Maggie: el robot traductor

F. Alonso-Martín[1], Arnaud A. Ramey[1], Miguel A. Salichs[1]

[1]Robotics Lab, Universidad Carlos III de Madrid;

Resmuen. El objetivo de este artículo es mostrar los pasos necesarios para mostrar al robot social Maggie, desarrollando en la Universidad Carlos III de Madrid, como un agente traductor, capaz de leer textos escritos y sintetizar mediante voz la traducción de dicho texto en cualquier idioma. Para ello, se han integrado dentro de la arquitectura de control del robot tres nuevos sistemas: una herramienta de reconocimiento visual de caracteres OCR, un sistema de síntesis de voz multilenguaje, capaz de generar/sintetizar voz en varios idiomas y con diferentes emociones y por último un sistema de traducción online de frases.

1. Introducción

Tradicionalmente los robots han tenido un sistema pobre de expresión verbal si lo comparamos con los robots que se ven en las películas de ciencia ficción. En muchas de ellas los robots se pueden comunicar con expresiones y emociones típicamente humanas. Podemos mencionar robots como HAL (2001: A Space Odyssey) o Jonny Five (de Short Circuit). El concepto de una máquina hablando será cada vez más real debido a los constantes avances en los sistemas de síntesis de voz; sin embargo, en muchos casos la naturaleza de la voz generada por estas herramientas es uniforme y lineal, en un tono neutral y desprovista de la capacidad de expresar emociones o sentimientos.

Numerosos estudios y artículos, que podemos encontrar en literatura, muestran que, al igual que la música puede transmitir emocions(Curtis and Bharucha 2010)(Thompson, Schellenberg, and Husain 2004), el estado emocional de los humanos afecta directamente en el patrón acústico de la señal de voz generada (K R Scherer 1995)(Johnstone)(Breazeal 2003). En el clásico "The Expression of the Emotion in Man and Animals" del evolucionista Darwin, se comenta como hay varios modos de expresar emociones, entre las que podemos incluir la voz. Desde este punto de vista evolucionista, se discute sobre como la evolución del aparato fonador ha permitido una mayor habilidad para transmitir emociones y expresiones verbalmente. Esta perspectiva es

respaldada por otros estudios sobre cómo se transmiten las emociones entre lenguajes y culturas (K. R. Scherer, Banse, and Wallbott 2001).

La expresión verbal de emociones permite añadir cierta información extra al contenido del mensaje, que puede afectar al comportamiento de los oyentes y proporcionar ciertas pistas sobre la intención y el contenido global del discurso (Juslin and Laukka 2003).

Muchos autores han resumido el conjunto de emociones en cinco básicas: felicidad, furia, tristeza, mido y disgusto (otros autores incluso incluyen el estado de enamoramiento también) (Juslin and Laukka 2003). Estas investigaciones se han focalizado en determinar que factores y como influye cada uno de ellos en la capacidad de la voz para expresar emociones. Las principales variables que se han estudiado han sido las siguientes:

- Rango de frecuencia fundamental: si es un rango estrecho se puede asocial con estados emocionales de tristeza mientras que si es ancho nosotros podemos pensar en estados de agitación, estrés y nerviosismo.

- Pitch: si es de un ritmo alto transmite sensaciones de nerviosismo mientras que si es lento sugiere estados de tranquilidad o tristeza.

- Intensidad: un alto valor de intensidad conlleva un sentimiento de potencia o agresividad en la emoción, mientras que un valor bajo puede transmitir valores de debilidad o calma.

Esos mismos estudios hablan sobre familias de estados agrupándolos a su vez en dos conjuntos: estados fríos (baja velocidad, volumen e intensidad) y estados calientes (alto velocidad y volumen). Además otros estudios se han focalizado en estudias las características de cada estilo de comunicación, por ej; noticias en la radio, el discurso de un político, una conversación informal…(Roekhaut et al.). El artículo (Roekhaut et al.) muestra también como los actuales sistemas de síntesis de voz (TTS) sin ninguna modificación para cada país y agente comunicativo son incapaces de mostrar suficiente expresividad.

