Análisis de Metodologías de Evaluación de Sistemas de Diálogo Multimodal

R. López-Cózar, Z. Callejas, M. Gea

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, zoraida@correo.ugr.es, mgea@ugr.es

Resumen: Los sistemas de diálogo multimodal constituyen un campo de investigación de gran interés en la actualidad, pues mejoran notablemente la interacción proporcionada por los sistemas de diálogo oral. Este artículo presenta, de forma resumida, el resultado de un amplio y profundo estudio con el que hemos pretendido conocer cómo se pueden evaluar tales sistemas, y en qué se diferencia esta evaluación de la realizada con sistemas de diálogo oral. Las conclusiones del estudio indican que no existen unos métodos claramente aceptados por la comunidad científica que puedan considerarse *estándares de evaluación*. Ello sumado a la gran diversidad de tareas y diferencias de complejidad de unos sistemas frente a otros, dificulta en gran medida la comparación de los sistemas de una forma objetiva. Por consiguiente, son necesarias técnicas estándar que faciliten la evaluación teniendo en cuenta las características específicas de cada sistema, así como métodos que faciliten la comparación de sistemas diseñados para tareas diferentes.
Palabras clave: Sistemas de diálogo, interacción multimodal, evaluación de sistemas.

Abstract: Multimodal dialogue systems represent a very challenging research field nowadays, as they enhance the interaction provided by spoken dialogue systems. This paper presents, in a summarised way, the results of a wide and deep study we have carried out to try to find the cues for the evaluation of these systems, as well as the main differences in comparison with the evaluation of spoken dialogue systems. Our conclusions indicate there are no methods clearly accepted by the scientific community that can be considered as *evaluation standards*. This fact, together with the diversity of tasks and differences in the system complexity, makes it very difficult to compare systems objectively. Thus, new techniques are necessary to ease the evaluation taking into account the specific features of each system, as well as new methods to compare systems designed for different tasks.
Keywords: Dialogue systems, multimodal interaction, system evaluation.

1 Introducción

Los sistemas de diálogo son programas de ordenador cuya finalidad es proporcionar un servicio a los usuarios, emulando en la medida de lo posible el comportamiento inteligente del ser humano que realiza dicha tarea. En nuestros días, diversas empresas e instituciones públicas y privadas usan estos sistemas con la finalidad de proporcionar información y otros servicios de forma automática, principalmente de forma oral a través del teléfono. Estos sistemas se pueden clasificar atendiendo a diversos criterios. Uno es el número de canales de comunicación sistema-usuario que se utilizan durante la interacción (López-Cózar 2003). Teniendo en cuenta este criterio, tales sistemas se pueden clasificar en dos tipos: *orales* y *multimodales*. En los sistemas de diálogo oral (*Spoken Dialogue Systems*, SDSs) sólo se utiliza un canal de comunicación de E/S (micrófono y altavoces del ordenador o teléfono) a través del cual la interacción se realiza mediante habla. En cambio, los sistemas de diálogo multimodal (*Multimodal Dialogue Systems*, MDSs) utilizan varios canales de comunicación (p. e. habla, texto, expresiones faciales, gestos, etc.) mediante micrófono y altavoces, cámaras para visión artificial, guantes de datos, pantallas sensibles al tacto, etc., proporcionando una interacción más flexible, cómoda y adaptativa, que permite reducir los errores en la comunicación.

El artículo está estructurado de la siguiente forma. La sección 2 presenta los

principales problemas de las técnicas de evaluación tradicionales (pensadas para SDSs) a la hora de evaluar MDSs. La sección 3 describe las técnicas de evaluación a nivel de sistema, mientras que la sección 4 se centra en la evaluación a nivel de componentes. Finalmente, la sección 5 presenta las conclusiones del artículo.

