La evolución de la Síntesis del Habla en Ingeniería La Salle

Francesc Alías[*]*, Ignasi Iriondo*

Departamento de Comunicaciones y Teoria de la Señal
Ingeniería y Arquitectura La Salle. Universitat Ramon Llull
{falias, iriondo}@salleURL.edu

Resumen

Este artículo resume la trayectoria del grupo de Tecnologías del Habla de Ingeniería La Salle (Universitat Ramon Llull) en el marco de la investigación y el desarrollo de sistemas de síntesis del habla. Partiendo del trabajo realizado en las últimas dos décadas, se presentan las líneas de investigación que se están desarrollando en la actualidad y se definen los objetivos planteados para un futuro próximo. La idea fundamental es conseguir un sistema de síntesis multimodal que haga más agradable el flujo de información desde el ordenador hacia el usuario. La materialización de estos objetivos se pretende llevar a cabo mediante el diseño y el desarrollo de un locutor virtual realista conjuntamente con el grupo de Visión por Computador de nuestro centro.

1. Antecedentes

Ingeniería La Salle ha sido uno de los centros pioneros en el campo de la síntesis del habla desde la década de los ochenta con los trabajos de Josep Martí [1] [2] [3]. Se llevaron a cabo trabajos de investigación y desarrollo en síntesis articulatoria, síntesis por formantes y síntesis basada en predicción lineal. Estos sistemas se aplicaron principalmente a productos orientados a personas invidentes. En concreto, la colaboración mantenida con la empresa CIBERVEU, S.A. del grupo ONCE se materializó en un conjunto de equipos sintetizadores de voz, capaces de leer la información proveniente de la pantalla y el teclado de un ordenador. Estos equipos tenían que ser externos al ordenador personal, debido a la poca capacidad computacional y de memoria que éste presentaba en aquel entonces. El producto más destacable de esta gama es el CIBER232P [4], un equipo portátil que incorpora el circuito integrado PCF8200 como módulo de síntesis. La comunicación con el ordenador se realiza vía serie y el equipo funciona como conversor texto-habla (CTH) para el catalán y el castellano.

La calidad de los sistemas de síntesis del habla hasta ese momento sólo estaba asociada a su grado de inteligibilidad, factor muy apreciado en un CTH orientado a personas invidentes. El siguiente reto fue incrementar su naturalidad, por lo que se centró el esfuerzo en el avance de estas tres áreas: 1) el análisis lingüístico del texto, 2) el modelado y la automatización de la prosodia, y 3) el procesado en la síntesis de habla. Cualquier deficiencia en una de estas tres áreas resulta crítica en el resultado final de la síntesis.

Seguidamente, el progresivo avance tecnológico conllevó una mayor capacidad de proceso y memoria en los ordenadores personales, por lo que se inició el desarrollo de sistemas CTH basados únicamente en *software*. Con el objetivo de mejorar el bloque de síntesis, se optó por la síntesis concatenativa basada en difonemas y trifonemas. Se implementó un sintetizador en catalán [5] [6], basado en la reciente publicación de la técnica PSOLA por parte de Moulines y Charpentier [7], que sirvió de base para los sistemas desarrollados posteriormente.

Después de este breve resumen sobre los primeros avances en la síntesis del habla llevados a cabo por nuestro grupo, el artículo describe los progresos en este ámbito obtenidos recientemente. El apartado 2 presenta el sistema CTH basado en síntesis por concatenación. Seguidamente, el apartado 3 describe los objetivos y las líneas de trabajo actuales, centradas en la síntesis multimodal del habla. A continuación, se exponen algunos de los proyectos llevados a cabo hasta el momento (apartado 4) y, por último, las conclusiones y líneas de futuro (apartado 5).