En este trabajo hemos integrado una herramienta de síntesis de voz TTS en nuestra arquitectura de control AD, una completa descripción de la arquitectura puede leerse en (Miguel Salichs et al. 2006). Esta arquitectura controla actualmente el robot social Maggie (ver **¡Error! No se encuentra el origen de la referencia.**) y en un futuro cercano nuevos robots. Partiendo de la herramienta que hemos integrado, que describiremos posteriormente en detalle, hemos generado cuatro estados emocionales con los que el robot se puede comunicar: felicidad, tristeza, nerviosismo y tranquilidad en los idiomas que el robot es capaz de hablar. Para obtener dichas emociones hemos ajustado algunos parámetros como: pitch, velocidad de palabras por minuto, volumen, énfasis, pausas y timbre, sin por ello lastrar la calidad de la voz sintetizada.

La integración del sistema de síntesis de voz, si bien es una pieza importante, no es la única tarea que hemos llevado a cabo para este trabajo. Como comentamos en el resumen hemos necesitado de otros dos componentes principales: un sistema de reconocimiento de caracteres (OCR) mediante visión y un sistema de traducción de frases entre idiomas. Con la unión de estos tres componentes, que pueden funcionar de manera independiente, les hemos dado una aplicación práctica y la mostramos como ejemplo en este trabajo, como sistema robótico de traducción de textos, combinando la parte visual y la parte verbal.

Por sistema de reconocimiento de caracteres nos referimos al sistema o aplicaciones dirigidas a la digitalización de textos. Identificando automáticamente símbolos o caracteres que pertenecen a un determinado alfabeto, a partir de una imagen para almacenarla en forma de texto plano. En los últimos años la digitalización de información (textos, videos, imágenes, etc.) se ha convertido en un punto de interés para la sociedad de la información. En el caso de los textos, existen enormes cantidades de información escrita no digitalizada, y por lo tanto difícilmente accesible desde medios electrónicos. En este contexto poder automatizar la introducción de caracteres evitando la entrada por teclado, implica un importante ahorro de recursos humanos y un aumento de la productividad, al mismo tiempo que se mantiene o se mejora la calidad de muchos servicios.

En la aplicación que le hemos dado al sistema de OCR como subconjunto de un sistema de traducción, nos posibilita una manera natural y cómoda de dirigirnos con el robot, ya que evita que el usuario tenga que usar el teclado para introducir los textos a traducir. Un sistema alternativo de entrada de información podría haber sido mediante la voz y un sistema automático de traducción de voz (ASR). Si bien este sistema también está integrado dentro de nuestra arquitectura de control, para esta aplicación en concreto presenta ciertos problemas: el usuario puede no saber pronunciar correctamente ciertos textos, el sistema de traducción debe soportar todos los idiomas a traducir mediante un sistema de reconocimiento no basado en gramáticas (dictado) sin necesidad de entrenamiento para el usuario, la tasa de fallos de las herramientas de ASR son mayores que las de OCR (sobre todo en ambientes ruidosos) y por último comentar que es una manera más lenta de entrada de información.

Por último y el más importante componente para el desarrollo de la aplicación aquí presentada es la del componente traductor de textos en varios idiomas. Para ello nos hemos valido del sistema de traducción online de Google1. Este sistema es capaz de traducir textos entre más de 50 idiomas, en forma de texto escrito. No es un sistema perfecto que proporcione siempre la mejor traducción, pero si es un sistema ligero de usar, lo suficientemente fiable y en continua mejora y evolución.

[1] http://translate.google.com/

En la sección 2 hablamos del robot Maggie a nivel físico y de software de control. En la sección 3 describimos el sistema de síntesis de voz y su integración dentro de la arquitectura de control. Dentro de la sección 4 mostramos el sistema de reconocimiento de caracteres y en la sección 5 hablamos sobre el sistema de traducción en-linea integrado. Finalmente en la sección 6 mostramos como funcionan todos estos componentes de manera coordinada para lograr que Maggie actúe como un traductor personal. Finalmente en la sección 7 se muestran las conclusiones del trabajo realizado.