2 Problemas de las técnicas de evaluación tradicionales

La evaluación es un aspecto clave en la tecnología de los sistemas de diálogo. Antes de implantar un nuevo sistema en el mundo real es necesario evaluarlo, para verificar que los usuarios acepten su funcionamiento. No obstante, la evaluación de los MDSs es un campo de trabajo abierto, dada la falta de estándares y métodos claramente aceptados por la comunidad científica. Ello se debe a varias razones. Por una parte, no existen suficientes bases de datos multimodales que puedan ser utilizadas como referencia en la evaluación. Por otra, los métodos usados tradicionalmente para evaluar de forma separada los diversos módulos de un MDS (p. e. reconocedor de habla, reconocedor de escritura a mano, detector de la cara del usuario, etc.) no tienen en cuenta uno de los aspectos más importantes de estos sistemas: la combinación de las diversas modalidades de interacción. Por ejemplo, en una interacción multimodal, los usuarios pueden realizar consultas empleando la voz a la vez que proporcionan información adicional, o más específica, señalando y/o gesticulando, haciendo por tanto más compleja la interacción y, en consecuencia, también la evaluación.

Asimismo, medir la contribución de cada módulo de reconocimiento en el rendimiento global del sistema no es algo tan inmediato como en el caso de los SDSs, puesto que estos módulos realizan tareas de muy distinta complejidad. Por ejemplo, el módulo que realiza el reconocimiento automático del habla (RAH) suele llevar a cabo una tarea muy compleja, mientras que el reconocedor de gestos suele trabajar con un conjunto muy reducido de posibles gestos a reconocer. En consecuencia, el segundo probablemente sea más fiable que el primero; sin embargo, el segundo debe tener un mayor peso al evaluar el rendimiento del sistema, puesto que la tarea que lleva a cabo es mucho más compleja. Además de emplear distintos pesos para los diferentes reconocedores, en la evaluación de los MDSs existen problemas de sincronización, pues será necesario decidir qué entradas multimodales deben ser consideradas síncronas (es decir, referentes a la misma intención del usuario) y cuales no.

3 Evaluación a nivel de sistema

Para realizar la evaluación de los MDSs a nivel de sistema se suelen emplear diversos enfoques, que están relacionados con el tipo de fusión de los datos multimodales de entrada (Wahlster 2002). Si se usa fusión a nivel de señal, la medidas de evaluación que suelen emplearse son parecidas a las usadas para evaluar de forma aislada los componentes, cuyas salidas se combinan en el proceso de fusión. Por ejemplo, en el caso de reconocimiento audio-visual del habla (*Audio-Visual Automatic Speech Recognition*, AVASR) (Rogozan y Deléglise 1998), las medidas de evaluación tratan de determinar la precisión del reconocimiento, al igual que ocurre con las medidas de evaluación del RAH y de la lectura de labios (*lipreading*). En cambio, si el proceso de fusión de realiza a nivel semántico, las medidas de evaluación que se suelen emplear intentan determinar el porcentaje de logro de tareas, tiempo requerido para realizar las tareas, naturalidad del diálogo, satisfacción del usuario, costes, etc.

Se pueden considerar tres enfoques para realizar la evaluación a nivel de sistema: experimental, predictivo, y basado en expertos.

3.1 Enfoque experimental

El enfoque experimental se basa en datos reales obtenidos de usuarios de test que realizan tareas reales. Al igual que ocurre en el caso de los SDSs, la evaluación de los MDSs se suele realizar mediante cuestionarios que los usuarios rellenan después de interactuar con el sistema a evaluar. Para ello suelen usarse escenarios

