http://www.virtualeduca.org
Palacio Euskalduna, Bilbao 20-23 de junio, 2006

SINCRONIZACIÓN DE LABIOS: MÉTODO SIN VISEMAS.

[1]Cabiedes F. [1]Pelczer, I., [1]Gamboa, F., [2]Bretón, J., [2]Rodríguez, S.

[1] Laboratorio de Interacción Humano Máquina y Multimedios,
Centro de Ciencias Aplicadas y Desarrollo Tecnológico, UNAM.
[2] Dirección de Investigación e Innovación-Enciclomedia, ILCE.

Introducción

En la ultima década ha crecido considerablemente el numero de desarrollos de sistemas tutoriales inteligentes. Una categoría de estas aplicaciones incluye agentes pedagógicos animados (personajes de computadora diseñados con el propósito de facilitar el aprendizaje). Un agente pedagógico animado tiene movimientos, expresividad emocional y, en muchas ocasiones, comunicación verbal de la información. Estas características no siempre son presentes en misma medida en las aplicaciones, por ejemplo en el trabajo de Conati (REF.) la información se trasnmite mediante cuadros de texto, igual que en ADELE (Shaw *et al.*, 1999), mientras que HERMAN (Lester, Stone y Stelling, 1999), COSMO (Lester *et al.,* 1999) y PPP PERSONA (André, Rist y Muller, 1999) son agentes pedagógicos con movimientos complejas, expresividad emocional y comunicación verbal.

Varios estudios confirman el efecto positivo que la presencia de los tutores tiene sobre el desempeño de un estudiante. En Lester *et al.* (1997b) los autores presentan los resultados de una evaluación a gran escala del impacto pedagogico de tal agente, mientras que en Lester *et al.* (1997a) se estudia el efecto de la presencia de tal agente sobre la motivación del estudiante. Las principales conclusiones de estos trabajo son los siguientes:

- Los estudiantes que interactuaron con un medio de aprendizaje con un agente pedagógico animado mostraron un crecimiento estadísticamente significativo desde pre-test al post-test;

- en los experimentos con agentes que emplean modalidades tanto visuales (animación) como verbales (habla) los estudiantes mostraron mejoramiento considerablemente superior en solución de problemas que en experimentos con agentes sin habla y estaticos.

- el agente pedagógico animado tiene un efecto de fuerte motivación para el estudiante.

A pesar de la complejidad de los medios (verbal, visual, afectivo) mediante cuales los agentes pedagógicos animados pueden transmitir la información, dichos sistemas presentan las siguientes limitaciones:

- toda la información que se transmite es predefinida por los creadores de los sistemas al momento de diseñar el sistema;

- los sistemas no prevén sincronización de labios con el mensaje verbal.

Consideramos que la posibilidad de añadir posteriormente archivos audio (con preguntas y/o explicaciones o versiones de idiomas) al sistema por el mismo usuario (o un profesor) es importante ya que la variedad de preguntas y explicaciones que los estudiantes requieren varia no solamente con el nivel de conocimiento y con los intereses que ellos tienen sino también según el contexto socio-cultural.

Por lo anterior en este trabajo nos proponemos atender esta necesidad y establecemos dos metas para seguir:

1. definir un proceso que permita añadir al sistema nuevos archivos audio, desde su registro hasta la integración en el funcionamiento del sistema;
2. definir un algoritmo que permita la asociación automática de los movimientos de labios al archivo de audio.

Por otra parte La búsqueda por producir personajes animados creíbles ha llevado entre otros, hacia resolución del problema de sincronización del audio con los movimientos bucales. Sobre este particular se ha desarrollado un número creciente de programas de cómputo y plug-ins de origen comercial. Estos sistemas se basan en la correlación entre un archivo de audio y uno de texto, (figura 1) en el cual se encuentra la trascripción del discurso del archivo de audio, lo que permite la correcta selección de visemas y fonemas. Sin embargo, ninguno de estos sistemas permiten realizar procesos automatizados o en tiempo real, ya que todos dependen de ajustes manuales de los fonemas en correlato con la aparición del discurso en el tiempo (figura 2).

