Tristeza		1,7%		**87%**	10,4%	0,9%
Neutra	1,7%	3,5%	2,6%	7,8%	**83,5%**	0,9%

Tabla 1: Matriz de confusión de emociones sintetizadas.

modificación en la señal generada, pudiendo obtener así una voz que exprese una emoción concreta a partir de una señal de voz neutra. Además, este método no requiere un corpus tan amplio como el anterior, puesto que sólo requiere un conjunto de frases de voz neutra, sobre la que se realizarán las modificaciones pertinentes, y un pequeño grupo de frases con las emociones que se desean sintetizar, a fin de obtener los parámetros para adaptar la voz neutra a la emoción objetivo. Así, basta con aplicar una serie de modificaciones sobre los elementos prosódicos de la voz original. (Barra et al., 2006) analiza las características de cuatro emociones básicas: alegría, tristeza, sorpresa y enfado, identificando los rasgos que permiten sintetizar una emoción a partir de voz neutra.

Las modificaciones planteadas sobre la voz neutra dependen de la emoción a sintetizar:

- La *alegría* necesita una modificación del ancho de banda de la señal original, así como una elevación del pitch y de su rango de variación, y un aumento de la velocidad de locución.
- La *tristeza* requiere una mayor lentitud en la expresión de la frase sintetizada y una reducción en la intensidad de la señal, además de un menor ancho de banda efectivo. Una mejora adicional consiste en modificar el pitch mediante la adición de un *jitter*, o pequeña variabilidad del mismo, de tal manera que se simula el temblor de la voz característico de una persona próxima a llorar.
- La *sorpresa* es especialmente difícil de sintetizar, puesto que se trata de una emoción transitoria que evoluciona rápidamente hacia otra emoción. Las modificaciones realizadas consisten en un aumento tanto del pitch como de su rango de variación, en un grado más acusado que en el caso de la alegría. Asimismo, se propone un contorno de frecuencia fundamental creciente hacia el final del enunciado, y una mayor duración de las sílabas tónicas.
- Por último, el *enfado* es una emoción con una importante componente no vocal, dado que casi siempre va acompañado de gestos corporales. La modificación planteada estriba en aumentar la intensidad de las sílabas tónicas y aumentar el rango de variación del pitch. Además, para simular el efecto de voz contenida y temblorosa característico del enfado en frío, se ha añadido una fuente de ruido aditivo síncrono con el pitch.

Este sistema de síntesis se ha evaluado presentando a un grupo de oyentes un conjunto de frases sintetizadas con diversas emociones, y solicitándoles que identificasen la emoción que, a su juicio, expresaba el locutor. Dicha emoción debía elegirse de un conjunto cerrado, que incluía las emociones sintetizadas, además de la voz neutra.

Los resultados de esta evaluación se muestran en la tabla 1. Se puede ver que la confusión es especialmente elevada entre alegría y sorpresa. Esto se debe a que, puesto que la sorpresa es un breve estado transitorio, si se pretende transmitir sorpresa en un enunciado largo, hay que mantener constantemente las modificaciones sobre la voz original, y dichas modificaciones son muy similares a las aplicadas para la síntesis de alegría, por lo que la confusión mutua entre ambas emociones aumenta significativamente. Además se observa cómo la voz que expresa tristeza está, a juicio de los oyentes, muy bien lograda, puesto que apenas presenta confusión con otras emociones.

4. Conclusiones

A la luz de los resultados mostrados en el presente trabajo, además de los resultados subjetivos obtenidos al emplear el robot en un contexto real, realizando las actividades propuestas con varios grupos de escolares entre 3 y 11 años, podemos afirmar que las prestaciones de los diferentes módulos que

componen nuestro robot lo hacen idóneo para cumplir una función fuertemente interactiva en el contexto de un museo de ciencias, no como sustituto de un guía humano, sino como un elemento más del museo al que se le añade una elevada capacidad de interacción con los visitantes.

El robot se desenvuelve de manera óptima en un entorno controlado (como puede ser una de las salas del museo) gracias al sistema de navegación.

Este control del entorno permite además el empleo de un vocabulario reducido, lo que asegura un número controlado de alternativas en el modelo de lenguaje empleado en el reconocedor de habla.

La medida de confianza básica se ha visto modificada mediante la definición de confianzas ponderadas y de palabras no confiables. Todas estas medidas de confianza son independientes de la tarea a realizar, lo que permite mantenerlas activas en cualquier entorno en el que se desee disponer del reconocedor de habla.

Las pruebas realizadas sobre el sistema demuestran que el cálculo modificado de medidas de confianza, junto con el empleo de un micrófono close-talk, han contribuido de manera importante a mejorar las tasas del reconocedor de habla y del sistema de comprensión, lo que permite que el robot responda a las intervenciones humanas con mayor eficacia, sin necesidad de volver a consultar con el interlocutor.

La capacidad del módulo de comprensión de aprender gradualmente de los ejemplos que se presentan a su entrada asegura unas tasas de Concept Accuracy muy elevadas en entornos controlados, además de no requerir una gramática previa o un conjunto de reglas para inferir los conceptos de una frase.

La inclusión de emociones en la voz sintetizada ha sido un gran acierto para hacer más atractivas las interacciones del robot con grupos de niños. Las modificaciones en los parámetros del sintetizador (valores medios y rangos del pitch, la amplitud, etcétera) han conducido a la obtención de una señal de voz capaz de expresar emociones. La evaluación de esta voz sintética demuestra cómo las modificaciones propuestas conducen a tasas significativas de reconocimiento de emociones por parte de oyentes no entrenados.