que definen una serie de objetivos que los usuarios deben tratar de lograr durante la interacción. Los cuestionarios se emplean para conocer la opinión de los usuarios respecto a varias medidas subjetivas, como por ejemplo, la facilidad de uso, satisfacción, implicación en la tarea, eficiencia, número de errores y facilidad para corregirlos, eficacia, naturalidad, etc. (Sturm et al. 2002). En el caso de los MDSs, los cuestionarios suelen incluir aspectos adicionales, específicos de estos sistemas, como por ejemplo, relativos a la personalidad del agente animado y su comportamiento durante el diálogo, velocidad y robustez de las distintas modalidades de interacción, etc. Estas medidas subjetivas se suelen asociar a afirmaciones con las que los usuarios de test deben manifestar su acuerdo (o desacuerdo), generalmente usando una escala de Likert. Por ejemplo, en una escala de cinco puntos se tendría: "Estoy en completo desacuerdo" = -2, "Estoy en desacuerdo" = -1, "Indiferente" = 0, "Estoy de acuerdo"=1, "Estoy en completo acuerdo"= 2. Una vez se tienen estas valoraciones, se suelen emplear métodos estadísticos para calcular los valores de las medidas consideradas, como por ejemplo, ANOVA, Tukey post hoc, Kruskal-Wallis, Mann-Whitney, etc. (Fabri et al. 2002).

3.2 Enfoque predictivo

Este enfoque pronostica el comportamiento del usuario considerando variables de rendimiento que tienen en cuenta suposiciones o parámetros de un modelo, sin emplear un sistema previamente implementado (Mellor y Baber 1997). La principal ventaja de este enfoque reside en que permite evaluar la interfaz del usuario en una etapa temprana del ciclo de desarrollo, de forma que el diseño pueda ser mejorado antes de que se lleve a cabo la implementación final. La principal desventaja es que las predicciones se basan en teorías hipotéticas y no en datos reales, con lo que éstas pueden ser poco precisas. Otro inconveniente es que la especificación de un modelo predictivo puede llevar prácticamente el mismo tiempo que implementar un prototipo del sistema. Algunas técnicas predictivas son las siguientes: CCT, ICS, KRI y *cognitive walkthrough*.

3.3 Enfoque basado en expertos

En este enfoque, un profesional experimentado emplea un prototipo del sistema y evalúa su especificación para determinar si cumple con los criterios de diseño predefinidos. La principal desventaja de este enfoque es la dificultad para encontrar al experto, por lo que, en su lugar, se suele recurrir a evaluadores cuya misión es intentar encontrar un número razonable de problemas de diseño (normalmente son necesarios al menos tres evaluadores para encontrar la mitad de estos errores). Entre otros investigadores, este enfoque ha sido utilizado por Almeida et al. (2002) para evaluar el sistema MUST (*MUltimodal, multilingual information Services for small mobile Terminals*), implementado en un dispositivo móvil. El sistema recibe como entrada voz y gestos realizados con el lápiz para señalar a objetos, y proporciona como salida voz, texto y gráficos. La evaluación, realizada por doce expertos, muestra que la mayoría de ellos comenzaron a interactuar empleando las dos modalidades por separado, y algunos de ellos ni siquiera intentaron combinarlas. Tras un cierto tiempo usando el sistema, cinco expertos comenzaron a emplear la voz y el lápiz simultáneamente. Por consiguiente, los experimentos mostraron que no era intuitivo, ni mucho menos obvio, que ambas modalidades de entrada podían ser empleadas simultáneamente. Ello sugiere que para usuarios principiantes, es conveniente realizar una explicación introductoria al servicio y a la interfaz, p. e. mediante un video o una animación breve.

3.3.1 PROMISE

Dadas las características y los problemas específicos de los MDSs, Beringer et al. (2002) propusieron el procedimiento de evaluación PROMISE (*PROcedure for Multimodal Interactive System Evaluation*), en el que, por una parte, se utilizan métodos empleados para evaluar SDSs, y por otra, métodos específicos para evaluar las propiedades características de los MDSs, como por ejemplo, la combinación de

gestos y habla en la entrada, la combinación de habla y gráficos en la salida, etc. Según este procedimiento, la evaluación se lleva a cabo de forma subjetiva definiendo una serie de medidas cualitativas y cuantitativas (denominadas *costes*) que tienen pesos asociados. En lugar de usar una regresión lineal (como en el caso del procedimiento PARADISE para la evaluación de SDSs), PROMISE emplea una correlación de Pearson calculada entre pares "Satisfacción del usuario – Coste", siendo algunos de estos costes objetivos y otros subjetivos. Para llevar a cabo la evaluación, los usuarios de test interactúan con el sistema y rellenan un cuestionario que incluye costes subjetivos. Algunos de estos costes son equivalentes a los utilizados en el procedimiento PARADISE, mientras que otros se utilizan para tratar de forma específica la multimodalidad y el comportamiento de usuarios no cooperativos.