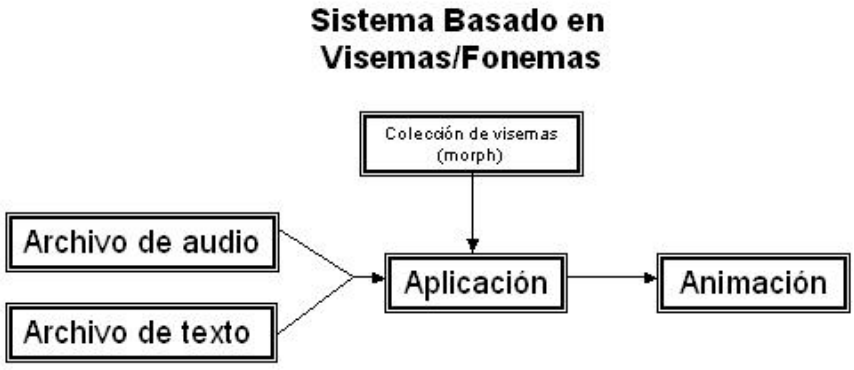
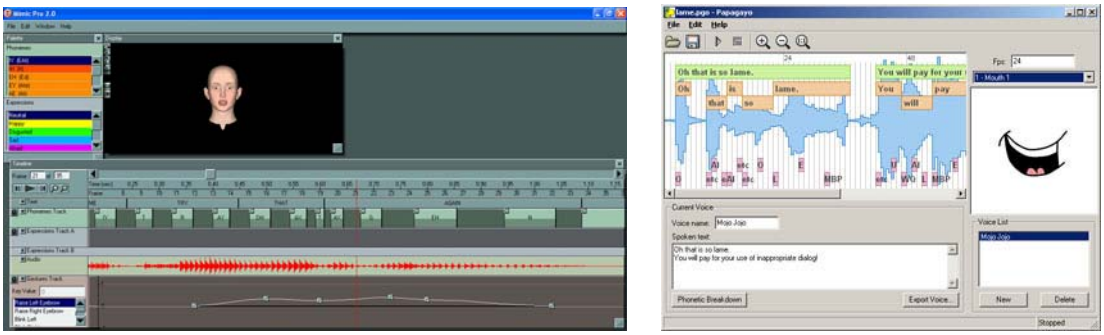


figura 1



a) *b)*

figura 2 sistemas de Lipsync comerciales basados en fonemas y visemas
a) Mimic 2.0 elaborado para figuras en 3D b) Papagayo 1.0 Elaborado para figuras en 2D (Flash)

A pesar de ser una solución en boga, resulta insuficiente cuando de lo que se trata es de producir movimientos faciales de manera automática o en tiempo real, además de requerir granes cantidades de horas-hombre para conseguir la correspondencia entre los visemas y el archivo de audio.

En el presente trabajo presentamos un sistema simple, que permite realizar análisis del audio base, generar un archivo de análisis y correlacionar dicho archivo con un grupo particular de bocas que a su vez cambian en el tiempo, lo que permite simular la acción de sincronización labial.

El problema.

La idea es construir un grupo de Avatares que intervienen bajo ciertas condiciones en un programa. Dichos avatares, serán capaces de exponer preguntas relacionadas con los materiales en los que se encuentran embebidos. La selección del género de los avatares es aleatoria, y las preguntas varían dependiendo de cada sección del software, de tal forma que una solución como la basada en visemas resulta inoperante. Así que se prepararon los avatares sin la boca, y se aplica esta última en el momento de presentarse la pregunta. El sistema de la boca es capaz de correlacionar el archivo de análisis con los grupos vocales almacenados y desplegarlos sincronizados con el audio.

El primer paso del trabajo es el análisis de un archivo que contiene una lectura en voz alta, el texto de la pregunta, este análisis permite la identificación automatizada de partes de habla y silencio. Después de haber identificado dichas partes, se integra el resultado a un programa que permite sincronización de movimientos de boca con el archivo de sonido. La animación automatizada de la boca en sincronía con el archivo se trata posteriormente en este trabajo.

Aunque simple en su formulación, estos pasos imponen buscar respuestas a varias preguntas, por ejemplo:

- ¿cual es la influencia del formato del archivo de sonido sobre el proceso de análisis?

- ¿cual es la resolución necesaria (o mínima) de muestreo para que tal tarea se pueda llevar a cabo?

