ENTORNO PARA EL DESARROLLO DE APLICACIONES MULTIMEDIA CON SÍNTESIS Y RECONOCIMIENTO DE VOZ

Autores: R. San-Segundo, J.M. Montero, J. Colás, J. Ferreiros, R. Córdoba, A. Gallardo, J. Macías-Guarasa, J.M. Gutiérrez, J. Pastor, J.M. Pardo.

Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica.
Universidad Politécnica de Madrid

E.T.S.I. Telecomunicación. Ciudad Universitaria s/n, 28040 Madrid.

Telf: 91-5495762 Ext: 579
Fax: 91-3367323

ÁREA IV : Servicios, Aplicaciones y contenidos multimedia. Tecnologías del habla y de la imagen.

RESUMEN

INTRODUCCIÓN

A medida que las prestaciones de los ordenadores van aumentando, cada vez se desarrollan aplicaciones multimedia más sofisticadas donde se permite la combinación de imágenes, sonidos y videos en tiempo real haciéndolas más atractivas para el usuario. El auge en este tipo de aplicaciones, junto con la expansión de Internet, han sido los elementos más importantes que han impulsado la llegada del ordenador a la mayoría de los hogares.

En la última década, las Tecnologías del Habla se han ido consolidando, dando lugar a una nueva forma de interacción con el ordenador mucho más natural y cómoda. El teclado, el ratón o incluso las pantallas táctiles no pueden competir con la flexibilidad ofrecida por la voz. La incorporación de estas tecnologías en las aplicaciones multimedia supone un salto cualitativo en la flexibilidad de la interacción que se produce con el usuario. A parte del mundo de los video-juegos, la tecnología multimedia encuentra en la información interactiva (guías de museos, rutas turísticas, etc...) y en la educación, campos de aplicación muy importantes [1][2]. A pesar de ello, para permitir una completa integración de las tecnologías multimedia y del habla en las aulas, museos o en las oficinas de turismo, es necesario disponer de herramientas que permitan a profesionales sin grandes conocimientos técnicos, como profesores de educación primaria/secundaria, psicólogos, guías de museos, historiadores, etc... desarrollar aplicaciones flexibles cuyos contenidos se adapten a la temática a exponer, la ruta a mostrar o a las dificultades específicas de un conjunto de alumnos.

DESCRIPCIÓN

En esta comunicación se presenta un entorno de programación implementado por el Grupo de Tecnología del Habla de la ETSIT de Madrid, para el desarrollo de aplicaciones multimedia con síntesis y reconocimiento de voz. Este entorno de desarrollo dispone de un lenguaje de alto nivel que permite el diseño rápido de aplicaciones sin necesidad de conocer aspectos técnicos de multimedia o de tecnología del Habla. De forma general, una aplicación multimedia especificada en nuestro lenguaje consta de las siguientes partes:

- *Inicialización de variables*: donde se configuran las variables generales del sistema (como por ejemplo el nombre del directorio con los ficheros multimedia a utilizar en la aplicación).
- *Tratamiento de errores estándar*: donde se indican las acciones a realizar por el sistema cuando se produzcan algunos de los errores considerados en la aplicación (máximo tiempo excedido sin que el usuario conteste a una pregunta de la aplicación).
- *Subrutinas*: donde se definen e implementan las subrutinas que se utilizarán en la aplicación.
- *Aplicación*: que constituye el conjunto de instrucciones que se irán ejecutando de forma secuencial. En este lenguaje se dispone de los siguientes tipos de sentencias o instrucciones:
    - *Gestión de Agentes Animados:* mediante un conjunto de instrucciones se ofrece el acceso a la librería de agentes animados que ofrece Microsoft® Agent, permitiendo realizar diversos movimientos, posicionar al agente en cualquier lugar de la pantalla y hacerle aparecer o desaparecer. La incorporación de agentes animados en las aplicaciones multimedia es muy importante porque ofrece la posibilidad de introducir humanidad emocional en la comunicación con el usuario y además permite dar una dimensión personal a la interacción con el ordenador.
    - *Gestión de Elementos Multimedia:* se dispone instrucciones para comenzar o parar la reproducción de videos, para carga y descarga de imágenes, y sentencias para la grabación y reproducción de ficheros de voz en formato WAVE.
    - *Síntesis y Reconocimiento de voz:* en este lenguaje se ofrece además primitivas de alto nivel para la síntesis de cualquier texto en español y primitivas para el reconocimiento de palabras aisladas y/o expresiones. La tecnología utilizada para implementar estas funciones ha sido desarrollada íntegramente por nuestro grupo de investigación [3][4]. En el caso del reconocimiento se disponen de

tasas superiores al 90% con vocabularios de 1000 palabras y/o expresiones, lo que ofrece una cobertura y fiabilidad suficiente para el desarrollo de este tipo de aplicaciones.