4 Evaluación de componentes

La evaluación a nivel de componentes se realiza usando medidas que determinan el funcionamiento de los diversos módulos de los MDSs de forma aislada. También se usan medidas para evaluar el comportamiento y los efectos de los agentes animados, que son específicos de estos sistemas. En esta sección describimos brevemente los métodos de evaluación utilizados.

4.1 Localización de la cara y seguimiento de la mirada

Dado que la localización de la cara (*face localisation*) requiere un paso previo de detección, generalmente se suele emplear la tasa de detección (ratio de detecciones correctas) como medida de evaluación. Por ejemplo, Rowley et al. (1998) emplearon esta medida para evaluar un sistema de localización basado en redes neuronales artificiales (*Artificial Neural Networks*, ANNs), obteniendo un 90,5% de detección para un corpus que contenía 130 imágenes en escala de gris. Yang et al. (1998) evaluaron la precisión a la hora de estimar la posición de la cabeza del usuario, considerando el error medio de rotación (medido en grados) y el error medio de traslación (medido en mm).

Una vez la cara del usuario ha sido localizada, se puede realizar el seguimiento de su mirada (*gaze tracking*). Existen diversas medidas de evaluación para esta tarea. La *precisión* (desviación respecto a la posición verdadera) y el *éxito de seguimiento* (proporción de tiempo en que la mirada es correctamente seguida) son las más empleadas. Por ejemplo, Yang et al. (1998) evaluaron un sistema de seguimiento de miradas usando la precisión (valor medio del error medido en grados). Otra medida de evaluación es el *tiempo* que se requiere para realizar una tarea dada, en comparación con el requerido usando otras modalidades de interacción. Por ejemplo, Sibert (2000) comparó los tiempos requeridos por el seguimiento de la mirada y el ratón en una tarea de selección de objetos, encontrando que la selección mediante la mirada era 338 ms más rápida que con el ratón.

4.2 Reconocimiento de gestos

En la literatura podemos encontrar diversos métodos para realizar el reconocimiento de gestos, basados principalmente en plantillas, ANNs, Modelos Ocultos de Harkov (*Hidden Markov Models*, HMMs), Redes Bayesianas, y técnicas de visión artificial. Los gestos pueden corresponderse con un único comando, una secuencia de comandos, una sola palabra o una frase, y pueden ser estáticos o dinámicos. La evaluación de los sistemas de reconocimiento de gestos se suele realizar mediante la *precisión* (proporción de gestos correctamente reconocidos). No obstante, algunos investigadores también han utilizado el tiempo de reconocimiento y la robustez (Hu et al. 2003, Wachs et al. 2002).

Podemos encontrar resultados muy diversos obtenidos de un amplio abanico de aplicaciones, como por ejemplo mapas interactivos, interacción en entornos de oficina, interacción con robots, etc. Por ejemplo, los experimentos realizados por Kettebekow y Sharma (2000) en el dominio de los partes meteorológicos y el marco de trabajo iMAP, proporcionaron tasas de reconocimiento de 78,1% y 79,6%,

respectivamente. Por otra parte, Montero y Sucar (2004) estudiaron un conjunto de gestos empleados en la interacción con otros objetos en un entorno de oficina. Los gestos, realizados con las manos, fueron detectados y seguidos mediante visión artificial, empleando una cámara e histogramas de color adaptativos. Los resultados mostraron una gran variabilidad en las tasas de reconocimiento, desde un 50% a más del 95%, en función de los parámetros considerados.

