Relada 5(3): 239-247, 2011
ISSN: 1988-5822

Utilización de las tecnologías del habla para facilitar el acceso a la web

David Griol. Víctor Corrales. José Manuel Molina.

Universidad Carlos III de Madrid. Departamento de Informática. Escuela Politécnica Superior.
Avda. de la Universidad, 30. 28911 Leganés. España.
dgriol@inf.uc3m.es, 100048294@alumnos.uc3m.es, molina@ia.uc3m.es

Resumen: En el diseño de aplicaciones para la web no siempre suele tenerse en cuenta el objetivo fundamental de maximizar la accesibilidad, especialmente para facilitar el acceso a personas con discapacidades. En este artículo presentamos un sistema de diálogo multimodal desarrollado como demostración del potencial que ofrece la tecnología XHTML+Voice para el acceso oral a la información en la red. El sistema integra diferentes módulos para el acceso a aplicaciones representativas de la red, centrándonos en esta comunicación en tres de ellas. Voice Dictionary permite realizar búsquedas de contenidos en la enciclopedia Wikipedia. Voice Pronunciations se ha diseñado para el aprendizaje de idiomas a través de divertidos juegos de palabras e imágenes. Voice Google es un completo interfaz multimodal al motor de búsqueda de Google. Todos los módulos de la aplicación permiten la interacción a través del teclado y ratón, o mediante la voz, mostrando también los resultados de forma multimodal.

Palabras clave: Sistemas de Diálogo. Multimodalidad. XHTML+Voice. Entornos Web. Interacción Oral.

Abstract: The main objective of maximizing accessibility is not always taken into account in the design of web applications, specifically to facilitate access for disabled people. In this paper we present a multimodal dialogue system to demonstrate the potential of the XHTML+Voice standard for oral access to the Internet. The system consists of several modules to access web information. This contribution is focused on three of them. Voice Dictionary allows users to access Wikipedia. Voice Pronunciations has been designed to facilitate language learning by means of games with words and images. Voice Google is a complete, fast and effective web search engine. All the modules of the web application can be accessed using the keyboard and mouse, or through the users' speech, also using multimodal ways to present the results.

Keywords: Dialogue Systems. Multimodality. XHTML+Voice. Web Interfaces. Oral Interaction.

INTRODUCCIÓN

El auge de las tecnologías de la información ha propiciado la posibilidad cada vez mayor de acceder a los datos desde cualquier lugar, en cualquier

momento y a una velocidad casi instantánea. Los avances tecnológicos han favorecido además la creación de dispositivos con tamaños cada vez más reducidos, capaces de ejecutar aplicaciones y acceder a los datos mediante conexiones inalámbricas, como por ejemplo, PDAs y teléfonos inteligentes, utilizados ampliamente hoy en día para el acceso a la web social.

Para facilitar el uso de estos dispositivos han surgido recientemente diversas tecnologías, entre ellas, los sistemas de diálogo multimodales (McTear, 2004; van Kuppevelt et al., 2005; Wahlster, 2006). Estos sistemas pueden definirse como programas informáticos diseñados para emular la capacidad de comunicación de un ser humano utilizando diversos modalidades de comunicación (como el habla, gestos, movimientos, mirada, etc.).

Con el desarrollo de este tipo de sistemas se persiguen tres objetivos fundamentales. En primer lugar, conseguir que mediante el habla, la comunicación con el sistema sea lo más natural posible y cercana a la comunicación persona-persona. En segundo lugar, permitir el uso de aplicaciones en entornos en los que no podrían ser utilizadas mediante los interfaces tradicionales como el teclado y ratón, por ejemplo, en un automóvil. Por último, facilitar el acceso a la web social a personas con discapacidades visuales o motoras, posibilitando su integración y la eliminación de las barreras de acceso a Internet (Beskow et al., 2009).

En este artículo describimos un sistema de diálogo multimodal, que hemos denominado Voice Applications (VA), desarrollado con el principal objetivo de servir como un marco de referencia para el estudio de la tecnología XHTML+Voice (X+V)[1] en la creación de sistemas multimodal que facilitan la accesibilidad a la información en la red. Este lenguaje combina las funcionalidades ofrecidas por XHTML[2] para la interacción utilizando modalidades visuales y las ofrecidas por el lenguaje VoiceXML[3] para la interacción mediante el habla.