- ¿como se puede caracterizar el silencio? Y en consecuencia ¿como se puede diferenciar el ruido de fondo del silencio?

- al considerar la parte de animación, hay que definir ¿que resolución de cuadros usar para la animación?

- ¿como reducir la granularidad de la información contenida en la muestra?

- ¿como elegir un diseño de boca con base a la información de muestreo?

- ¿cual es el numero mínimo de animaciones bucales necesarias para obtener movimientos percibidos como naturales?

Responder a estas preguntas impone en mismo tiempo determinar el proceso de trabajo, desde la concepción general de como se integra un componente de sincronización a un proyecto mas grande hasta determinar los detalles referentes al archivo de audio por ejemplo; codificación usada, resolución de muestreo, condiciones generales para registrar el archivo etc.

Hay que resaltar que el trabajo, formulado en los términos anteriores, se basa en una hipótesis muy fuerte, lo cual es que es posible detectar zonas de habla y silencio y hacer la animación de la boca (sincronización de los movimientos de la boca) en sincronía con el archivo de sonido basándose solamente sobre la información contenida en el muestreo.

Vamos a partir de tal hipótesis aunque evidentemente es una simplificación bastante fuerte. Los humanos percibimos los movimientos de la boca y tenemos la expectativa natural que estos correspondan al texto o sonido que se escucha, es decir, la expectativa es de ver visemas asociadas adecuadamente al sonido que se escucha. Sin embargo, la información contenida en un archivo de sonido se refiere a la frecuencia en un momento de muestreo y por lo tanto no retiene ninguna información sobre que se dice en la grabación. En otras palabras; tenemos que asociar movimientos de boca basándonos solamente sobre una visualización grafica del archivo de sonido.

Varias investigaciones confirman la dificultad (incluso, la imposibilidad) de la tarea de sincronización sin tener el texto leído disponible. Nosotros, partimos de la hipótesis que es posible hacer una aproximación satisfactoria de los movimientos de la boca, sin la necesidad de llegar al nivel de visemas.

El articulo esta organizado de siguiente manera. En el capitulo 2 discutimos aspectos relacionados con el formato del archivo audio, mientras que el capitulo 3 contiene la descripción de la parte de vincular el resultado del análisis con la animación. En la última parte se describen las conclusiones (recomendaciones de trabajo) del trabajo y se mencionan las líneas para seguir en el futuro.

A continuación se describe la manera en cual hemos abordado las metas presentadas. Primero se presentará el algoritmo para asociar movimientos de labios con el archivo audio y luego el proceso de integración a un sistema.

Metodología

1. Análisis del archivo de sonido

1.1. Formato del archivo

Los formatos más usuales para archivos de sonido son el WAV y MPEG, de cual lo más conocido es el MP3. El formato MP3 se volvió muy popular dado que la codificación usada permite reducir drásticamente la dimensión del archivo de sonido y guardar en mismo tiempo la calidad del sonido. Por lo tanto, también han proliferado los programas para leer estos archivos. El principio de organización consiste en guardar la información organizada en cuadros (frames), cada cuadro tiene al principio las especificaciones (header). Tal organización permite escuchar pedacitos del archivo de sonido, porque toda la información necesaria para tocarla esta contenido en el mismo cuadro. En mismo tiempo cada cuadro tiene longitud, bit rate y sample rate variables.

El formato WAV está estandarizada, lo que representa una gran ventaja cuando se trata de analizar los datos contenidos. El principio de organización es diferente del MPEG, aquí tenemos una organización global de la información. EEEL archivo de sonido tiene una parte descriptiva (header) que contiene información sobre todo el archivo de sonido. Esta parte del archivo esta seguida por los datos correspondientes al muestreo. Por la organización del archivo tenemos datos correspondientes a momentos de tiempo iguales. La gran desventaja de un archivo de sonido en este formato es la limitación del tamaño, dado que en la parte de descripción del formato hay disponibles solamente 4 bytes para guardarlo.