2. Síntesis por concatenación

Durante el año 1997, se inició el desarrollo de un sistema de conversión texto-habla en catalán de alta inteligibilidad, a partir de la experiencia acumulada y gracias a un proyecto financiado por Televisió de Catalunya, S.A. (TVC). Se fijó el objetivo de mejorar los diferentes módulos que componían el CTH existente. En primer lugar, se mejoró el preprocesado del texto. Seguidamente se desarrolló un lenguaje de reglas que permitiese programar la conversión grafema a fonema y generar automáticamente la prosodia. En cuanto al módulo de síntesis se realizaron mejoras en la concatenación de las unidades y se grabó, segmentó y etiquetó un nuevo corpus de difonemas y trifonemas para el catalán (1207 unidades). Con este proyecto se consiguió un CTH con muy buena inteligibilidad y una calidad aceptable para su integración en muchas aplicaciones. Sin embargo, dicha calidad no es suficiente para aplicaciones que requieran calidad *broadcast*, como por ejemplo el doblaje de documentales para televisión. A continuación se detallan los módulos que se obtuvieron mediante este proyecto.

2.1. Editor de Mensajes Orales con Voz Sintética: EMOVS

Se diseñó e implementó una aplicación en entorno Windows para la síntesis automática de voz a partir de texto (ver figura 1). El objetivo inicial fue el diseño de una herramienta que permitiese la modificación manual de la prosodia, para su ajuste previo a la utilización de la voz sintética en un determinado medio. Los parámetros editables son la energía, la frecuencia fundamental, la duración de fonemas y pausas y la trascripción fonética. Se permite la modificación de los parámetros prosódicos para una frase, para un fonema o para un conjunto de fonemas consecutivos. Esta herramienta ha sido utilizada tanto para la mejora manual de locuciones sintéticas, como para la extracción y el modelado de patrones prosódicos.

[*]Este trabajo se ha realizado con el apoyo del DURSI de la Generalitat de Catalunya, mediante la beca número 2000FI-00679.

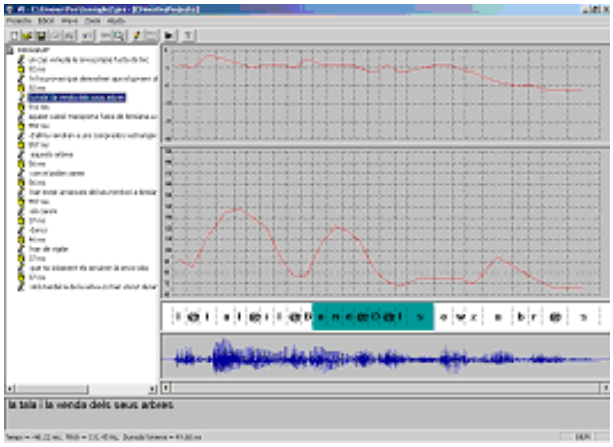

Figura 1: *Ejemplo del EMOVS. Se observa el texto organizado en frases y la información fonética y prosódica asociada a la frase seleccionada*

2.2. Lenguaje de reglas: SINCAT/2

La automatización del proceso de generación de la prosodia a partir del texto a sintetizar no es una tarea sencilla. De hecho, se han presentado en los últimos años diferentes aproximaciones para solucionar este problema, sin que ninguna de ellas se pueda considerar definitiva. Nuestra solución se basó en la generación e implementación de un lenguaje para la interpretación de reglas, que permitiese la conversión de un texto en su correspondiente trascripción fonética y la asignación a cada unidad de su prosodia (entonación, ritmo y energía). Estas reglas no dependían del propio código del programa, sino que se diseñaron para que fueran sencillas, claras y fácilmente modificables por el usuario.

Las características principales que debía cumplir el lenguaje fueron:

- Reglas estructuradas en conjuntos o módulos de reglas, de forma que puedan estructurarse según su cometido. Por ejemplo, reglas fonéticas, reglas de duración, reglas de energía, etc.
- Ficheros de reglas en modo texto, para facilitar su edición.
- Las reglas se compilarán previamente a su utilización para garantizar que la sintaxis sea correcta.
- Las reglas consultan y actualizan una estructura de datos que representa la secuencia de fonemas y sus propiedades.