4.3 Reconocimiento de escritura a mano

La evaluación del reconocimiento de la escritura a mano (*handwriting*) se realiza principalmente mediante la *tasa de reconocimiento* de símbolos (caracteres, dígitos o palabras). Por ejemplo, Yasuda et al. (2000) obtuvieron una tasa de reconocimiento de 85,35% en el reconocimiento de varios conjuntos de datos que contenían caracteres japoneses Kanji (primera categoría JIS), Katakana y Hiragana. Por otra parte, Vo y Wood (1996) comprobaron que el sistema *Jeanie* era capaz de realizar el reconocimiento de la escritura a mano, realizada de forma continua (cursiva) e independiente del usuario, con una tasa de reconocimiento de más de 90% para un vocabulario de 20.000 palabras. Además de la tasa de reconocimiento, se utilizan otras medidas de evaluación. Por ejemplo, LeCun et al. (1995) se centraron en el *tiempo de reconocimiento* y los *requisitos de memoria* al evaluar diversos algoritmos de reconocimiento.

4.4 Fusión de información multimodal

La medida de evaluación más usual para la fusión de información multimodal es el porcentaje de *interpretaciones semánticas* correctas, las cuales se pueden obtener mediante la combinación y cooperación de las diversas modalidades de entrada.

4.4.1 Habla y gestos

Entre otros, Vo y Wood (1996) utilizaron este criterio para evaluar el sistema *Jeanie*, tomando como referencia un conjunto reducido de interacciones del usuario que contenía 77 palabras, 57 gestos realizados con el lápiz sobre una pantalla táctil, y 52 combinaciones de ambas modalidades. Los resultados obtenidos indican que, en ausencia de errores de reconocimiento de ambos reconocedores, el módulo de fusión obtiene un 80% de interpretaciones semánticas correctas. Sin embargo, si el porcentaje de reconocimiento de palabras era tan sólo del 76%, la tasa de comprensión disminuía hasta un 62%. La diferencia de 18% se debía a un 15% de errores de RAH y a un 3% de errores en el reconocimiento de gestos.

4.4.2 Habla e información visual

La información visual extraída de la cara del usuario es muy importante para el RAH, y en especial en aquellas situaciones donde el canal auditivo está degradado por el ruido, filtrado por el ancho de banda o limitado por discapacidades auditivas. Diversos autores han evaluado la fusión de ambas modalidades en sistemas de AVASR centrándose en *estudios de percepci*ón. Este es el caso de los primeros trabajos de Risberg y Lubker (1978) en los que se muestra cómo un conjunto de sujetos que ve a otra persona hablar, sin poder escucharla, es capaz de percibir el 1% de las palabras pronunciadas. Si los sujetos sólo oyen la voz, degradada por un filtro paso-baja, perciben un 6% de las palabras, y si recibe ambas señales (visual y acústica) perciben el 45% de las palabras.

Otros autores han evaluado la fusión de ambas modalidades considerando la *exactitud en el reconocimiento* de símbolos. Por ejemplo, Yang et al. (1998) evaluaron un sistema de AVASR entrenado con 170 secuencias de datos acústico/visuales, con objeto de reconocer 30 secuencias en las que se había añadido ruido blanco. La Tabla 1 muestra los resultados obtenidos, los cuales indican que el uso de ambos tipos de información (acústica y visual) mejora claramente tas tasas de reconocimiento, especialmente cuando la SNR (Signal-to-Noise Ratio) es baja (8 dB).

