Procesamiento del Lenguaje Natural, núm. 30 (2003), pp. 89-97 recibido 1-03-2003; aceptado 20-05-2003

Uso de Canales de Comunicación Adicionales en Sistemas Conversacionales

R. López-Cózar

Dpto. Lenguajes y Sistemas Informáticos, E.T.S. Ingeniería Informática
18071 Universidad de Granada, Tel.: +34 958 240579, Fax: +34 958 243179
E-mail: rlopezc@ugr.es

Resumen: Durante los últimos años ha crecido notablemente el interés en la utilización conjunta del habla y otras modalidades de comunicación en los sistemas conversacionales, dando lugar al desarrollo de los denominados *sistemas conversacionales multimodales*. Disponiendo de varias modalidades de interacción, los usuarios pueden recibir mayor información del sistema conversacional en cada momento y éste puede recibir mayor información de los usuarios, reduciéndose por tanto el número de errores durante la interacción. En este trabajo se presenta una breve introducción a tecnología de la interacción multimodal, la cual pretendemos incorporar en los sistemas conversacionales desarrollados en nuestro laboratorio. El artículo describe las características más importantes de los sistemas conversacionales multimodales, las principales técnicas usadas para desarrollarlos y algunos de los problemas a los que deben hacer frente. Finalmente, describe brevemente el funcionamiento de dos sistemas conversacionales multimodales.
Palabras clave: Sistemas de diálogo multimodal, interacción hombre-máquina, reconocimiento de voz, síntesis de voz, control del diálogo.

Abstract: In recent years there has been an increasing interest concerning the integration of speech and other communication media in conversational systems, originating the so-called *multimodal conversational systems*. When several interaction modalities are available, users can receive more feedback from a conversational system and it can also receive more information from users, leading to a reduction of the interaction errors. This paper presents a brief introduction to the multimodal interaction technology, which we plan to set up for the conversational systems developed in our lab. The paper focuses on the most interesting features of multimodal systems, addressing the design techniques and some of the main problems to solve. Finally, it presents briefly the performance of two multimodal conversational systems.
Keywords: Multimodal dialogue systems, man-machine interaction, speech recognition, speech synthesis, dialogue management.

1 Introducción

Los sistemas conversacionales (o sistemas de diálogo) son sistemas diseñados para interactuar con los usuarios mediante el habla a fin de proporcionar determinados servicios, como por ejemplo, acceso a bases de datos, reservas de viajes, información meteorológica, localización de números de teléfono en directorios, localización de información en Internet, compra de productos, etc. (Bonafonte et al. 2000; López-Cózar et al. 2001, 2002; Rodríguez-Liñares L. 2002; Torres et al. 2002). Dado que el habla constituye la forma más natural de comunicación entre las personas, estos sistemas pueden aumentar la rapidez, efectividad y facilidad a la hora de realizar dichas tareas de forma automática.

La comunicación humana se basa en el uso de diversos canales de información (p. e. voz, miradas, gestos, expresiones faciales, etc.). Las personas usamos toda esta información, de forma inconsciente a veces, para añadir, modificar o sustituir información en la comunicación oral, lo cual nos permite conseguir gran exactitud en el reconocimiento de las palabras, incluso cuando existen problemas de comunicación en el entorno.

El objetivo fundamental de los sistemas conversacionales multimodales es usar diversos canales de información para mejorar la interacción con los usuarios. En una interacción multimodal, el sistema puede

ISSN 1135-5948

utilizar varios dispositivos de entrada (p. e. teclado, ratón, micrófono, cámara de visión artificial, pantalla sensible al tacto, etc.). La información de entrada se procesa a varios niveles de abstracción, obteniéndose diversos niveles de comprensión de la misma. Asimismo, el sistema multimodal puede utilizar diversos canales de salida para proporcionar información al usuario (p. e. voz, texto, gráficos, etc.) a fin de estimular varios sentidos del usuario de forma simultánea (Gustafson et al. 1999; Oviatt, 1996).