Una vez implementado el lenguaje se crearon los módulos de reglas para llevar a cabo la trascripción fonética y la asignación de los parámetros prosódicos de la energía y la duración de las unidades. Además se desarrolló un analizador morfológico simple con el fin de generar patrones entonativos muy básicos. Sin embargo, se concluyó que la generación automática de la curva de entonación requería de un proceso más complejo a partir de un análisis más profundo del texto (morfológico, sintáctico y semántico).

2.3. Mejoras en la calidad segmental

Con la finalidad de mejorar la calidad segmental de la síntesis del habla por concatenación de unidades, se abordó la solución a los problemas debidos a la unión entre difonemas y a la modificación de su duración.

El hecho de trabajar con síntesis del habla por concatenación de unidades comporta que la evolución temporal del espectro de la señal generada sufra discontinuidades importantes en el centro de cada alófono. Estas discontinuidades provocan una degradación importante en la calidad del habla sintética. Por esta razón se estudiaron distintas alternativas para disminuir este efecto y mejorar la calidad del sistema. Se aplicaron técnicas de interpolación entre tramas y se modificó el punto de concatenación de las unidades, permitiendo empalmes en cualquier instante del fonema [8]. Además estudió, por otra parte, la posibilidad de utilizar selección de unidades mayores, con la finalidad de disminuir el número de errores generados [9].

Para modificar la duración de los fonemas, hasta entonces se utilizaban las técnicas convencionales (repetición de señal en la parte estable del alófono, repetición de señal intercaladamente, etc.) lo que generaba una señal poco natural cuando las modificaciones de duración eran importantes. La solución que se propuso se basaba en un sistema de transformación y combinación de tramas de segmentos adyacentes proporcionando una calidad de voz superior [8].

3. Síntesis multimodal

La síntesis multimodal, o audiovisual, se puede definir como la generación automática de una señal de voz conjuntamente con su correspondiente información visual. Tiende a emular la comunicación entre personas, la cual es básicamente visual y hablada. Por lo tanto, se utilizará fundamentalmente en interfaces hombre-máquina, haciendo las funciones de emisor del mensaje audiovisual generado por el sistema. La síntesis multimodal supone un valor añadido para la aplicación que la utilice, ya que permite mejorar tanto su accesibilidad como el grado de comprensión del mensaje emitido. Cabe destacar que la correcta sincronización entre la información sonora y visual es crucial para conseguir las mejoras que aporta este tipo de síntesis.

3.1. Objetivos planteados

En la actualidad, la interacción hombre-máquina está en constante evolución tecnológica y tiende hacia la consecución de sistemas multimodales y multilingües. Se plantea el objetivo de mejorar la componente oral del mensaje multimodal, en lo que se refiere a la producción automática del habla. Concretamente, se pretende desarrollar un nuevo sistema de síntesis de voz que permita mejorar las prestaciones obtenidas con el sistema concatenativo basado en difonemas (ver apartado 2). Estas mejoras se deberán centrar en:

- *Incremento de naturalidad*. La señal de voz sintetizada debe asemejarse a la real para conseguir la máxima verosimilitud posible. Será primordial mejorar la calidad del habla sintética, partiendo de la alta inteligibilidad obtenida por el conversor texto-voz basado en difonemas, hasta conseguir una mayor naturalidad siguiendo la filosofía de los sistemas de síntesis más recientes [10].
- *Multimodalidad*. La captación y generación de información audiovisual será la base comunicativa de la mayoría de los sistemas y aplicativos utilizados en un futuro no muy lejano. El progresivo incremento en la autonomía y capacidad computacional de los ordenadores personales, así como la aparición de nuevas técnicas de procesado digital del habla y de la imagen, permiten plantearse esta

posibilidad [11]. Por lo tanto, el sistema CTH deberá generar información de sincronismo para adaptar la imagen al habla, por ejemplo, debe notificar el fonema que se está emitiendo para que el módulo de procesado de imagen lo articule correctamente.

- *Multilingualidad.* Dentro del desarrollo de los futuros sistemas de síntesis de voz, una de las tendencias que parece más interesante es la síntesis de textos multilingües [12]. De este modo el sistema permite afrontar la síntesis de distintos idiomas. Referente a este aspecto, se plantea la mejora del sistema CTH existente para el catalán y el castellano.