La mayoría de estas instrucciones funcionan de modo asíncrono, de forma que se pueden simultanear varias de ellas. Este hecho permite, por ejemplo, realizar un movimiento con el agente animado a la vez que se reproduce un video y se sintetiza un párrafo explicativo sobre las imágenes observadas. Un ejemplo se puede observar en la figura siguiente.

Figura 1: Instante de la ejecución de una aplicación multimedia sobre la historia del fútbol.

- *Funciones Adicionales:* además de las instrucciones comentadas anteriormente se ofrecen funciones para la realización de operaciones aritméticas, manejo de cadenas, gestión de directorios y archivos (copia, borrado, cambio de nombre,...), acceso a bases de datos y envío de correo electrónico.
- *Funciones de Gestión de línea telefónica:* en este entorno de programación se permite además el diseño de aplicaciones telefónicas con síntesis y reconocimiento de voz. Para ofrecer esta funcionalidad adicional se dispone de un módulo para la gestión de la línea telefónica. Este módulo ofrece al usuario las funciones de colgar/descolgar la línea telefónica, esperar una llamada y marcar. Además de estas instrucciones comentadas, el sistema ofrece análisis del progreso de llamada (*CPA: Call Progress Analysis*) configurable, redirección de llamadas dentro de una misma centralita (Alcatel o Ibercom) y activación de un hilo musical en caso de accesos largos a bases de datos.

En la comunicación se describirá más extensamente este lenguaje, se presentarán ejemplos de aplicaciones desarrolladas y se detallará la implementación de las instrucciones de alto nivel.

Además de este lenguaje, el entorno de desarrollo dispone de las herramientas necesarias para cubrir el ciclo de vida de una aplicación: diseño, compilación, depuración y ejecución:

- Editor de texto con todas las utilidades de edición para la escritura de las aplicaciones multimedia (genera ficheros *.pro).
- Un compilador de aplicaciones que permite la detección de errores de escritura y ayuda al desarrollador en el aprendizaje de la sintaxis del lenguaje. Una vez corregidos los posibles errores se traduce el fichero de texto a un fichero binario (*.est) con la especificación del autómata de estados de la aplicación multimedia desarrollada.
- El ejecutor de aplicaciones permite la carga de ficheros binarios (*.est), los interpreta y ejecuta la aplicación correspondiente.

- El entorno permite también la ejecución paso a paso de la aplicación con el fin de facilitar su depuración. En este modo de ejecución se permite la introducción de puntos de ruptura y se posibilita la visualización y modificación de los valores de la variables definidas en la aplicación.
- Se ofrecen herramientas para la grabación de ficheros de voz por parte del desarrollador, de forma que puedan ser utilizadas en la aplicación multimedia.
- El entorno dispone también de opciones para la generación de diccionarios de reconocimiento con las palabras y/o expresiones que se utilizarán en la interacción con la aplicación a desarrollar.

CONCLUSIÓN

En esta comunicación se presenta un entorno para el desarrollo de aplicaciones multimedia con la posibilidad de incorporar la funcionalidad de síntesis y reconocimiento de voz. A través de este entorno se acerca el diseño de aplicaciones multimedia a colectivos sin gran nivel de conocimientos técnicos en programación como profesores de educación primaria, psicólogos, guías turísticos, etc... para los cuales estas tecnologías le son de gran ayuda y que, poco a poco, se están consolidando como herramientas de trabajo imprescindibles. En esta comunicación se propone la alianza de dos tecnologías, multimedia y del habla para dar lugar a una nueva generación de interfaces de usuario y aplicaciones interactivas.

REFERENCIAS

- [1] http://www.irabia.org/
- [2] Sutton, S. Cole, R.A. de Villiers, J. Schalkwyk, J. Vermeulen, P., Macon, M., Yan, Y. Kaiser, E. Rundle, B., Shobaki, K., Hosom, J.P., Kain, A., Wouters, J., Massaro, M. And Cohen, M. "Universal Speech Tools: the CSLU Toolkit". Proceeding of the International Conference on Spoken Language Processing page 3221-3224, Sydney Australia, November 1998.
- [3]. J.M. Pardo et al “Spanish text to speech: from prosody to acoustic” International Conference on Acoustic 95 vol III, 1995
- [4]. J.Macías-Guarasa, A. Gallardo, J. Ferreiros. “Módulo de búsqueda rápida para el reconocedor de habla aislada y grandes vocabularios”. Informe interno. Grupo de Tecnología del Habla. DIE. UPM. Madrid, 1996.
- [5]. J. Macías-Guarasa, M.A. Leandro, J. Colás, A. Villegas, S. Aguilera, J.M. Pardo. “On the Development of a Dictation Machine for Spanish: DIVO”. ICSLP’94, S22-26, pp. 1343-1346. 1994