Algunos sistemas multimodales permiten incluso que los usuarios puedan elegir entre las diversas modalidades de entrada para llevar a cabo la interacción, permitiendo así una cierta adaptación a las condiciones ambientales de luz, ruido, etc. Además, esta ventaja permite que personas con determinadas discapacidades puedan usar este tipo de sistemas utilizando alguna de las modalidades de interacción disponibles (Beskow et al. 1997; Cole et al. 1999). Diversos estudios muestran que los usuarios no sólo prefieren los sistemas multimodales a los unimodales (basados únicamente en el habla), sino que además, la interacción multimodal permite reducir el porcentaje de errores y el tiempo requerido por los usuarios para realizar las tareas (Granström et al. 1999, 2002).

El uso de varias modalidades de comunicación puede permitir que éstas compensen sus respectivas limitaciones, lográndose una mayor velocidad de interacción y un mayor ancho de banda. Así por ejemplo, el habla puede compensar algunas de las limitaciones de las interfaces gráficas, pues posibilita referenciar objetos que no se encuentran en pantalla en un momento dado. Asimismo, las interfaces gráficas pueden compensar algunas de las limitaciones del habla, al hacer visibles los resultados de las acciones sobre los objetos, mostrando qué objetos y acciones son realmente importantes para una determinada tarea en un momento dado.

El habla constituye un canal de entrada muy importante en los sistemas conversacionales multimodales. No obstante, este artículo se centra en tres canales de información característicos de la comunicación multimodal: por una parte, *gestos* y *expresiones faciales* de los usuarios (usados en la interfaz de entrada), y por otra, *agente animado* (usado en la interfaz de salida). Por consiguiente, el artículo deja al margen las características y los problemas característicos de la comunicación oral, ampliamente discutidos en múltiples trabajos acerca de sistemas conversacionales unimodales.

El artículo está estructurado de la siguiente forma. La sección 2 proporciona una visión general de la información multimodal de entrada, centrándose en el reconocimiento del movimiento de labios y en el de gestos, así como en la fusión y representación de la información multimodal. La sección 3 está dedicada a la información multimodal de salida, haciendo especial hincapié en el agente animado. Esta sección también hace referencia a la necesidad de sincronización de los canales de salida y describe el funcionamiento de dos sistemas conversacionales multimodales. Finalmente, la sección 4 presenta las conclusiones.

2 Información multimodal de entrada

Además de usar diversos canales de salida para proporcionar información a los usuarios, los sistemas multimodales utilizan diversos canales de entrada para obtener información de ellos. Principalmente, estos canales aportan información acerca del movimiento de sus labios y de sus gestos.

2.1 Movimiento de labios

El movimiento de los labios permite obtener información acerca de la pronunciación de las palabras. Existen diversas técnicas para obtener este tipo de información (Wolff et al. 1994; Meier et al. 1999). Un método consiste en grabar y analizar el movimiento de labios de diversos sujetos reales a fin de extraer parámetros relativos a su forma y disposición, que permitan reconocer los ítems fonéticos asociados. Los problemas más importantes de este proceso son los relacionados con la localización de los labios en las imágenes, el seguimiento de su movimiento y la extracción de los parámetros deseados. Por ejemplo, en (Agelfors et al. 1998) se indica que teniendo en cuenta únicamente el movimiento de los labios, es posible obtener un 100% en la distinción de tres vocales, un 70% en la distinción de los diez dígitos, y un 25% en el reconocimiento de algunas frases completas.

No obstante, para realizar el reconocimiento del movimiento de labios, el sistema

multimodal debe comenzar por detectar la cara del usuario. En la bibliografía pueden encontrarse diversas técnicas aplicables (Collobert et al. 1996; Wang y Brandstein, 1999; Darrel et al. 2000). No obstante, las técnicas mayormente empleadas usan vectores de intensidad de color (Figura 1, izquierda) o vectores de perfiles faciales (Figura 1, derecha). Este último enfoque se basa en la detección de características faciales importantes, como la forma de los ojos, la nariz o los labios.