3.2. Líneas de trabajo

Para conseguir los objetivos planteados en esta nueva etapa, resulta necesario el desarrollo de nuevos módulos dentro del proceso de conversión texto-habla y, a su vez, la mejora de algunos de los módulos ya existentes.

3.2.1. Bloque de Procesado del Lenguaje Natural:PLN

Este bloque es el encargado de analizar el texto de entrada del sistema para obtener su trascripción fonética y sus características lingüísticas y prosódicas. Actualmente este proceso se realiza mediante un lenguaje de reglas (ver apartado 2.2) junto a un módulo de preprocesado del texto, necesario para normalizar las abreviaciones, números, acrónimos... que puedan aparecer en el texto.

Generalmente, el bloque de PLN está constituido por los siguientes módulos: preprocesador del texto, analizador morfosintáctico, fonetizador y generador prosódico. En una primera fase, se pretende mejorar el bloque de PLN actual en lo que se refiere al análisis morfosintáctico, el cual se divide en dos procesos. Primero, el analizador morfológico examina cada una de les palabras del texto y les asigna todas sus posibles interpretaciones morfológicas y su lema. Éste debe estar acompañado por un desambiguador que seleccione la interpretación correcta de cada forma según el contexto sintáctico en el que aparece. A continuación, el analizador sintáctico estructura el texto en sintagmas simples organizados jerárquicamente, a partir de la gramática asociada al idioma tratado.

Ambos módulos se han obtenido de la colaboración con el Centro de Lingüística Computacional (CLiC) de la Universidad de Barcelona (UB). En la actualidad se ha iniciado el proceso de integración de estos módulos en nuestro sistema. Se espera obtener mejoras importantes en lo referente al análisis del texto y a la generación automática de la prosodia (energía, duración y frecuencia fundamental) de las unidades a sintetizar.

3.2.2. Módulo de Selección de Unidades

En los últimos años, los sistemas de síntesis concatenativa han pasado de utilizar una única realización por unidad a disponer de grandes corpus de voz donde existen múltiples realizaciones por unidad [13, 14]. Estas técnicas mejoran las limitaciones de síntesis de los sistemas basados en difonemas, ya que disminuyen el número de puntos de concatenación no naturales y reducen las modificaciones prosódicas a realizar durante el proceso de síntesis.

Estos sistemas necesitan de un módulo que sea capaz de seleccionar las mejores realizaciones para obtener la secuencia de sonidos deseada (objetivo). Este proceso lo lleva a cabo el módulo de selección de unidades, basado en programación dinámica (algoritmo de Viterbi), que escoge el conjunto de unidades que minimiza una determinada función de coste. Esta función toma en consideración tanto la similitud entre la unidad candidata y la objetivo (coste de unidad) como el grado de continuidad entre unidades consecutivas (coste de concatenación) [15]. En la actualidad existe todavía mucho trabajo a realizar en la selección de los parámetros a considerar, el modo de calcular los subcostes de la función y el ajuste de los pesos que los ponderan [16]. También resulta esencial un buen diseño y dimensionado del corpus de voz [9]. Ambos conceptos son objeto de estudio e investigación en nuestro grupo con el objetivo de desarrollar un nuevo sintetizador basado en corpus.

3.3. Corpus de voz

En el contexto de los sistemas de síntesis de voz basados en selección de unidades, la mejora en la calidad de la señal generada requiere el uso de un corpus de voz de grandes dimensiones (mínimo 50.000 unidades) comparado con el sistema basado en difonemas y trifonemas (1.207 unidades). Se pretende diseñar y grabar un nuevo corpus de voz para la síntesis con selección de unidades, contando con la colaboración del Departamento de Comunicación Audiovisual y Publicidad (CAP) de la Universidad Autónoma de Barcelona (UAB).