El proceso de análisis debe rastrear el archivo y extraer los datos de muestreo. Hemos efectuado varios experimentos con los dos formatos y hemos llegado a la conclusión que el formato MPEG regresa datos que no corresponden con el sonido (al escuchar). En la figura 1 se presenta la comparación entre los resultados del análisis del mismo archivo de sonido, pero en los dos formatos, WAV y MP3. Se puede observar que hay diferencias fuertes aunque estamos hablando del mismo archivo. En frente de tal situación hemos decidido de usar para el análisis el formato WAV.

Archivo WAV

0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3100,6968,-1270,-1236,1901,2794,6090,3583,3286,-2791,-1225,8243,8630,-370,-219,-1162,30,0,0,0,0,0,0,0,0,0,-3,4635,-5663,-983,1598,2322,10471,-680,-389,-419,-10675,-14,-13269,6239,5948,99,1407,-34,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,

Archivo MP3

1,-1,0,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1025,-1,-1,0,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1025,-1,-1,0,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1025,-1,-1,0,-1,-5754,-29546,-5066,451,-30288,6151,25799,11056,-19292,22319,3372,32445,1267,17661,26380,-1025,11448,-5790,0,-17758,-27779,-1104,2944,-6384,-2139,22037,- 1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1025,-1,-1,0,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1025,-1,-1,0,-1,-1,-1,-1,-1,-1,-1,-1,-1,0,0,0,0,0,8224,

Figura 3 comparación de lectura
Ambos análisis corresponden a la misma grabación.

1.2. Ruido de fondo

Para hacer un análisis adecuado se necesita saber sobre la presencia del ruido de fondo. Al considerar que el ruido puede ser interpretado como dato real, es necesario saber de su presencia. Por lo tanto, hemos decidido que , en el proceso de grabación se incorpore un segundo de "silencio" para registrar el ruido inherente del micrófono y al ambiente. Otra alternativa es de hacer las grabaciones en condiciones de estudio, lo cual reduce enormemente las posibilidades de hacerlo sin una infraestructura adecuada.

1.3. Resolución de los datos

El número de muestras por segundo es un parámetro del formato WAV, por lo tanto hay que leerlo e identificarlo adecuadamente para extraer los datos. A cada segundo vamos a tener un gran número de datos. En caso de tener ruido de fondo, guardado en el primer segundo del archivo, se ajustan los datos leídos para eliminar el ruido de fondo. Una vez obtenidos las muestras nos encontramos con el problema de reducir la granularidad de la información, es decir extraer un número menor de datos que se puedan considerar como representativos para un periodo de tiempo deseado (un segundo o fracciones de segundo).

El problema se torna complejo, por que hay que determinar criterios para comprimir datos. Esto lo convierte en un problema que pertenece al área de compresión y codificación de datos.

Vamos a ilustrar el problema con el siguiente ejemplo: supongamos que tenemos 25 datos correspondientes a un cuarto de segundo y queremos asignar un solo valor que contenga la "esencia" de los 25. Consideramos los siguientes 4 casos (para simplificar vamos a usar solamente 1 y 0's, donde 1 marca la presencia de un dato y 0 representa "silencio").

a. 0001110001110001110001111

b. 00000000000000111111111111

d. 11111111111000000000000000

d. 11111111111100000000000000

El propósito es asignar un solo valor a estas cadenas. ¿Qué criterios se podrían usar? En principio se podría pensar en decidir con base en el promedio de 1's, si es mayor a la 0.5 entonces el valor asignado sea 1, en caso contrario 0. Si aplicamos este criterio solamente en el caso d. tendremos 1. Sin embargo es evidente que estamos perdiendo una información valiosa, dado que en todos los casos el promedio es muy cercano a 0.5.

Otra idea sería la de decidir dependiendo de la secuencia de 1's y 0's tomando un criterio mas complejo con algún promedio ponderado (según un peso asociado de manera subjetiva con la posición de aparición en la cadena o otro criterio). La conclusión es que no hay garantía de que nuestros criterios funcione satisfactoriamente en todas la situaciones. Esta conclusión tiene un gran impacto sobre el proceso de trabajo, porque significa que podemos buscar soluciones adecuadas a una situación. Para tal propósito vamos a tener que analizar y aprovechar las particularidades del contexto en cual nos encontramos y necesitamos evaluar el costo de una error de evaluación en el desempeño de la animación final.