Fig. 1 El robot Maggie

2. Entorno de trabajo

El robot Maggie (Miguel Salichs et al. 2006) es una plataforma de investigación en el estudio de la interacción humano-robot (HRI). Los asuntos de estudio están focalizados en encontrar nuevos modos de adaptar el potencial de los robots para proporcional a los usuarios humanos nuevas formas de trabajar, aprender y divertirse con ellos. Seguidamente vamos a describir brevemente el software y el hardware incluidos en el Robot Maggie.

2.1 Hardware

Maggie está diseñada como un robot de 1.35 metros de altura. Su base está constituida por dos ruedas motorizadas y una rueda de apoyo, además dicha base está dotada con 12 parachoques capaces de detectar el contacto con los objetos. Sobre la base, se encuentra un telémetro laser infrarrojo (Clase 1) capaz de detectar con precisión la distancia a los objetos más cercanos. Dentro de la "barriga" del robot se encuentra un emisor/receptor de infrarrojos programable que permite al robot controlar ciertos aparatos domésticos como televisores, cadenas de música, aires acondicionados, etc.

La parte superior del robot incorpora varios sensores de tacto capacitivos distribuidos por la superficie del cuerpo a modo de "piel sensitiva". Un tablet-pc esta situado en su pecho, proporcionando comunicación bidireccional entre el robot y los usuarios. A ambos lados se encuentran los brazos, dotados únicamente de un grado de libertad. En lo alto de la plataforma encontramos la cabeza del robot, con un diseño atractivo y con dos grados de libertad. Dentro de la cabeza se encuentra un lector de radio frecuencia (RFID) capaz de leer etiquetas de radiofrecuencia. La boca del robot está compuesta por una serie de LEDs capaces de iluminarse cuando el robot habla y dentro de ella se encuentra una webcam (Logitech QuickCam Pro 9000) capaz de proporcionar imágenes en tiempo real. Esto, junto con unos parpados animados, conforman el aspecto amigable de la cabeza de Maggie. Ver Fig. 2.

Recientemente, las capacidades físicas el robot se han ampliado gracias a la inclusión del sensor de Microsoft, Kinect[2]. Este potente sensor es capaz de proporcionar imágines a color y mapas de profundidad del entorno simultáneamente, adicionalmente se puede usar como micrófono para captar la voz de los usuarios. Entre otras cosas, puede ser usado para seguir personas y/o obtener la postura de los humanos que están en su campo de visión (esqueletización del cuerpo humano).

2 www.xbox.com/kinect

El mecanismo de captura de audio está basado en un micrófono auricular inalámbrico, aunque recientemente se está experimentando con el Kinect y con otros array de micrófonos integrados en el cuerpo del robot. Maggie puede hablar gracias a los altavoces incorporados en su cuello.

Maggie es controlada por un ordenador portátil situado dentro de su cuerpo. En este ordenador reside la arquitectura de control del robot que se describe en la siguiente sección. Notar que no es necesario que toda la arquitectura de control corra dentro de este ordenador, ya que dada la naturaleza distribuida de la arquitectura, cualquier componente de la misma puede estar en ejecución en diferentes ordenadores.

Fig. 2 La cabeza de Maggie

2.2 Arquitectura Software de Control AD

La arquitectura software que controla el robot ha sido desarrollado dentro del mismo grupo investigador que ha desarrollado el robot Maggie: el RoboticsLab[3] y recibe el nombre de Arquitectura Automática-Deliberativa (AD). AD está compuesta por dos capas de abstracción: el nivel automático y el nivel deliberativo. El nivel automático es donde se llevan a cabo las acciones de más bajo nivel de control. Aquí se localizan las primitivas de control que proporcionan la comunicación y control de los sensores, los motores y cualquier otro hardware. El nivel deliberativo corresponde con los módulos y habilidades que proporcionan un cierto grado de razonamiento y decisión de mayor nivel de complejidad.