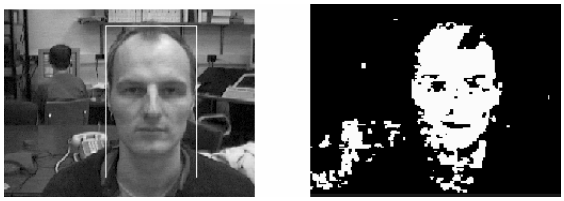

Figura 1. Reconocimiento de la cara del usuario

2.2 Gestos

Los gestos constituyen un canal de información muy integrado en el proceso de comunicación humana, que permite especificar muy eficientemente objetos y operaciones. Existen diversos tipos de gestos (Scherer y Ekman, 1982). Los gestos *simbólicos* se usan para transmitir mensajes (p. e. mover la cabeza para afirmar o negar). Los gestos *deícticos* se utilizan para señalar a un punto en el espacio (p. e. señalar con el dedo a un determinado lugar a la vez que se dice *"pon eso allí"*). Los gestos *icónicos* se emplean para describir objetos, relaciones o acciones visualmente. Los gestos *metafóricos* se usan para manipular objetos o herramientas abstractas. Finalmente, los gestos *rítmicos* se utilizan para marcar el ritmo de las frases.

Para realizar el reconocimiento de los gestos se pueden usar dispositivos intrusivos o no intrusivos, dependiendo de los requisitos de cada aplicación. Por ejemplo, en una aplicación de realidad virtual es admisible usar guantes de datos (Figura 2), en cuyo caso, el reconocimiento de los gestos se suele realizar en tres etapas. En la primera se realiza una abstracción de características mediante la que se obtienen diversos datos del guante, como la posición de los dedos y la orientación y movimiento de la mano. En la segunda etapa se construyen unas estructuras de datos denominadas *gestlets* (Chu et al. 1997), que almacenan información acerca de la postura, orientación y movimiento a partir de los datos obtenidos en la etapa anterior. Finalmente, la tercera etapa consiste en analizar los *gestlets* para determinar el tipo de gesto realizado por la mano del usuario.

Figura 2. Guante de datos

En otras aplicaciones, el uso de dispositivos intrusivos no es aceptable, debiendo usarse dispositivos no intrusivos (como cámaras, por ejemplo) y técnicas de visión artificial para segmentar y clasificar las imágenes. Este tipo de reconocimiento requiere usar técnicas sofisticadas para localizar al sujeto que realizará los gestos, detectar sus manos y brazos, y finalmente, clasificar sus gestos (Krumm et al. 2000).

2.3 Fusión de información multimodal

Para que un sistema conversacional puede usar la información multimodal de entrada es necesario combinar los datos provenientes de los diversos canales, a fin de obtener la interpretación correcta de cada acción del usuario. Este proceso, denominado *fusión*, se puede realizar a distintos niveles, que van desde el nivel de señal hasta el nivel semántico (Fisher et al. 2000).

La fusión a nivel de señal se suele emplear en el reconocimiento audio-visual del habla, combinando la información acústica y de movimiento de labios. Para realizar la fusión a este nivel se suelen emplear diversas tecnologías, como por ejemplo, redes neuronales artificiales (Hershey y Movellan, 1999) entrenadas de forma independiente para el reconocimiento acústico y visual. También se suelen usar Modelos Ocultos de Markov (Su y Silsbee, 1996) y transformadas rápidas de Fourier (McAllister et al. 1997) para analizar el espectro de voz y encontrar la correlación

existente entre la forma de los labios y la forma básica del espectro (Plumbley, 1991; Deco y Obradovic, 1996).

La información multimodal también se puede fusionar a nivel semántico. En este caso, la combinación de las entradas multimodales se realiza teniendo en cuenta su significado (Koons et al. 1993). Este tipo de fusión se suele realizar en dos etapas. En la primera, los eventos de las diversas modalidades se combinan en un módulo de interpretación a bajo nivel, y en la segunda, el resultado de la combinación anterior se envía a un módulo de interpretación de nivel superior que extrae su significado. Este módulo obtiene información acerca de la acción que el usuario desear realizar, la cual se envía al módulo de gestión del diálogo para que se inicie su ejecución. Un ejemplo de fusión a este nivel se produce cuando se combinan las informaciones obtenidas del análisis de la frase del usuario *"pon eso ahí"* y del gesto de su mano cuando pronuncia esta frase.