Por otro lado, el hecho de trabajar con grandes corpus de voz y la necesidad de conocer las características prosódicas de las unidades que las componen, hacen necesario la utilización de técnicas automáticas de etiquetado. Con esta finalidad, se han desarrollado las siguientes herramientas informáticas:

- *Segmentador automático.* Este módulo es el encargado de determinar la posición de las unidades que forman cada una de las frases del corpus de voz. De este modo se determina la posición temporal de cada unidad para su posterior recuperación en el proceso de síntesis y para la extracción de características prosódicas, como por ejemplo, su duración. Este programa utiliza un algoritmo basado en los modelos ocultos de Markov y se ha desarrollado a partir del *Hidden Markov Models ToolKit* (HTK) [17]. Después del proceso de entrenamiento de los modelos con un conjunto de frases etiquetadas y revisadas manualmente, el sistema es capaz de segmentar todo el corpus con un alto grado de acierto. Esta automatización, junto con la corrección experta de errores, consigue acelerar notablemente la segmentación de un gran corpus de voz respecto al proceso manual equivalente.
- *Marcador de* pitch. Las marcas de *pitch* se sitúan siguiendo el período fundamental (T_0) de la señal de voz en el dominio temporal. Son necesarias en las aplicaciones *pitch*-síncronas de procesado de la señal de voz y en la extracción de características prosódicas, como por ejemplo la frecuencia fundamental del extremo de la unidad. El método implementado está basado en dos conceptos: la energía de la señal y la ubicación óptima de las marcas mediante programación dinámica [18]. Este método trabaja directamente sobre la señal de voz, aunque puede ser interesante probar su funcionamiento sobre la señal electroglotal (EGG) extraída durante el proceso de grabación del nuevo corpus oral.
- *Interfície de Tractament de la Parla (ITP).* De la necesidad de integrar todos los algoritmos y cálculos relacionados con el análisis de un corpus de voz para síntesis, surge ITP o interfaz de tratamiento del habla. Esta herramienta se ha diseñado partiendo de una aplicación inicial desarrollada en MATLAB, que se aplicó al eti-

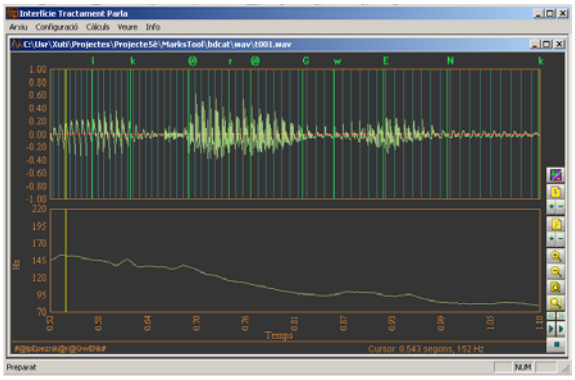

Figura 2: *Ejemplo de funcionamiento de ITP. Se presentan distintas informaciones: las marcas de segmentación, las de pitch, la trascripción fonética y la curva de pitch.*

Figura 3: *Apariencia del Locutor Virtual para diferentes expresiones y.*

quetado manual de un corpus de voz. Posteriormente, se le añadieron distintos módulos para la extracción de datos prosódicos de la señal, como pueden ser la duración de las unidades, su frecuencia fundamental media, su energía, etc. ITP hereda esta filosofía e incorpora los módulos de etiquetado automático que se acaban de describir, permitiendo también la edición manual de las marcas, si ésta resulta necesaria. Es una herramienta en Visual C++ en continuo desarrollo, cuya pantalla principal se presenta en la figura 2. Actualmente, permite consultar y editar la trascripción fonética de la frase, sus marcas de segmentación y de *pitch*, obtener la curva de *pitch* y el sonograma de la señal. A corto plazo está previsto añadir a ITP los módulos de extracción de información prosódica de las unidades.