El sistema VA integra diferentes aplicaciones, centrándonos en esta comunicación en describir tres de ellas. Voice Dictionary recibe del usuario el contenido que se desea buscar en la enciclopedia Wikipedia, recoge el resultado de la búsqueda, lo procesa y lo muestra/narra al usuario, permitiéndole también realizar una nueva búsqueda o seleccionar mediante la voz cualquiera de los enlaces que aparecen en el resultado. Voice Google realiza búsquedas en el navegador web, remitiendo la información al buscador, procesando la información resultante y mostrarla al usuario, facilitando del mismo modo el acceso multimodal a los enlaces obtenidos como resultado de la búsqueda. Finalmente, Voice Pronunciations incluye diferentes juegos diseñados para el aprendizaje de idiomas y basados en la pronunciación de palabras a través de la visualización de imágenes, definiciones y frases con palabras que el usuario debe completar.

[1] http://www.w3.org/TR/xhtml+voice
[2] http://www.w3.org/TR/xhtml1/
[3] http://www.w3.org/TR/voicexml20/

EL SISTEMA VOICE APPLICATIONS

Tal y como se ha descrito en la introducción, la principal finalidad del sistema Voice Applications es proporcionar una demostración de la tecnología X+V para aumentar la accesibilidad a aplicaciones representativas de Internet.

El sistema está implementado mediante un conjunto de documentos X+V. Algunos de ellos se encuentran almacenados desde el inicio en el servidor de documentos de la aplicación, mientras que otros se generan dinámicamente utilizando programación en PHP y tienen en cuenta las características y preferencias de los usuarios, p. ej. sexo (masculino/femenino) y lenguaje de interacción preferido (p. ej. inglés), así como información extraída de diversas bases de datos MySQL.

Para visualizar los documentos X+V que componen la aplicación, los usuarios deben disponer en su PC o dispositivos móviles de un navegador que soporte la interacción oral y la especificación X+V. En nuestra implementación hemos utilizado el navegador Opera[4], que permite además la navegación mediante la voz (como recargar una página, detener el acceso a Internet, volver hacia atrás, etc.).

A través del acceso a la página principal del entorno, mostrada en la Figura 1, encontramos un portal amigable que alberga las aplicaciones desarrolladas. El contenido de la página principal (a la cual los usuarios pueden acceder desde cualquier página del entorno pronunciando la palabra “Home”) consta de una sencilla presentación de cada una de los módulos del sistema. Cada módulo se muestra con su imagen representativa a la izquierda, y alineado a la derecha un texto que describe brevemente el contenido de la aplicación. Ambos elementos, texto e imagen, contienen el enlace correspondiente a la aplicación.

La totalidad de contenidos de la aplicación son leídos al usuario a través de la interfaz oral de la página, incluyendo la posibilidad de interrumpir este diálogo en cualquier momento, así como la utilización del resto de posibilidades de navegación multimodal ofrecidas por el entorno.

El desarrollo de interfaces orales implementados en X+V implica la definición de gramáticas, que establecen los límites de recepción de información del motor oral de la aplicación. Para poder cubrir el mayor rango de posibilidades de búsqueda en los diferentes módulos de la aplicación se estableció una estrategia que favoreciese esta la maximización. Esta estrategia se basa en aspectos como la generación automática de gramáticas una vez se dispone de los resultados generados por la aplicación, la utilización de gramáticas con frases completas para favorecer la naturalidad de la interacción con el sistema, y la utilización incluso el deletreado de las palabras en los casos en los que requiere no acotar la búsqueda o en las situaciones en las que se haya detectado fallos continuados de reconocimiento.

[4] http://www.opera.com/

Figura 1. Pantalla inicial del sistema Voice Applications.

Voice Dictionary

Tal y como se ha introducido previamente, mediante la aplicación Voice Dictionary (VD) accedemos a un entorno sencillo mediante el cual pueden realizarse búsquedas en la enciclopedia Wikipedia, con la particularidad de que en VD el contenido resultante de la búsqueda es narrado íntegramente al usuario. Además, esta búsqueda de contenidos se puede realizar tanto de manera tradicional (utilizando el teclado y el ratón) como mediante el uso de la voz.

Una vez que el contenido de la búsqueda inicial se muestra por pantalla y la interfaz oral comienza a narrarlo, el usuario puede perfectamente visitar cualquier otra aplicación interrumpiendo la narración, o acceder a cualquiera de los enlaces destacados en el texto. Esta funcionalidad se consigue mediante la generación dinámica de las gramáticas correspondientes, en las que se incorporan directamente los vínculos encontrados en el contenido resultante de la búsqueda.