2.4 Representación de información multimodal

Para representar la información multimodal de entrada se pueden utilizar diversos enfoques. Por ejemplo, (Faure y Julia, 1993) proponen utilizar un formalismo sintáctico según el cual los eventos multimodales se representan mediante tripletes de la forma *{verbo, objeto, localización}*. Este formalismo es muy apropiado para representar información oral junto con información deíctica expresada mediante gestos.

Por otra parte, (Nigay y Coutaz, 1995) proponen una técnica denominada mezcla de botes (*melting pots*), según la cual, las informaciones multimodales se encapsulan junto con marcas temporales. Según esta aproximación, cuando se combinan los botes, la fusión puede ser de tipo *microtemporal*, *macrotemporal* o *contextual*. En el primer caso, se combinan informaciones producidas simultáneamente o muy próximas en el tiempo; en el segundo, se combinan informaciones secuenciales próximas entre sí y complementarias, y en el tercer caso, se combinan informaciones en base a restricciones semánticas.

Finalmente, otro enfoque para representar la información multimodal de entrada consiste en utilizar estructuras semánticas denominadas *frames* (Allen, 1995). En este caso, la información de cada modalidad se interpreta separadamente y se transforma en frames cuyos *slots* determinan los parámetros de la acción a realizar. Los frames pueden contener información parcial, representada mediante slots vacíos. Durante la fusión los frames se combinan entre sí, obteniéndose frames con todos los slots completos.

3 Información multimodal de salida

Los sistemas conversacionales multimodales suelen contar con un módulo de gestión de eventos de salida multimodales que decide cómo proporcionar la información al usuario (p. e. de forma acústica, visual o de ambas formas simultáneamente). La finalidad de combinar información visual y oral en la salida es facilitar que el usuario cree un modelo adecuado del funcionamiento del sistema, lo que facilita la corrección de los errores y la gestión del dialogo.

3.1 Agente animado

Los sistemas conversacionales multimodales suelen incluir en su interfaz de salida la imagen de un agente animado con aspecto humano a fin de permitir que la salida sea más expresiva y amigable (Figura 3) (Cassel et al. 2000). Este agente contribuye a incrementar la inteligibilidad de la salida oral generada por el sistema, siendo particularmente útil en entornos ruidosos y en las situaciones en que son frecuentes las conversiones cruzadas de diversas personas.

Figura 3. Interfaz gráfica de un sistema conversacional multimodal

En (House et al. 2001) se indica que el uso de un agente animado permite incrementar la inteligibilidad global de los estímulos VCV (vocal-consonante-vocal) en un entorno ruidoso en un 17% (desde 30% cuando sólo se utiliza el sintetizador de voz hasta 47% cuando se utiliza el sintetizador junto con el agente animado).

3.1.1 Representación del agente animado

Los primeros agentes animados estaban basados en la digitalización e interpolación de imágenes de caras con diferentes expresiones. Si bien esta técnica permitía obtener buenos resultados, añadir una nueva expresión requería crear un nuevo modelo y digitalizarlo, lo cual conlleva bastante tiempo. Actualmente, se suelen utilizar enfoques basados en modelos paramétricos, fisiológicos, procedurales o de deformación libre.

En los modelos paramétricos se utilizan parámetros de forma (relativos a la posición y tamaño de la nariz, los ojos, etc.) y parámetros expresivos (relativos al movimiento de cejas, boca, párpados, etc.) (Parke, 1982). La animación se realiza mediante el cambio de los valores de los parámetros oportunos. Este método es muy simple, requiere reducidas exigencias computacionales y resulta muy apropiado para la animación de los labios. Su principal inconveniente reside en que no permite modelar ni la propagación del movimiento ni el movimiento de los músculos.

Los modelos fisiológicos modelan muy bien las propiedades de la piel y las acciones musculares, pues se basan en estudios biológicos que permiten obtener un gran realismo y naturalidad en la animación; sin embargo, requieren un mayor coste computacional. Estos modelos pueden ser divididos, a su vez, en tres tipos: estructurales, basados en músculos y basados en el tejido facial.