El componente principal de la arquitectura es la "habilidad". Una habilidad es una entidad con la capacidad de razonamiento, procesamiento de información

[3] http://roboticslab.uc3m.es/roboticslab/

y capaz de llevar a cabo acciones al mismo tiempo que puede comunicarse con otras habilidades. Por ejemplo, la habilidad *laserSkill* maneja la información que proporciona el sensor de distancias y es capaz de hacerla accesible al resto de habilidades que necesiten, por ejemplo, hacer navegación por el entorno. Una descripción profunda y detallada sobre la arquitectura AD puede encontrarse en (R. Barber and Ma Salichs 2002).

3. El sistema de síntesis de voz

Muchos sistemas de síntesis de voz han sido usados y estudiados en el campo de investigación de la Interacción Humano-Robot. Nosotros hemos analizado y comparado los más populares y con mayor potencial para ser integrados dentro de nuestra arquitectura de control y por lo tanto para el trabajo aquí expuesto. Una comparación de los mismos puede verse en la Tabla 1.

Tabla 1Resumen de sistemas de TTS

Motores de TTS	Parámetros				
	Tono	Duración	Volumen	Calidad Voz	Notas
BabTTS	Uniforme	Pausas y velocidad configurables	Volumen general	pobre	Configurable el volumen de cada palabra
Natural Voices		Pausas y velocidad configurables a nivel de frase y de oración	Volumen controlable a nivel de frase y de oración	Pobre	
DECtalk	Variable: con acentuación de palábras	Pausas y velocidad configurable	Volumen general		
Naxpres					Todo controlado por el sistema automáticamente
Loquendo	Muy configurable, capaz de acentuar palabras concretas	Velocidad y pausas configurables		Muy bueno con todos los idiomas soportados	Con gestos de voz y controlable todo desde el propio esto
RealSpeak	Puede acentuar frases	Pausas y velocidad configurables	Volumen configurable	Bueno	
Festival	Acentuación a nivel global o de palabra			Muy bueno en inglés	
GnuSpeech			Controlable la amplitud		Configurable a bajo nivel
Google TTS				Algunos idiomas muy	Ningún parámetro controlable, voz uniforme.

				pobre.	

La más importante cualidad a analizar es la calidad de la voz, pero analizar esta característica puede ser una tarea bastante subjetiva. Nosotros hemos probado con los que hemos examinado en detalle los que sintetizan la voz lo más natural y fácil de entender posible, en este sentido hemos trabajado con Loquendo TTS, Nuance RealSpeak, Festival, Mbrola y Google TTS. Además estas herramientas pueden trabajar con varios idiomas, no siendo de igual calidad el idioma generado para todos los idiomas que dan soporte.

De los sistemas analizados, por el que nos hemos decantado inicialmente ha sido por el sistema Loquendo TTS, por su elevada calidad de síntesis y expresividad para todos los lenguajes soportados y por su relación calidad/precio. Este framework puede sintetizar voz en más de 50 idiomas, entre los que se incluyen español, inglés americano, inglés británico, etc. además de ser la herramienta con mayor potencial de configuración y personalización. Es posible cambiar el timbre de voz, el volumen, la velocidad de habla, la duración de las pausas, la expresividad de ciertas palabras, además de permitir expresar ciertos gestos de voz como son sonrisas, bostezos, cosquillas, silbidos, lloros, besos … que dotan al sistema de una enorme capacidad de expresividad.

Como dijimos en la introducción el objetivo perseguido es conseguir la expresividad propia de los humanos, y por lo tanto ser capaces de reflejar en la voz ciertos matices o emociones lingüísticas, por ello, hemos ajustado los parámetros que nos posibilita Loquendo TTS para conseguir cuatro emociones básicas: felicidad o alegría, tristeza, tranquilidad y nerviosismo. Todas ellas están disponibles en los idiomas que somos capaces de sintetizar. Actualmente los idiomas que soportamos son el Español y el Inglés (británico y americano).