2. Sincronización de boca

Al considerar los ejemplos presentados en los puntos a-d y al pensar que hay que asociar un diseño de boca a ellos, nos encontramos otra vez con el problema de la reducción de granularidad. Como no hay una solución universalmente satisfactoria, tenemos que elegir una que sea adecuada para la meta particular que tenemos.

La particularidad de la situación para cual buscamos solución esta ligada con la percepción humana. Tenemos que analizar varios aspectos:

- cuantas animaciones (de bocas diferentes) por segundo se perciben como "movimiento natural" de la boca;

-como se percibe si hay una falla, en el sentido de tener una boca cerrada en lugar de boca abierta;

- cuantos diseños de bocas necesitamos para tener una variedad que asegure la "naturalidad";

- como decidir el diseño particular por usar en un momento de tiempo

- como decidir cuando se pueden repetir los diseños.

Implementación

En la figura 4 se presenta un esquema del sistema implementado para leer archivos WAV y cuya salida es una animación bucal en Flash.

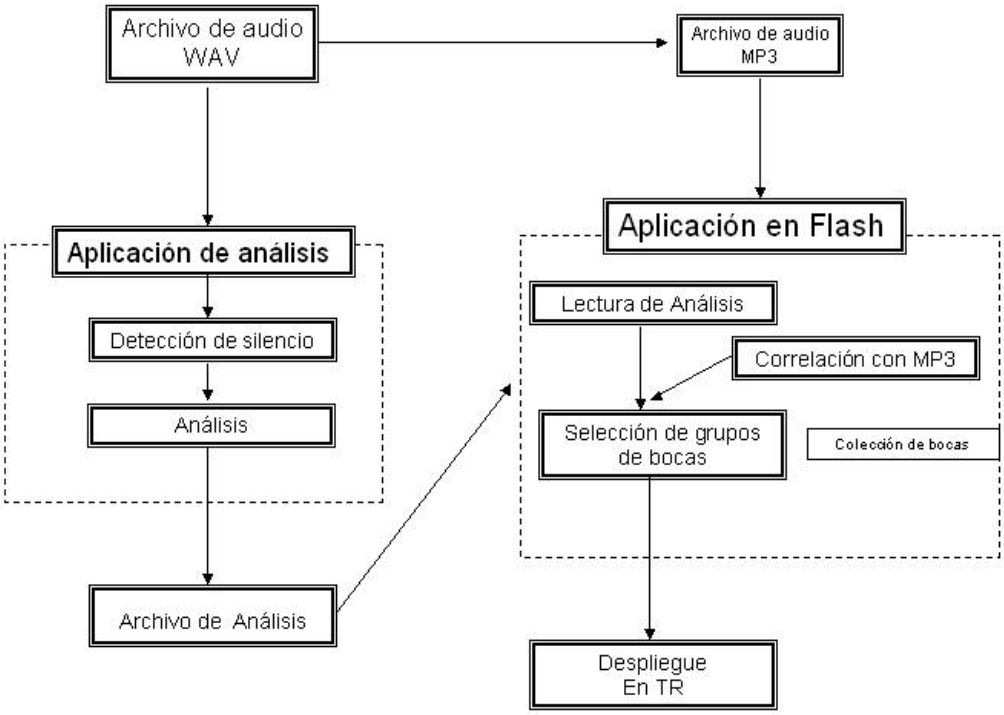


Figura 4.

Diagrama del sistema completo
Panel Izquierdo, Unidad de Análisis. Panel derecho Unidad de Despliegue Gráfico

En el panel derecho se diagrama la secuencia de trabajo de la Unidad de Análisis, la cual abre el archivo WAV, determina el silencio en el primer segundo, determina la velocidad de muestreo del análisis y produce un archivo de texto similar al panel superior de la figura 3, el archivo entonces es guardado en la base de datos de las preguntas junto con la versión en MP3 del audio, de tal suerte que al ser llamado por la Unidad de Despliegue Gráfico correspondan el audio y el análisis.