- En los modelos estructurales la cara del agente está formada por una jerarquía de regiones (p. e. frente, labios, barbilla, etc.) y subregiones (p. e. labio superior, labio inferior, etc.), de forma que las regiones se pueden contraer por los efectos de los músculos, o bien, pueden ser afectadas por los movimientos de regiones adyacentes.
- Los modelos basados en músculos combinan características anatómicas de los músculos junto con propiedades de la cara, como la elasticidad de la piel, por ejemplo. Cada músculo se representa mediante un vector que indica una dirección, una magnitud y una zona de influencia.
- En el modelo basado en tejido facial se considera que la piel está formada por varias capas de diferente grosor, que se simulan mediante mallas.

Los modelos procedurales no se basan en estudios biológicos, sino en datos empíricos que permiten simular la acción de los músculos mediante procedimientos especializados similares a los utilizados en el modelo FACS (*Facial Action Coding System*) (Ekman y Rosenberg, 1997), los cuales tienen en cuenta el efecto de la contracción de los músculos en las regiones de la cara. La principal ventaja de este método reside en que permite definir una estructura jerárquica de acciones para producir expresiones faciales y movimientos de los labios; sin embargo, no tiene en cuenta la propagación de los movimientos.

Finalmente, los modelos de deformación libre modelan las acciones musculares mediante las denominadas *cajas de deformación* (*deformation box*), compuestas por conjuntos de puntos que se pueden estirar, contraer, doblar, etc. Cada caja afecta sólo a los puntos de una región, y por tanto, simula el efecto de un solo músculo. Para realizar la animación de la cara del agente se utilizan diversas funciones de deformación que simulan el comportamiento de los diferentes músculos (Kalra et al. 1992).

3.1.2 Sincronización de canales

Un problema muy importante relacionado con la animación del agente es la sincronización de la salida oral y visual a fin de lograr sensación de realismo. En (Guiard-Marigny et al. 1996) se indica que la falta de sincronización es perceptible para el ser humano si el canal de audio se adelanta en más de 130 ms o se retrasa en más de 260 ms respecto al canal de vídeo. El margen es más reducido para los sonidos agudos; así por ejemplo, en el caso de un martillo golpeando contra un bloque de acero, la falta de sincronización es perceptible si la señal de audio se adelanta en más de 75 ms o se retrasa en más de 188 ms.

Para lograr la sincronización entre el canal de audio y el de vídeo, el primero se suele usar

como “reloj síncrono”, existiendo diversas aproximaciones para “visualizar” los fonemas. Por ejemplo, en (House et al. 2001), para cada fonema, el módulo de audio envía una señal al módulo de vídeo para que éste calcule la disposición correspondiente de los labios (denominada *visema*). Para lograr sensación de realismo es fundamental mostrar los visemas importantes durante la locución; por ejemplo, durante la pronunciación del fonema /p/ es muy importante que el agente animado cierre claramente los labios.

Por otra parte, en (Beskow, 1995) el problema de la sincronización se resuelve mediante un conjunto de reglas que tienen en cuenta los efectos coarticulatorios. Los gestos y las expresiones faciales se controlan mediante procedimientos que “encriptan” la actualización conjunta de diversos parámetros (p. e. tiempo, valor, dirección, duración, etc.). Esta técnica permite al agente animado realizar gestos complejos que requieren la actualización simultánea de un gran número de parámetros.

3.2 Algunos sistemas conversacionales multimodales

3.2.1 Olga

Olga es un sistema conversacional multimodal desarrollado en la Universidad de Estocolmo (Suecia) cuya finalidad es proporcionar información acerca de hornos microondas (McGlashan y Axlin, 1996). El sistema emplea información multimodal sólo para la salida; la información de entrada únicamente se proporciona mediante el habla. Además de los módulos relacionados con la interfaz oral, el sistema consta de una interfaz gráfica para proporcionar información y permitir la navegación a través de menús. El módulo de gestión del diálogo se encarga de interpretar las frases de los usuarios y coordinar la generación de las salidas en ambos canales de comunicación (oral y gráfico).

La interfaz gráfica consta de un agente animado (Figura 4) que puede mover los labios, la lengua y la barbilla sincronizadamente con la salida oral. De esta forma, el sistema puede enfatizar determinadas frases e incrementar la inteligibilidad de las mismas. Además, el agente puede mirar a los diversos objetos mostrados en la pantalla cuando hace referencia a ellos. Para decidir los canales de salida que debe utilizar en cada momento, el módulo de gestión del diálogo usa un conjunto de reglas que tienen en cuenta los objetivos del sistema, así como el tipo de información que se debe proporcionar al usuario.