La herramienta de TTS de Loquendo ha sido integrada, por lo tanto, en nuestra arquitectura en forma de "skill". Esta habilidad la hemos bautizado con el nombre de "emotional Text To Speech Skill" (eTTS Skill), y permite que cualquier componente de la arquitectura pueda sintetizar voz de manera controlada.

Cuando hablamos de que la síntesis de voz debe hacerse de manera controlada, nos referimos a la supervisión necesaria por parte de esta habilidad para que no haya dos componentes de la arquitectura generando voz simultáneamente, o que las frases a sintetizar se entremezclen las de un módulo con las de otro o se corten inesperadamente. También entendemos por control del sistema de voz, la capacidad para saber cuando una locución empieza a sintetizarse, cuando finaliza, cuantas locuciones se encuentran encoladas en el sistema, identificar a que componente del sistema pertenece cada locución, etc. Todos estos mecanismos que ofrece la habilidad desarrollada los vamos a detallas continuación:

Sintetizar voz con parámetros (texto, idioma, emoción, modo): el sistema es capaz de sintetizar cualquier texto en voz. Para ello se necesita indicar el texto a sintetizar, el idioma al que pertenece dicho texto (actualmente manejamos 2 idiomas), la emoción con la que se quiere expresar (actualmente tenemos 4 posibles emociones) y el modo (encolar la locución al final de la cola o introducirla en el primer lugar).

Parar/pausar/reanudar la síntesis de voz: el sistema puede pausar o parar definitivamente las locuciones, bien parando la locución actual en el momento exacto en el que está o bien esperando que acaba de sintetizarse. Si la síntesis de voz se para también se puede reanudar posteriormente.

Control del volumen: mediante la habilidad de síntesis de voz podemos especificar que se suba o se baje el volumen con el que está actualmente hablando el robot.

Control de la velocidad de locución: cada emoción lleva asociado una velocidad de síntesis (palabras por minuto), pero se puede variar par que el robot hable más lento o más rápido.

Consulta de locuciones en cola: el sistema es capaz de mantener una estructura de datos con la información de las locuciones pendientes de encolar. Hay que tener en cuenta, que no se deben sintetizar más de una locución al mismo tiempo (ya que el efecto sobre la interacción robótica sería de coarticulación) y sí que es posible que varios módulos de la arquitectura soliciten "hablar" antes de que la última locución haya sido sintetizada. Toda esta información se puede consultar.

Modo de secuestro de la voz: si un módulo de la arquitectura quiere acaparar el sistema de voz, puede hacerlo. De esta forma ningún otro modulo interferirá en la interacción por voz hasta que el secuestrador libere el sistema de voz.

Consulta si capacidad de síntesis disponible inmediatamente: muchas veces interesa sintetizar voz únicamente si la frase a pronunciar se puede hacer en ese mismo instante, ya que si se encola, pierde totalmente su significado. Ej: si una persona toca una parte del robot, el robot se puede expresar con una carcajada en el mismo instante en el que es tocado.

Como comentamos en la introducción, para conseguir simular la expresión mediante emociones, hemos ajustado los siguientes parámetros:

Tabla 2 Parámetros del sistema de síntesis de voz

	Feliz español/inglés	Tranquilo español/inglés	Nervioso español/inglés	Triste español/inglés
Tono	37/0	4/0	50/4	-8/-30
Velocidad	-3/-15	-11/-30	24/10	-11/-11
Timbre	28/27	27/27	33/31	27/27
Volumen	56/56	56/56	56/56	56/56

Recientemente y gracias a las posibilidades que nos permite nuestra arquitectura de control en dos capas, hemos podido integrar de manera sencilla nuevos motores de síntesis de voz, que permiten también sintetizar voz, de tal manera que el usuario puede elegir en todo momento que motor de síntesis de voz usar por la habilidad ettsSkill. Ver Fig. 3. Estas nuevas herramientas de síntesis de voz gratuitas, permiten crear nuevos "personajes" distintos a los usados por Loquendo, así como usar el sistema de voz en el caso de no poseer licencia de uso de la herramienta comercial de Loquendo TTS. Las herramientas que hemos añadido recientemente son Festival y Google TTS.