Voice Pronunciation

La idea original de la aplicación Voice Pronunciation surgió a partir del objetivo fundamental de desarrollar una aplicación web con la que facilitar el aprendizaje y mejora de la pronunciación de idiomas, así como la adquisición de nuevo vocabulario. Una vez accedemos a la aplicación desde cualquier lugar del entorno VA mediante el comando “Pronunciation”, el usuario se encuentra con una bienvenida a través de discurso oral, y la posibilidad de elegir entre tres opciones: Words, Pictures e Instructions (Figura 2) Al pronunciar el comando “Instructions” accedemos a la sección dónde se describen todas las instrucciones necesarias para aprender el uso de la aplicación.

Figura 2. Aplicación Voice Pronunciation para el aprendizaje de idiomas.

Los dos menús inferiores de la aplicación VP, Words y Pictures, permiten el acceso a los juegos desarrollados, el primero de ellos para un uso más dirigido a la interfaz oral y el segundo más encaminado a la interfaz visual, aunque ambos juegos posibilitan la interacción completa con el usuario de forma multimodal. Words se basa en ir mostrando sucesivamente en pantalla una de las más de cien mil palabras almacenadas en la base de datos de la aplicación, conjuntamente con la definición de la misma. La interfaz oral sólo narra la descripción de la palabra y el usuario debe pronunciar correctamente la palabra mostrada. La aplicación Pictures, en lugar de palabras, muestra una

imagen elegida de entre las diversas dificultades especificadas, cuyo nombre exacto debe averiguar el usuario y pronunciarlo correctamente para continuar en el juego y aumentar su puntuación (Figura 3).

Figura 3. Aplicación Pictures en Voice Pronunciation.

Voice Google

Google es actualmente una de las empresas más importantes en el tratamiento de la información en Internet debido a su motor de búsqueda y multitud de aplicaciones y servicios ofrecidos en la red. Sin embargo, aunque está en implantación actualmente una aplicación oral a través de la cual poder consultar el correo de Gmail, no existe actualmente la posibilidad de utilizar el motor de búsqueda únicamente con la voz. De este modo, la aplicación Voice Google (VG) se ha desarrollado con el objetivo fundamental de posibilitar la búsqueda de información a través de Google utilizando la voz, tal y como muestra la Figura 4.

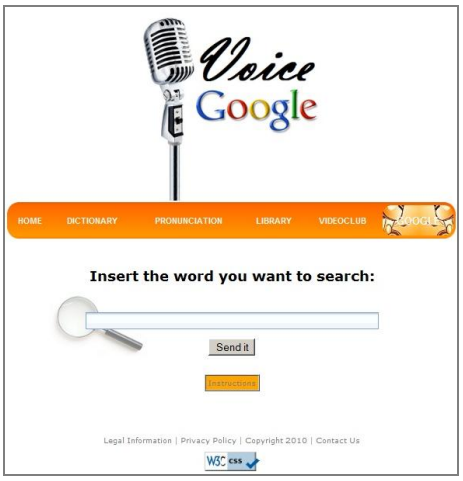

Figura 4. Aplicación Voice Google para la búsqueda de información.

La interfaz de la aplicación recibe el contenido facilitado por el usuario y muestra los resultados de la búsqueda tanto visual como oralmente. Además, la aplicación permite seleccionar también de forma multimodal cualquiera de los enlaces mostrados en la página de resultados (Figura 5).