Se ha logrado que los movimientos del brazo y el rostro del robot (párpados y labios) contribuyan a una mayor expresividad del mismo, variando su posición de manera simultánea a la síntesis de voz, humanizando así sus intervenciones. Las pruebas realizadas con varios grupos de escolares demostraron que la identificación de la emoción se ve potenciada cuando ésta no sólo se expresa con la voz, sino también mediante gestos corporales.

En resumen, se ha logrado que el robot genere un mayor interés en el ámbito de un Museo de Ciencias.

Agradecimientos

El presente trabajo ha sido parcialmente financiado por el Ministerio de Educación y Ciencia, bajo los contratos DPI2007-66846-C02-02 (ROBONAUTA), DPI2004-07908-C02 (ROBINT) y por la UPM_CAM, bajo el contrato CCG06-UPM/CAM-516 (ATINA).

Los autores desean agradecer la colaboración de Nuria Pérez Magariños, así como el trabajo desarrollado por Ramón Galán y Diego Rodríguez-Losada, responsables de la estructura y el guiado del robot.

Bibliografía

Barra, R., Montero, J.M., Macías, J., D'Haro, L.F., San Segundo, R. and Córdoba, R., '*Prosodic and Segmental Rubrics in Emotion Identification*'. Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP'06) Pag. 1085-1088. 2006.

Breazeal, C., '*Toward Sociable Robots*'. Robots and Autonomous Systems, n 42. Pag. 167-175. 2003.

Dautenhahn, K. and Werry, I., '*Issues of Robot-Human Interaction Dynamics in the Rehabilitation of Children with Autism*'. Proceedings of the Sixth International Conference on the Simulation of Adaptive Behavior (SAB2000). Pag. 519-528. 2000.

Druin, A., Montemayor, J., Hendler, J., McAlister, B., Boltman, A., Fiterman, E., Plaisant, A., Kruskal, A., Olsen, H., Revett, I., Plaisant Schwenn, T., Sumida, L. and Wagner, R., '*Designing PETS*: *a Personal Electronic Teller of Stories*'. Human Factors in Computing Systems (CHI 99). ACM Press. Pag. 326-329. May 1999.

Fernández, F., Ferreiros, J., Sama, V., Montero, J.M., San Segundo, R., Macías, J. and García, R., '*Speech Interface for Controlling an Hi-Fi Audio System based on a Bayesian Belief Networks Approach for Dialog Modeling*'. Proceedings of the 9th Conference on Speech Communications and Technology (INTERSPEECH 2005). Pag. 3421-3424. September 2005.

Ferreiros, J., San Segundo, R., Fernández, F., D'Haro, L.F., Sama, V., Barra, R. and Mellén, P., '*New Word-Level and Sentence-Level Confidence Scoring using Graph Theory Calculus and its Evaluation on Speech Understanding*'. In Proceedings of the 9th Conference on Speech Communication and Technology (INTERSPEECH 2005). Pag. 3377-3380. September 2005.

Fong, T., Nourbakhsh, I. and Dautenhahn, K., '*A Survey of Socially Interactive Robots*'. Robots and Autonomous Systems, n 42. Pag. 143-166. 2003.

Huang, X., Acero, A. and Hon, H., '*Spoken Language Processing. A Guide to Theory, Algorithm and System Development*'. Prentice Hall. New Jersey. 2001.

Plaisant, C., Druin, A., Lathan, C., Dakhane, K., Edwards, K., Maxwell Vice, J. and Montemayor, J., '*A Storytelling Robot for Pediatric Rehabilitation*'. Proceedings of the Fourth International ACM Conference on Assistive Technologies. Pag. 50-55. 2000.

Rodríguez-Losada, D., '*SLAM Geométrico en Tiempo Real para Robots Móviles en Interiores basado en EKF*'. PhD Thesis (Unpublished). Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid. 2004.

Saldien, J., Goris, K., Vanderborght, B., Verrelst, B., Van Ham, R. and Lefeber, D., '*ANTY*: *The Development of an Intelligent Huggable Robot for Hospitalized Children*'. Vrije Universiteit Brussel (http://anty.vub.ac.be). 2006.

Sama, V., Ferreiros, J., Fernández, F., San Segundo, R., Pardo, J.M., '*Utilización de medidas de confianza en sistemas de comprensión del habla*'. Procesamiento del lenguaje natural Nº 35, pp. 229-234, ISSN 1135-5948. Septiembre 2005.

Shibata, T., Mitsui, T., Wada, K., Touda, A., Kumasaka, T., Tagami, K. and Tanie, K., '*Mental Commit Robots and its Application to Therapy of Children*'. Proceedings of the IEEE/ASME International Conference on Advanced Intelligence Mechatronics. Pag. 1053-1058. 2001.

Silva, A., Vala, M. and Paiva, A., '*Papous*: *the Virtual Storyteller*'. Intelligent Virtual Agents. Springer. 2001.

Willeke, T., Kunz, C. and Nourbakhsh, I., '*The History of the Mobot Museum Robot Series*: *An Evolutionary Study*'. American Association for Artificial Intelligence (www.aaai.org). 2001.

drodri http://www.disam.upm.es/~drodri/, 2008.

disam http://www.disam.upm.es/control/, 2008.