3.3.1. Síntesis del habla expresiva

Aunque la mayor parte de los sistemas actuales de síntesis del habla se caracterizan por una buena inteligibilidad, sólo algunos de ellos consiguen cierta naturalidad. Principalmente, esta tendencia se debe a la mejora en la modelado de las características suprasegmentales del habla (prosodia), y al incremento en la calidad segmental (ver 3.2.2). Sin embargo, no hay resultados concluyentes en lo que se refiere al habla expresiva, es decir, a la transmisión de un estado emocional a partir de la voz sintetizada. Consideramos que esta es una línea de investigación muy interesante y de aplicación inmediata a la síntesis multimodal del habla. Gracias a la colaboración con el CAP de la UAB, se inició una línea de investigación en el campo del modelado y generación automática del habla expresiva o emocionada. Se partió de un modelo acústico de la expresión emocional [19] para llevar a cabo una validación de dicho modelo mediante síntesis del habla [20]. Se concluyó que la síntesis de ciertas emociones requiere de un módulo de procesado digital de la señal que permita grandes variaciones prosódicas. La síntesis utilizada hasta el momento no tenía suficiente versatilidad para conseguir resultados satisfactorios. Por este motivo se decidió utilizar un modelo híbrido entre PSOLA [7] y los modelos armónicos-estocásticos [21] que permitiese variaciones prosódicas de mayor calidad [22]. Actualmente se está implementando un módulo de síntesis del habla basado en dicho modelo híbrido que en un futuro se pretende incorporar a un sistema basado en selección de unidades.

4. Proyectos y aplicaciones

Para llevar a cabo los objetivos planteados en nuestro departamento, se iniciaron una serie de líneas de investigación en el ámbito del procesado de la señal de voz y en el del tratamiento de la imagen. Éstas han permitido la integración de los bloques de síntesis de voz y de imagen en distintas aplicaciones multimedia. Además, se han establecido distintas colaboraciones con otros grupos de investigación para cubrir algunas necesidades específicas y que son complementarias al trabajo que se está desarrollando dentro del grupo.

4.1. Locutor Virtual

A partir de los primeros resultados obtenidos en el campo de la síntesis multimodal, se realizó una primera toma de contacto con el sector informático para comprobar el interés que podían suscitar estos avances. Las empresas consultadas respondieron satisfactoriamente, por lo que se decidió solicitar una ayuda dentro del marco del *Plan de Investigación Científica, Desarrollo e Innovación Tecnológica de la Investigación Técnica (PROFIT)*. El proyecto FIT-150500-2002-410, titulado ”Locutor Virtual”, ha permitido obtener la financiación necesaria para llevar a cabo las líneas de trabajo que nos habíamos propuesto inicialmente (ver apartado 3.2). En concreto, la finalidad del proyecto ha sido el desarrollo de un prototipo correspondiente a una interfaz hombre-máquina multimedia, basada en un locutor de apariencia real. Permitirá reproducir texto de forma automática, sincronizando la voz con la gesticulación facial [23].

A continuación se describe, a grandes rasgos, la filosofía del módulo de síntesis integrado en el Locutor Virtual, junto a su interacción con el módulo de imagen.

4.1.1. Procesado digital de la imagen

La componente visual del locutor virtual parte de un modelo facial 2D parametrizado basado en imágenes reales y que, a su vez, es personalizable [24]. Es un modelo que permite la incorporación de expresiones faciales y de ciertos movimientos de la cabeza, características que consiguen mejorar la naturalidad del locutor. El proceso de entrenamiento del modelo se fundamenta en técnicas de seguimiento robusto, las cuales posibilitan la personalización específica del locutor de forma bastante sencilla.

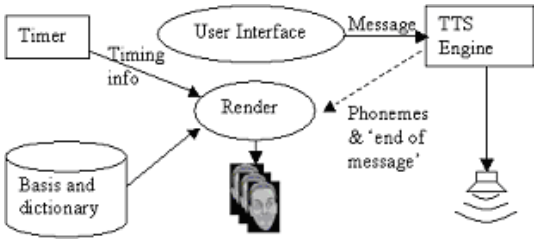

Figura 4: *Diagrama de bloques de la interacción de los módulos que componen el Locutor Virtual.*

4.1.2. Arquitectura e interacción de los bloques

El Locutor Virtual se ha diseñado sobre la interficie SAPI (*Speech Application Interface*) de Microsoft [25]. De este modo será sencillo incorporar nuevas voces y nuevos idiomas a esta aplicación multimodal. La interacción entre los distintos módulos que componen el locutor se presenta en la figura 4. Tal y como se describe en ella, el sistema de procesado de la imagen recibe notificaciones del sistema de síntesis del habla con información sobre qué fonema sonará y exactamente cuándo lo hará, cosa que permite la obtención de pronunciaciones sincronizadas.