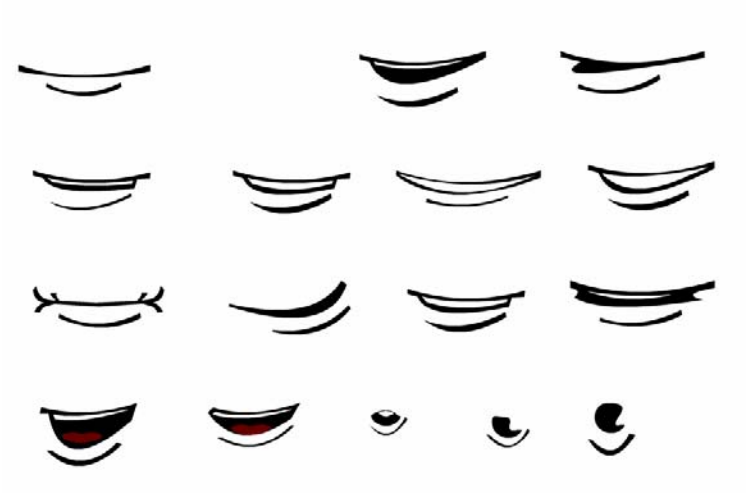


Figura 5 Colección de bocas
Primera a la izquierda se muestra "boca cerrada", cuyo valor en el análisis es 0

En el panel derecho se presenta la secuencia de pasos de la Unidad de Despliegue Gráfico, donde es leído el archivo de Análisis y correlacionado con el MP3 del audio, en cada paso de muestreo, se intercambian las bocas asumiendo que 0 es la que corresponde a boca cerrada, éste módulo contiene la colección de las bocas en acción (ver figura 5) y las presenta conforme se leen sincronizadas con el archivo de audio, permitiendo el despliegue gráfico en tiempo real.

Discusión

Independientemente de los problemas inherentes a la granuralidad del muestreo, el sistema actualmente presenta una solución satisfactoria a los requerimientos, dado que se presenta la animación de la boca en el contexto del avatar, el cual a su vez se encuentra animado.

La sincronización obtenida recuerda las caricaturas comerciales de los sesenta, dando una credibilidad mínima en los gestos bucales, con respecto al audio.

En futuros trabajos, estaremos obligados a mejorar el análisis del audio con la meta de poder utilizar el micrófono de cualquier máquina, con el fin de utilizar una modificación de la unidad de despliegue gráfico en la representación bucal en tiempo real de avatares que se encuentren en sistemas 3D, y poder así producir comunidades virtuales colaborativas en las que se comuniquen los usuarios a través de avatares "parlantes", lo cual permitirá relegar las comunicaciones de tipo "sólo texto" a una opción y no la única opción.

Bibliografía

André, E.; Rist, T. y Muller, J.: "Employing AI methods to control the behavior of animated interface agents". *Applied Artificial Intelligence*, 13: 415-448, 1999.

Conati, C. y X. Zhao. "Building and evaluating an intelligent pedagogical agent to imporve the effectiveness of an educational game". *International Conference on Intelligent User Interfaces, Proceedings of the 9th international conference on Intelligent User Interfaces*, Funchal, Madeira, Portugal, ACM Press, 2000.

Lester, J.C; Converse, S.A.; Kahler, S.E.; Barlow, S.T.; Stone, B.A. y Bhogal, R.: "The persona effect: Affective impact of animated pedagogical agents." *Proceedings of the Conference on Human Factors in Computing Systems*, 1997a.

Lester, J.C; Converse, S.A.; Stone, B.A., Kahler, S.E. y Barlow, S.T.: "Animated pedagogical agents and problem solving effectiveness: A large scale empirical evaluation". In *Proceedings of the Eighth World Conference on Artificial Intelligence in Education*, pp. 23-30, IOS Press, 1997b.

Lester, J.C.; Voerman, J.R.; Towns, S.G. y Callaway, C.B.: "Deictic believability: Coordinating gesture, locomotion and speech in life-like pedagogical agents". *Applied Artificial Intelligence*, 13: 383-414, 1999.

Lester, J.C.; Stone, B.A. y Stelling, G.D.: "Life-like pedagogical agents for mixed-initiative problem solving in constructivist learning environments". *User Modeling and User-Adapted Interaction*, 9: 1-44, 1999.

Shaw, E.; Ganeshan, R.; Johnston, W.L. y Millar, D.: "Building a case for agent-assisted learning as a catalyst for curriculum reform in medical education". In *Proceedings of the Ninth World Conference on Artificial Intelligence in Education*, IOS Press, 1999.