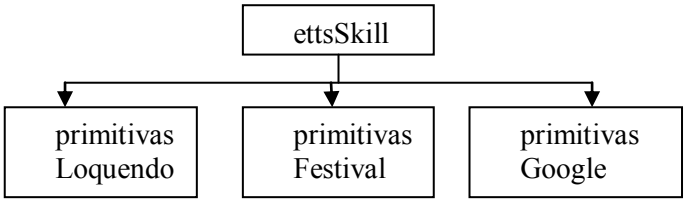

Fig. 3 Sistema de Voz en AD

La inclusión de Festival en nuestra arquitectura de control requiere la instalación de dicho framework para que nuestro motor de síntesis pueda usar su API de programación, en cambio el sistema de síntesis de Google es un sintetizador "en la nube", esto es requiere de conexión a internet. La generación de frases mediante el Google TTS no es un sistema actualmente sujeto a una API propiamente dicha, pero si se puede usar, haciendo una petición HTTP al servicio de traducción que describiremos en la siguiente sección y obteniendo mediante el comando de Linux "wget" el archivo de audio que se recibe al sintetizar el texto que enviamos para traducir.

Estos dos nuevos sistemas de síntesis incluidos en AD, gozan del mismo control que proporcionaba la habilidad *ettsSkill* anteriormente sólo a las primitivas de voz construidas sobre el API de Loquendo, esto es, consulta de locuciones en cola, secuestro de la voz, capacidad de encolar locuciones, etc.

4. El sistema de reconocimiento de caracteres

Un sistema de reconocimiento de texto identifican automáticamente símbolos o caracteres que pertenecen a un determinado alfabeto, a partir de una imagen para almacenar en forma de texto plano la interpretación realizada de la imagen.

Partiendo de una imagen perfecta, es decir, una imagen con sólo s niveles de gris, el reconocimiento de estos caracteres se realizará básicamente comparándolos con unos patrones o plantillas que contienen todos los posibles caracteres. Pero en la vida real existen numerosas dificultades, entre las que mencionamos las siguientes:

- Localizar en que parte de la imagen se encuentra el texto.
- La tipografía usada en los textos no siempre es la misma (incluso puede ser escritura manual).
- El espaciado entre los caracteres de una palabra no siempre es constante.
- La resolución de las imágenes puede variar.
- El tamaño del texto también es variable.
- El sistema de captación de la imagen puede introducir ruido.

Casi todos los algoritmos de OCR se basan en 5 etapas:

1. Localización del texto en la imagen.
2. Binarización.
3. Fragmentación o segmentación de la imagen.
4. Adelgazamiento de los componentes.
5. Comparación con patrones.

Para nuestros objetivos hemos integrado un sistema de reconocimiento OCR llamado Tesseract[4] que realiza los pasos 2 al 5. Tesseract es una herramienta libre que fue inicialmente desarrollada por Hewlett Packard pero actualmente está siendo desarrollado por Google y está considerado como uno de los OCR de mayor precisión disponibles. Tesseract puede procesar aproximadamente los 10 idiomas más importantes pero puede ser entrenado con diccionarios para funcionar con otros idiomas.

Para darle mayor facilidad a Tesseract y darle directamente la parte de la imagen donde debe encontrarse el texto, el paso 1, lo hemos desarrollado nosotros mismos. Para ello, debemos escribir el texto dentro de un recuadro con bordes negros y gruesos. Nuestro sistema encuentra el texto escrito escrito buscando en la imagen un recuadro de dichas características. Una vez se ha

[4] http://code.google.com/p/tesseract-ocr/

encontrado dicha zona, es necesario obtener una imagen en planta de dicho texto, para ello es necesario hacer una transformación en perspectiva.