4.2. SAPI 4.0 y SAPI 5.1

Speech API (SAPI) de Microsoft se creó con el objetivo de ser un marco de normalización o estandarización de la comunicación entre las aplicaciones de CTH y los motores de voz que éstas necesitan. Permite generalizar la integración de los algoritmos desarrollados por grupos de investigación, empresas tecnológicas, etc. en aplicaciones multimedia que requieran el uso de reconocimiento o síntesis de voz. A partir de los resultados obtenidos con el CTH incorporado en EMOVS, Televisió de Catalunya S.A. financió un proyecto para la integración del CTH basado en difonemas al entorno SAPI 4.0, implementando la mayor parte de sus interfaces para síntesis.

A partir de la experiencia obtenida con este proyecto y habiendo utilizado el motor de voz *SAPI-compliant* en la primera versión del Locutor Virtual, se decidió continuar trabajando en esta línea y aprovechar las ventajas planteadas por la nueva versión de la API, SAPI 5.1. En la actualidad, prácticamente se ha finalizado la migración del código aprovechando el mejor diseño y arquitectura de esta nueva versión de la interfaz. En breve estará disponible para su integración en la nueva versión del Locutor Virtual.

4.3. Utilización *on-line* en la Web

Debido al creciente ancho de banda disponible en Internet, la gran mayoría de sus usuarios tiene la posibilidad de visualizar en su ordenador personal vídeo en *streaming*. Actualmente, se está desarrollando una aplicación interactiva para su funcionamiento en la Web, de forma que el usuario introduzca el texto y las propiedades que desee (ritmo, entonación y estado emocional) para que sea reproducido por el Locutor Virtual. Esta aplicación se podría acoplar a un sistema de diálogo para contestar preguntas de los usuarios en un determinado contexto o como ayuda a la navegación en una Web.

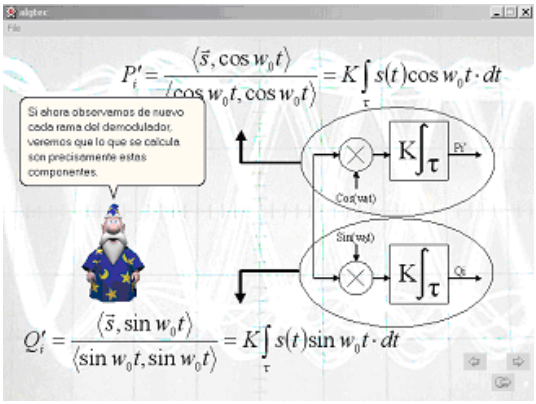

Figura 5: *Lección de la aplicación ALGTEC.*

4.4. Aplicaciones educativas: ALGTEC

ALGTEC (ÁLGebra y TECnología) [26] es una aplicación multimedia que nace con el objetivo de hacer más atractivas las explicaciones en la asignatura de Álgebra Lineal, aprovechando las posibilidades tecnológicas existentes en la actualidad. Pretende complementar las lecciones impartidas por el profesor y motivar al alumno en su aprendizaje de los conceptos algebraicos, haciendo referencia a distintas aplicaciones de éstos en el mundo tecnológico.

En la aplicación existe un 'profesor virtual' que es el encargado de explicar la aplicación técnica del concepto estudiado. En un diseño inicial, se utilizaron unos personajes denominados *agents* de Microsoft [27] que trabajan con motores de voz *SAPI-compliant* (ver figura 5). En un nuevo diseño se ha decidido integrar el Locutor Virtual en la aplicación, para conseguir un entorno más amigable y realista.