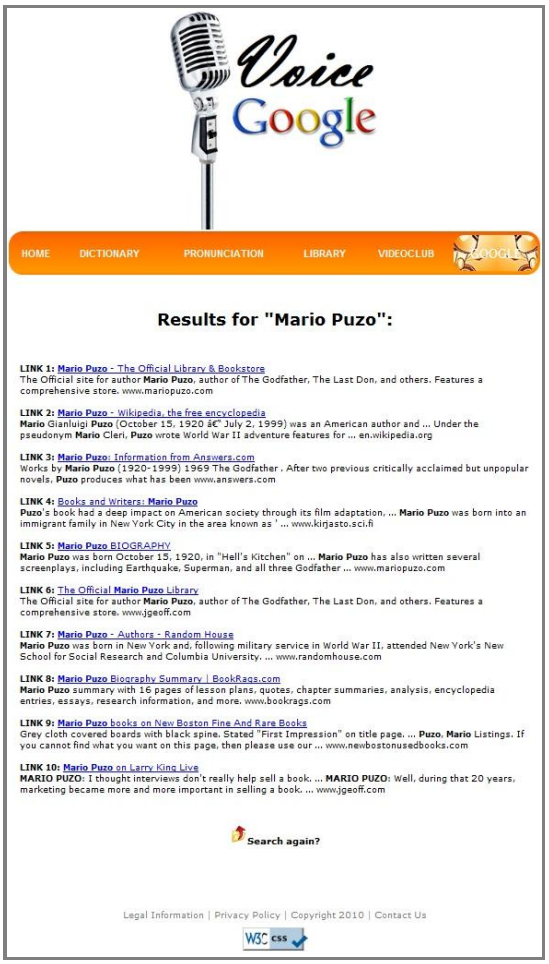

Figura 5. Resultado de una búsqueda utilizando Voice Google.

Evaluación preliminar

Se han llevado a cabo numerosas pruebas para maximizar el grado de funcionalidad de los diferentes módulos de la aplicación y posibilitar la detección y corrección del mayor número de errores. Una de las principales dificultades detectadas consiste en la generación de incongruencias cuando coinciden palabras de pronunciación similar a las palabras reservadas por el navegador o definidas para la propia aplicación. Se han limitado estas incoherencias al máximo de forma que las posibles concordancias entre las palabras seleccionadas se hayan eliminado.

Los resultados de la evaluación preliminar de Voice Applications con la participación de profesores y alumnos de nuestra universidad muestran que se valora muy positivamente la facilidad de obtener la información solicitada mediante la interacción con el sistema, así como el adecuado ritmo de interacción durante el diálogo y el funcionamiento individual de cada uno de los módulos del sistema. Entre los aspectos que se mencionan que deben mejorarse se incluye la corrección de los errores de sistema y una mejor clarificación de las acciones esperadas por el sistema en cada momento de la interacción.

CONCLUSIONES

El entorno Voice Applications se ha desarrollado como un marco de referencia para el estudio de las tecnología XHTML+Voice en el desarrollo de sistemas que posibiliten una mayor accesibilidad a la información en Internet. Sincronizados a través de eventos XML, los lenguajes XHTML y VoiceXML se encargan respectivamente de facilitar el diseño visual de la aplicación y permitir el diálogo hablado con el usuario. De esta forma, se integra la capacidad de interacción multimodal tanto para la entrada como salida del sistema. La utilización de lenguajes como PHP y JavaScript, así como de gestores de bases de datos como MySQL, facilita además la incorporación de las funcionalidades de adaptación y generación dinámica de los contenidos de la aplicación.

Las diferentes aplicaciones descritas posibilitan respectivamente la búsqueda de contenidos en la enciclopedia Wikipedia con nuestra propia voz y obtener los resultados de manera visual y oral; el aprendizaje de idiomas a partir de divertidos juegos que ayudan a la correcta pronunciación de las palabras; y la implementación de un interfaz oral a un motor de búsqueda en Internet.

En el diseño de este conjunto de aplicaciones se ha prestado una especial atención a potenciar la accesibilidad de la información y la facilidad de navegación en el propio sistema, a través de la incorporación de instrucciones detalladas, mensajes de ayuda y menús que facilitan la interacción entre los diferentes módulos y componentes de los mismos.

Entre las líneas de trabajo futuro cabe destacar como trabajos fundamentales la adaptación del sistema desarrollo para la interacción en castellano, el desarrollo de una evaluación más detallada de cada uno de los

módulos del sistema, así como la incorporación en cada uno de ellos de nuevas funcionalidades.

BIBLIOGRAFÍA

Beskow, J.; Edlund, J.; Granström, B.; Gustafson, J.; Skantze, G.; y Tobiasson, E. (2009). The MonAMI Reminder: a spoken dialogue system for face-to-face interaction. Proc. of InterSpeech'09, pp.296-299

McTear, M. F. (2004). Spoken Dialogue Technology: Towards the Conversational User Interface. Springer.

van Kuppevelt, J.; Dybkjaer, L. y Bernsen, N. O. (2005). Advances in natural multimodal dialogue systems. Springer.

Wahlster, W. (2006). SmartKom: Foundations of Multimodal Dialogue Systems. Springer.

Recibido: 11 marzo 2011.
Aceptado: 11 abril 2011.