5. El sistema de traducción automática

El éxito del sistema global de traducción robótica, que aquí presentamos, depende en gran medida de la calidad y la velocidad de la traducción realizada. Este sistema debe ser capaz de traducir textos completos, no siendo para nuestros propósitos válidos un traductor que únicamente traduzca frases de manera aislada.

Para conseguir tales propósitos nos hemos valido del traductor en-linea de google que mediante una petición HTTP, permite traducir textos entre los más de 50 idiomas que soporta. Es importante resaltar que no es un traductor perfecto, si bien para frases sencillas, la traducción suele ser la correcta y para frases un poco más complejas, si bien muchas veces la traducción no es exacta, si nos permite hacernos una idea del contenido del mensaje.

El uso de Google Translate está en constante evolución, por lo cual, es previsible una mejora continuada de las traducciones proporcionadas por este servicio, con la consiguiente mejora de la aplicación presentada, y en la actualidad no hemos encontrado ningún otro traductor libre/comerial que proporcionase mejores traducciones de frases en tiempo real. Si bien cabe resaltar que nos obliga a que el robot tenga permanentemente una conexión de datos a Internet.

6. DESARROLLO DE LA HABILIDAD ROBOT-TRADUCTOR

Hasta ahora hemos descrito los tres componentes principales del sistema que aquí presentamos de manera independiente, en este apartado vamos a describir cómo estas tres habilidades pueden funcionar conjuntamente y coordinadamente para construir una habilidad capaz de que el robot funcione como un agente traductor.

El usuario deberá mostrarle al robot el texto que desea traducir, para ello debe escribir el texto dentro de un recuadro negro grueso y situarlo a una distancia menos a dos metros de la boca del robot, donde se encuentra localizada la cámara web. Para las pruebas, las traducciones las hemos hecho de español a inglés. Es decir el texto se escribe en español y el robot es capaz de sintetizarlo en inglés, si bien cualquier otra configuración es posible. El texto puede ser escrito a mano, o es posible coger cualquier texto imprimido con anterioridad.

Cuando el usuario finalizada la escritura del texto dentro del recuadro negro, le debe indicar al robot que ya está disponible dicha información. Este mensaje se puede hacer mediante voz o simplemente tocando al robot. Una vez el robot sabe que dicha información está disponible, mediante el módulo de reconocimiento de caracteres es capaz de determinar el texto que ha sido introducido. Dicha información la coloca en memoria compartida y envía un evento dentro de la arquitectura de control, de esta manera cualquier habilidad puede trabajar con ese texto.

Una vez que la habilidad que controla la traducción recibe el evento generado por el módulo anterior, hace una petición HTTP a la API de google translate para traducir dicho texto al idioma oportuno, en este caso al inglés. Con el resultado de esta petición, escribe nuevamente en memoria compartida la traducción del texto y se genera un nuevo evento.

El sistema de síntesis de voz se encarga de sintetizar el texto traducido en el idioma oportuno, en este caso en inglés, para ello lee dicho texto de memoria compartida cuando recibe el evento de que el texto ya está traducido. Para la síntesis de voz se puede usar los diversos motores de síntesis que hemos integrado: Loquendo, Festival o Google TTS, ya que los tres disponen de voces en inglés claramente entendibles, si bien Loquendo es el único que nos permite expresarse con emociones: felicidad, tristeza, nerviosismo o tranquilidad.

7.CONCLUSIONES

Hemos presentado al robot personal y social Maggie como un traductor personal capaz de ayudar a traducir textos en cualquier idioma. Para ello basta con presentarle el texto a traducir dentro de un recuadro de bordes negros, tocarle en alguna parte de su cuerpo y casi al instante el robot es capaz de expresar verbalmente la traducción del contenido del texto en el idioma establecido.