Figura 4. Agente animado del sistema Olga

El sistema emplea mensajes orales y gestos para expresar los objetivos que tienen como finalidad mantener el control del diálogo y proporcionar retroalimentación al usuario. Por ejemplo, para mostrar la comprensión de las frases del usuario, el agente realiza un gesto de asentimiento con la cabeza; en cambio, para expresar la incomprensión abre la boca y levanta las cejas, a la vez que solicita una clarificación de forma oral. Asimismo, si tras consultar su base de datos, el sistema no tiene datos que proporcionar la usuario, el agente informa de ello mediante un mensaje oral que se acompaña de un gesto de tristeza del agente. La información de productos obtenida de la base de datos se transmite mediante mensajes orales y gráficos, y la información detallada sobre algún producto se muestra en pantalla a la vez que el agente genera un resumen oral de la misma.

3.2.2 SmartKom

SmartKom es un sistema conversacional multimodal en desarrollo actualmente en el Centro para la Inteligencia Artificial (Alemania), que permite la interacción del usuario mediante habla y gestos de la mano (Wahlster, 2001). El sistema puede interactuar con el usuario mediante habla y gráficos, sirviéndose de un agente animado que mueve los labios y el cuerpo, y realiza gestos sincronizadamente. La finalidad del sistema es proporcionar información acerca de la

programación de películas en salas cinematográficas, así como gestionar reservas para asistir a las mismas (Figura 6).

Para presentar la información visual al usuario, el sistema utiliza una superficie plana (p. e. una mesa) sobre la que proyecta imágenes de la sala cinematográfica, del agente animado y otras. Para reservar una butaca, el usuario señala sobre la imagen de la sala (sin tocar la superficie) y dice *"Quiero este asiento"*. El sistema usa una cámara de visión artificial para reconocer el movimiento de la mano del usuario, y emplea la información deíctica obtenida junto con la oral para comprender su intención. Como resultado de la acción del usuario, el sistema muestra a continuación el plano de la sala, en el que aparece marcado el asiento seleccionado.

Figura 6. Usuario interactuando con el sistema SmartKom

El sistema consta de una arquitectura multiplataforma en la que se realiza de forma paralela el análisis de cada modalidad de entrada. Para procesar la información generada por cada uno de los módulos del sistema se utiliza el lenguaje XML, muy apropiado para facilitar la representación y el intercambio de información multimodal. Usando este lenguaje, cuando el usuario señala en un mapa la ubicación de la sala cinematográfica a la que desea asistir, se genera una representación XML como la mostrada a continuación, en la que se puede observar que existen dos hipótesis para el objeto referenciado mediante el gesto del usuario ("dynId30" y "dynId28"). El primer objeto es el que tiene una mayor prioridad (1) y se corresponde con el cine "Europa" ubicado en unas determinadas coordenadas del mapa.

```
<gestureAnalysis>
[...]
  <type> tarrying </type>
  <referencedObjects>
    <object>
    <displayObject>
      <contentReference> dynId30 </contentReference>
    </displayObject>
    <priority>  1  </priority>
  </object>
  <object>
  <displayObject>
      <contentReference> dynId28 </contentReference>
    </displayObject>
    <priority>  2  </priority>
  </object>
</referencedObjects>
<representationContent>
[…]
  <movieTheater structID=dynId30>
   <entityKey> cinema_17a </entityKey>
  <name> Europa </name>
    <geoCoordinate>
      <x> 225 </x>   <y> 230 </y>
    </geoCoordinate>
[...]
</gestureAnalysis>
```

Durante la interacción con el usuario, las estructuras XML proporcionadas por los módulos de reconocimiento de voz y de gestos se envían a un módulo que realiza la fusión de las mismas. Tras un proceso de inferencia, el módulo de gestión del diálogo decide cual debe ser la reacción del sistema, seleccionado las modalidades de salida que se han de usar para interactuar con el usuario de forma coherente y natural.