Tipo test	Solo info. visual	Solo info. acústica	Info. combinada
Datos limpios	55%	98,4%	99,5%
SNR 16 dB	55%	59,6%	73,4%
SNR 8 dB	55%	36,2%	66,5%

Tabla 1. Resultados de reconocimiento

4.4.3 Habla, gestos e información visual

Varios investigadores han estudiado la fusión de habla, gestos e información visual para mejorar el rendimiento de los MDSs en varias facetas. Por ejemplo, Adelhardt et al. (2003) estudiaron estas modalidades de entrada para obtener información acerca del estado emocional del usuario. Su idea es que la prosodia de la voz puede indicar si el usuario está (o no) frustrado, la expresión facial puede indicar si el usuario está (o no) satisfecho, y los gestos pueden revelar su posible inseguridad. Los autores realizan la evaluación según la *exactitud en el reconocimiento* de cada modalidad. Los resultados obtenidos fueron de 32%, 76% y 77% para el reconocimiento de las expresiones faciales, la prosodia y los gestos, respectivamente. Según los autores, estos resultados posibilitan identificar el estado emocional del usuario, aunque en muchos casos no es necesario usar las tres modalidades simultáneamente.

4.5 Agentes animados

Los agentes animados se suelen usar en los MDSs para incrementar la inteligibilidad del los mensajes orales generados mediante síntesis de habla, y para mejorar la apariencia visual de la interfaz. Los resultados experimentales que ofrecen varios autores muestran que la "apariencia humana" del sistema induce a un comportamiento más social en algunos usuarios. La evaluación de estos agentes es una tarea muy compleja; aún no existen métodos estándar de evaluación, debido en gran parte a la gran complejidad y variedad de agentes, aplicados a diferentes dominios y aplicaciones (por ejemplo, guías virtuales, vendedores, etc.). La mayoría de las medidas de evaluación usan escalas de Likert o sentencias del tipo "Confío en el agente", "el agente parece una 'persona' amable", "el agente tiene mal carácter", "el movimiento de los ojos, labios, cabeza y otras partes del cuerpo parecen naturales".

4.5.1 Efecto del incremento de efectividad y naturalidad

Un método básico para evaluar agentes animados es identificar en qué medida influyen realmente en el diálogo, es decir, qué efectos producen en la comunicación con el usuario. Ello puede parecer obvio en algunas aplicaciones, pero no es tan fácil en otras. Algunos autores han estudiado las respuestas de los usuarios a diferentes tipos de agentes animados. Por ejemplo, Cassell y Thorisson (1999) estudiaron tres clases de agentes: unos que sólo suministran respuestas en forma transaccional, otros que incluyen además información acerca de su estado emocional, y por último, agentes que proporcionan una respuesta transaccional acompañada de un estado emocional, mediante gestos, miradas, etc. El resultado del estudio indicaba que, según los usuarios de test, el último tipo parece más natural y proporciona una interacción más eficiente.

4.5.2 Incremento de inteligibilidad

Otros autores han evaluado el incremento de inteligibilidad del habla cuando se usan agentes animados, comparado con la inteligibilidad cuando sólo se usan estímulos acústicos. Por ejemplo, LeGoff et al. (1996) realizaron tests de inteligibilidad considerando cinco niveles de degradación de la señal acústica a causa de ruido. Usando sólo la voz, la inteligibilidad fue comparada con un modelo de movimiento de labios, un modelo facial y la cara del locutor original. Los resultados confirmaron la importancia de la información visual en la percepción del habla: la visión de la cara completa proporcionó dos terceras partes de la inteligibilidad acústica cuando la transmisión de ésta era degradada o se perdía; el modelo facial (sin movimientos de la lengua) proporcionaba la mitad, y únicamente el movimiento de labios proporcionaba una tercera parte.

Usando la misma medida de evaluación, Beskow (1997) evaluó los agentes animados Parke y Olga. En el experimento

participaron 18 sujetos y se usaron dos voces sintéticas (de hombre y mujer) además de voz natural (de hombre y mujer) en ocho condiciones acústicas diferentes, con la señal de audio degradada por ruido a una SNR de 3dB. Los resultados indicaron que con la voz sintética de hombre, la tasa de inteligibilidad era del 30% cuando únicamente se usaba el canal del audio, de 45% en el caso audiovisual (agente Parke) y de 47% en el mismo caso (agente Olga). Usando la voz natural, la inteligibilidad variaba desde el 62% (sólo audio) al 70% con el agente Parke.