Con esta utilidad cumplimos un doble objetivo, por un lado dotamos a Maggie de una interesante herramienta, que ya podemos ver en algunos teléfonos inteligentes, y por otro lado, la tarea de investigación y desarrollo ha derivado en que integremos dentro de nuestra arquitectura de control de tres importantes mecanismos que permiten aumentar las posibilidades de interacción de cualquier robot que use la arquitectura AD, como son: sistema de sistema complejo de síntesis de voz, un sistema de reconocimiento de caracteres y finalmente un sistema de traducción automática.

Agradecimientos

Los autores agradecen enormemente la financiación del gobierno de España a través del proyecto llamado “Peer to Peer Robot-Human Interaction" (R2H), del ministerio de Ciencia e Investigación (MEC), y al proyecto “A new approach to social robotics" (AROS), del MICINN (Ministerio de Ciencia e Innovación).

Referencias

Barber, R., and Ma Salichs. 2002. A new human based architecture for intelligent autonomous robots. In *Intelligent autonomous vehicles 2001 (IAV 2001): a proceedings volume from the 4th IFAC Symposium, Sapporo, Japan, 5-7 September 2001*, 81. Pergamon. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+new+human+based+architecture+for+intelligent+autonomous+robots#0.

Breazeal, Cynthia. 2003. Emotive qualities in lip-synchronized robot speech. *Advanced Robotics* 17, no. 2 (May): 97-113. doi:10.1163/156855303321165079. http://www.ingentaconnect.com/content/vsp/arb/2003/00000017/00000002/art00003.

Curtis, Meagan E, and Jamshed J Bharucha. 2010. The minor third communicates sadness in speech, mirroring its use in music. *Emotion (Washington, D.C.)* 10, no. 3 (June): 335-48. doi:10.1037/a0017928. http://www.ncbi.nlm.nih.gov/pubmed/20515223.

Johnstone, T. *Emotional speech elicited using computer games*. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP ’96*. IEEE. doi:10.1109/ICSLP.1996.608026. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=608026.

Juslin, Patrik N, and Petri Laukka. 2003. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological bulletin* 129, no. 5 (September): 770-814. doi:10.1037/0033-2909.129.5.770. http://www.ncbi.nlm.nih.gov/pubmed/12956543.

Roekhaut, Sophie, Jean-philippe Goldman, Anne Catherine Simon, Université De Mons Umons, Université De, Département De Linguistique, Université De Genève, and Institut Langage. A Model for Varying Speaking Style in TTS systems 4 . Implementing various speaking styles within the eLite TTS system, no. Table 1: 4-7.

Salichs, Miguel, Ramon Barber, Alaa Khamis, Maria Malfaz, Javier Gorostiza, Rakel Pacheco, Rafael Rivas, Ana Corrales, Elena Delgado, and David Garcia. 2006. *Maggie: A Robotic Platform for Human-Robot Social Interaction*. *2006 IEEE Conference on Robotics, Automation and Mechatronics*. IEEE, December. doi:10.1109/RAMECH.2006.252754. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4018870.

Scherer, K R. 1995. Expression of emotion in voice and music. *Journal of voice : official journal of the Voice Foundation* 9, no. 3 (September): 235-48. http://www.jvoice.org/article/S0892-1997(05)80231-0/abstract.

Scherer, K. R., R. Banse, and H. G. Wallbott. 2001. Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. *Journal of Cross-Cultural Psychology* 32, no. 1 (January): 76-92. doi:10.1177/0022022101032001009. http://jcc.sagepub.com/cgi/content/abstract/32/1/76.

Thompson, William Forde, E Glenn Schellenberg, and Gabriela Husain. 2004. Decoding speech prosody: do music lessons help? *Emotion (Washington, D.C.)* 4, no. 1 (March): 46-64. doi:10.1037/1528-3542.4.1.46. http://www.ncbi.nlm.nih.gov/pubmed/15053726.