4 Conclusiones

En este artículo se ha presentado una visión panorámica de la tecnología de los sistemas conversacionales multimodales, centrada principalmente en los aspectos que los diferencian de los sistemas conversacionales *tradicionales*, basados únicamente en el habla como canal de comunicación.

En primer lugar se han mencionado brevemente las principales ventajas que aportan los canales de comunicación característicos de la interacción multimodal. A continuación se ha analizado el procesamiento de la información multimodal de entrada proporcionada por dos de estos canales: movimiento de labios y gestos. Seguidamente,

se ha examinado el proceso de representación y fusión de la información multimodal de entrada, describiendo tres de las principales técnicas empleadas. A continuación se ha estudiado el proceso de generación de información multimodal de salida, prestando especial atención al denominado *agente animado*. Se han descrito los principales enfoques existentes para modelar dicho agente y se ha analizado el problema de la sincronización de la información proporcionada por el mismo. Para concluir, se ha descrito brevemente, a modo de ejemplo, el funcionamiento de dos sistemas conversacionales multimodales.

Bibliografía

Agelfors E., Beskow J., Dahlquist M., Granström M., Lundeberg K.-E., Spens K.-E., Öhman T. 1998. Synthetic faces as a lipreading support. Proc. ICSLP, Sydney, Australia

Allen J. 1995. Natural language understanding. The Benjamin/Cummings Publishing Company Inc.

Beskow J. 1995. Rule-based visual speech synthesis. Proc. Eurospeech, Madrid, pág. 299-302

Beskow J., Dahlquist M., Granström B., Lundeberg M., Spens K.-E., Öhman T. 1997. The telephace project – multimodal speech communication for the hearing impaired. Proc. Eurospeech, Rodas, Grecia

Bonafonte A., Aibar P., Castell N., Lleida E., Mariño J. B., Sanchís E., Torres I. 2000. Desarrollo de un sistema de diálogo oral en dominios restringidos. Primeras Jornadas en Tecnología del Habla, Sevilla

Cassel J., Bickmore T., Campbell L., Hannes V., Yan H. 2000. Conversation as a system framework: Designing embodied conversational agents. Cassell J.l, Sullivan J., Prevost S., Churchill E. (Eds). Embodied Conversational Agents, Cambridge Ma, The MIT Press

Chu C. C., Dani T. H., Gadh R. 1997. Multisensory Interface for a Virtual Reality Based Computer Aided Design System, Computer-Aided Design, 29 (10), pág. 709-725

Cole R., Massaro D. W., de Williers J., Rundle B., Shobaki K., Wouters J., Cohen M., Beskow J. Stone P., Connors P. Tarachov A., Solcher D. 1999. New tools for interactive speech and language training: Using animated conversational agents in classrooms of profoundly deaf children. Proc. ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education, pág. 45-52

Collobert M., Feraud R., LeTorurneur G., Bernier O., Viallet J. E., Mahieux Y. Colobert D. 1996. Listen: a system for locating and tracking individual speakers. Proc. Second International Conference on Face and Gesture

Darrel T., Gordon G. G., Harville M., Woodfill J. 2000. Integrated person tracking using stereo, color and pattern detection. IJCV, 37(2), junio, pág. 199-207

Deco G., Obradovic D. 1996. An information theoretic approach to neural computing. Springer-Verlag

Ekman P., Rosenberg E. L. 1997. What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS). New York: Oxford University Press

Faure C., Julia L. 1993. Interaction homme-machine par la parole et le geste pour l'édition de documents. Proc. International Conference on Real and Virtual Worlds, pág. 171-180

Fisher J. W., Darrell T., Freeman W. T., Viola P. 2000. Learning statistical models for audio-visual fusion and segregation. Advances in Neural Information Processing Systems, 13

Granström B., House D., Lundeberg M. 1999. Prosodic cues in multimodal speech perception. Proc. ICPhS, pág. 655-658

Granström B., House D., Swerts M. G., 2002. Multimodal feedback cues in human-machine interactions. Proc. Speech Prosody Conference, pág. 11-13