4.5.3 Efecto de la expresión facial

Otros investigadores han estudio el efecto que producen en el diálogo las expresiones faciales del agente animado. Por ejemplo, Koda y Maes (1996) estudiaron el efecto producido por un sistema diseñado para jugar al póquer, considerando como medidas de evaluación la atención requerida, el compromiso y la distracción causada en los usuarios. Los autores estudiaron asimismo el tipo de características faciales (p. e. edad, realismo de movimientos, etc.) que hacen parecer inteligente al agente, y confortable durante la interacción. Los autores también estudiaron las opiniones de los usuarios acerca del agente animado, con objeto de determinar si éstas venían determinadas por su apariencia, su funcionamiento o por ambos factores a la vez. Asimismo, estudiaron si las opiniones de los usuarios dependían de su género, o de sus criterios acerca del fenómeno conocido como *personificación* [1] (*persona effect*). Los resultados de la evaluación, realizada mediante una escala Likert de 7 puntos, indicaron que los usuarios consideraron al sistema igualmente inteligente con independencia de que usara el agente animado o no. Ello sugiere que usar un agente animado no incrementa la percepción de inteligencia de un sistema de diálogo. Sin embargo, cuando el sistema contaba con el agente animado, éste era considerado más "deseable", más atractivo y más confortable a la hora de jugar con él, con independencia de la opinión de los usuarios acerca de la personificación.

Los agentes animados también han sido evaluados teniendo en cuenta su éxito (o no) a la hora de transmitir emociones específicas mediante expresiones faciales. Por ejemplo, Fabri et al. (2002) evaluaron el uso de un pequeño conjunto de unidades de acción FACS (*Facial Action Coding System*) (Ekman y Friesen 1978) para generar seis expresiones básicas (sorpresa, temor, disgusto, ira, felicidad y tristeza) además de una expresión neutra. Tras mostrar aleatoriamente 28 imágenes de agentes animados a 29 sujetos, se realizó un estudio estadístico que mostró una tasa de reconocimiento de las emociones del 62,2%.

5 Conclusiones

En la literatura puede encontrarse una gran variedad de métodos de evaluación que son aplicados por los investigadores para medir aspectos concretos de sus sistemas de diálogo. Sin embargo, no existen unos métodos claramente aceptados por la comunidad científica, que puedan considerarse *estándares de evaluación*. Ello, sumado a la gran diversidad de tareas, diferencias de complejidad de unos sistemas frente a otros (en cuanto a modalidades de interacción, características del agente animado, etc.) provoca que sea muy difícil comparar los sistemas entre sí de una forma objetiva. Por ejemplo, dada la disparidad de criterios, algunos autores proporcionan resultados que muestran claramente la conveniencia de utilizar agentes animados, mientras que otros autores proporcionan datos en otra dirección. Por consiguiente, son necesarias técnicas estándar que faciliten la evaluación de tales sistemas, teniendo en cuenta sus características específicas, así como métodos que faciliten la comparación de sistemas diseñados para realizar tareas diferentes.

[1] Por este fenómeno se conoce al supuesto efecto beneficioso producido por los agentes animados en la interacción con sistemas de diálogo.

6 *Referencias*

Adelhardt, J., *et al.* 2003. Multimodal user state recognition in a modern dialogue system. Proc. 26th German Conference on Artificial Intelligence

Almeida, L., *et al.* 2002. Implementing and evaluating a multimodal and multilingual tourist guide. Proc. ISCA Tutorial and Research workshop on multimodal dialogue in mobile environments

Beringer, N., Katerina, L., Penide-López, V., Türk, U. 2002. End-to-end evaluation of multimodal dialogue systems – can we transfer established methods?. Proc. 3rd Language Resources and Evaluation Conference, pág. 558-563

Beskow, J. 1997. Animation of talking agents. Proc. ESCA Workshop on Audio-visual Speech Processing, pág. 149-152

Cassell, J., Thorisson, K. R. 1999. The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents. Applied Artificial Intelligence, 13, pág. 519-538

Ekman, P., Friesen, W. 1978. Facial action coding system. Consulting Psychologist Press

Fabri, M., Moore, D. J., Hobbs, D. J. 2002. Expressive agents: Non-verbal communication in collaborative virtual environments. Proc. Workshop on embodied conversational agents – let's specify and evaluate them!

Hu, C., Meng, M. Q., Liu, P. X., Wang, X. 2003. Visual gesture recognition for human-machine interface of robot teleoperation. Proc. IEEE/RSJ International conference on Intelligent Robots and Systems

Kettebekov, S., Sharma, R. 2000. Understanding gestures in multimodal human computer interaction. Proc. Int. Journal on Artificial Intelligence Tools, pág. 205-224

Koda, T., Maes, P. 1996. Agents with faces: The effect of personification. Proc. 5th IEEE Int. Workshop on robot and human communication, pág. 189-194

LeGoff. B., Guiard-Marigny, T., Benoît, C. 1996. Analysis-synthesis and intelligibility of a talking face. In: J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (Eds.) Progress in speech synthesis, Springer-Verlag

López-Cózar, R. 2003. Uso de canales adicionales en los sistemas conversacionales. Procesamiento del Lenguaje Natural, nº 30, pág. 89-97

LeCun, Y. *et al.* 1995. Comparison of learning algorithms for handwritten digit recognition. Proc. International Conference on Artificial Neural Networks, pág. 53-60

Mellor, B., Baber, C. 1997. Modelling of speech-based user interfaces. Proc. Eurospeech, pág. 2263-2266

Montero, J. A., Sucar, L. E. 2004. Feature selection for visual gesture recognition using Hidden Markov models. Proc. Fifth International Conference in Computer Science, pág. 196-203

Risberg, A., Lubker, J. L. 1978. Prosody and speechreading. Quarterly Progress & Status Report 4, Speech Transmission Laboratory, KTH, Stockholm, Sweden

Rogozan, A., Deléglise, P. 1998. Adaptive fusion of acoustic and visual sources for automatic speech recognition. Speech Communication, 26 (1-2), pág. 149-161

Rowley, H. A., Baluja, S., Kanade, T. 1998. Neural network-based face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, nº 1, pág. 23-28

Sibert, L. 2000. Evaluation of eye gaze interaction. Proc. Human Computer Interaction, pág. 281-288

Sturm, J., Bakx, I., Cranen, B., Terken, J., Wang, F. 2002. Usability evaluation of a Dutch multimodal system for train timetable information. Proc. Language Resources and Evaluation Conference

Vo, M. T., Wood, C. 1996. Building and application framework for speech and pen input integration in multimodal learning interfaces. Proc. International Conference on Acoustics, Speech and Signal Processing, Atlanta, GA

Yang, J., Stiefelhagen, R., Meier, U., Waibel, A. 1998. Real-time face and facial feature tracking and applications. Proc. Workshop on audio-visual speech processing, pág. 79-84

Yasuda, H., Takahashi, K., Matsumoto, T. 2000. A discrete HMM for online handwriting recognition. Pattern recognition and Artificial Intelligence, 14(5), pág. 675-689

Wachs, J., Kartoun, U., Stern, H., Edan, Y. 2002. Real-time hand gesture telerobotic system using fuzzy C-Means clustering. Proc. 12th Annual Conference of Industrial Engineering and Management

Wahlster, W. 2002. SmartKom: Fusion and fission of speech, gestures, and facial expressions. Proc. First International Workshop on Man-Machine Symbiotic Systems, pág. 213-225