Guiard-Marigny T., Tsingos N., Adjoundani A., Benoît C., Gascuel M. P. 1996. 3D models of the lips for realistic speech animation. Proc. Second ESCA/IEEE workshop on Speech Synthesis, pág. 49-52

Gustafson J., Lindberg N., Lundeberg M. 1999. The August Spoken Dialogue System. Proc. Eurospeech, pág. 1151-1154

Hershey J., Movellan J. 1999. Using audio-visual synchrony to locate sounds. S. A. Solla, T. K. Leen, K. R. Miller (Eds). *Advances in neural Information Processing Systems* 12, pág. 813-819

House D., Beskow J., Granström B. 2001. Timing and interactions of visual cues for prominence in audiovisual speech perception. Proc. Eurospeech, Aalborg, Dinamarca, pág. 387-390

Kalra P., Mangili A., Magnenat-Thalmann N., Thalmann D. 1992. Simulation os muslce action using rational free form deformations. Proc. Eurographics, Computer Graphics Forum 2(3), pág. 59-69

Koons D. B., Sparrell C. J., Thorisson K. R. 1993. Integrating simultaneous input from speech, gaze, and hand gestures. M. Maybury (Ed). Intelligent Multimedia Interfaces, pág. 257-275

Krumm J., Harris S., Meyers B., Brummit B., Hale M. Shafer S. 2000. Multi-camera multi-person tracking for easyliving. Proc. IEEE Workshop on Visual Surveillance

López-Cózar R., Segura J.C. De la Torre A., Rubio A. J. 2001. Una Nueva Técnica para Evaluar Sistemas Conversacionales Basada en la Generación Automática de Diálogos. Procesamiento del Lenguaje Natural, nº 27, pág. 255-260

López-Cózar R., Rubio A. J., J. E. Díaz-Verdejo, López-Soler J. M. 2002. Validación de un Sistema de Diálogo Mediante el Uso de Diferentes Umbrales de Poda en el Proceso de Reconocimiento Automático de Voz. Procesamiento del Lenguaje Natural, nº 29, pág. 205-211

McAllister D., Rodman R., Bitzer D., Freeman A. 1997. Lip synchronization of speech. Proc. AVSP Workshop, Rodas, Grecia

McGlashan S., Axling T. 1996. Talking to agents in virtual worlds. Proc. Third UK Virtual Reality Conference

Meier U., Stiefelhagen R., Yang J., Waibel A. 1999. Towards unrestricted lipreading. Proc. Second Conference on Multimodal Interfaces (ICMI)

Nigay L., Coutaz J. 1995. A generic platform for addressing the multimodal challenge. Proc. International Conference on Human-Computer Interaction, ACM, pág. 98-105

Oviatt S. 1996. Multimodal interfaces for dynamic interactive maps. Proc. Conference on Human Factors in Computing Systems

Parke F. I. 1982. Parametrized models for facial animation. IEEE Computer Graphics, 2(9), pág. 61-68

Plumbley M. 1991. On information theory and unsupervised neural networks. Technical report CUED/F-INFENG/TR. 78, Cambridge University Engineering Department, UK

Rodríguez-Liñares L. 2002. Un Sistema de Diálogo para la Consulta de Correo Electrónico en Lenguaje Natural. Procesamiento del Lenguaje Natural, nº 29, pág. 181-187

Scherer K. Ekman, P. 1982. Handbook of Methods in Nonverbal Behavior Research. Cambridge University Press

Su Q., Silsbee P. 1996. Robust audiovisual integration using semicontinuous Hidden Markov Models. Proc. ICSLP

Torres F., Sanchís E., Segarra E. 2002. Desarrollo de un gestor de diálogo basado en modelos estocásticos y dirigido por la semántica.. Procesamiento del Lenguaje Natural, nº 28, pág. 175-180

Wahlster, W. 2001. SmartKom: Multimodal dialogs with mobile web users. Proc. International Cyber Assist Symposium. Tokyo Internation Forum, pág. 33-40

Wang C., Brandstein M. 1999. Multi-source face tracking with audio and visual data. IEEE International workshop on multimedia signal processing

Wolff G., Prasad K. V., Stork D. G., Hennecke M. 1994. Lipreading by neural networks. IEEE International Workshop on Multimedia Signal Processing