5. Conclusiones y líneas de futuro

En este artículo se ha presentado un resumen histórico del trabajo realizado por el grupo de investigación del área de Tecnologías del Habla de Ingeniería La Salle en el ámbito de la síntesis del habla. También se han expuesto los distintos objetivos planteados en cada uno de los períodos, los cuales han intentado estar acorde con las líneas de investigación más novedosas del momento. A partir de estos trabajos, en una fase inicial se desarrollaron tanto equipos como algoritmos, todos ellos relacionados con la síntesis del habla. Este proceso culminó en la consecución de un sistema de conversión texto-habla en catalán que presentaba un alto grado de inteligibilidad.

Recientemente, con el afán de continuar progresando en este ámbito, se ha planteado mejorar el sistema en cuanto a su naturalidad y, a corto plazo, permitir la síntesis en distintas lenguas. En el proyecto del Locutor Virtual se han tomado en consideración todos estos propósitos y, además, se les ha añadido el objetivo de la multimodalidad. Por ello se incluye un módulo de síntesis de la imagen facial, cosa que permite plantearse la consecución de un sistema de conversión texto a mensaje audiovisual. De este modo se potencia la integración e interacción de las líneas de trabajo de los distintos grupos de investigación de nuestro departamento, cuyos primeros resultados se verán reflejados en una aplicación *on-line* y en otra para la ayuda al estudio.

Para desarrollar el nuevo módulo de síntesis, se han iniciado investigaciones en los ámbitos de la síntesis basada en selección

de unidades y la síntesis híbrida entre PSOLA y los modelos armónicos-estocásticos, estudiando su aplicación a la síntesis de emociones junto al grupo de investigación del CAP de la UAB. Por otro lado, se pretende mejorar el bloque de PLN mediante la colaboración establecida con el grupo de investigación del CLiC de la UB. En cuanto a la integración del nuevo CTH en arquitecturas *normalizadas*, se está terminando la migración del código del sintetizador por concatenación de difonemas a SAPI 5.1 y a su vez se ha iniciado un estudio preliminar de *Festival Synthesis Speech System* [28] para una futura integración del sistema. Con todas estas líneas de trabajo, junto a las que también se llevan a cabo para desarrollar el bloque de síntesis de imagen, se pretende conseguir una síntesis multimodal y multilingüe, que tendrá como máximo exponente la consecución del Locutor Virtual.

6. Referencias

[1] J. Martí, *Estudi acústic del català i síntesi automàtica per ordinador*, Ph.D. thesis, Universidad de Valencia, 1985.

[2] J. Martí, "Síntesis del habla: Evolución histórica y situación actual," in *Reconocimiento automático del habla*, Boixareu Marcombo, Ed. Casacuberta, F. and Vidal, E., 1987.

[3] J. Martí, "Estado actual de la síntesis de voz," in *Estudios de Fonética Experimental*, 1990, number 4, pp. 147–168.

[4] J. Llisterri, N. Fernández, F. Gudayol, J.J. Poyatos, and J. Martí, "Testing user's acceptance of Ciber232, a text to speech system used by blind persons," in *Granström, B., Hunnicutt, S., Spens, K.-E. (Eds) Speech and Language Technology for Disabled Persons. Proceedings of an ESCA Workshop*, Stockholm, Suecia, 1993, pp. 203–206.

[5] J. Camps, G. Bailly, and J. Martí, "Synthèse à partir du texte pour le catalan," in *Proc. 19èmes Journées d'Études sur la Parole*, Bruxelles, Francia, 1992, pp. 329–333.

[6] R. Guaus, F. Gudayol, and J. Martí, "Conversión texto-voz mediante síntesis PSOLA," in *Jornadas Nacionales de Acústica*, Barcelona, España, 1996, pp. 355–358.

[7] E. Moulines and F. Charpentier, "Pitch-Sinchronous waveform processing techniques for text-to-speech synthesis using diphones," in *Speech Communication*, 1990, number 9, pp. 453–467.

[8] R. Guaus, J. Oliver, H. Moure, I. Iriondo, and J Martí, "Síntesis de voz por concatenación de unidades: Mejoras en la calidad segmental," in *TecniAcústica*, Lisboa, Portugal, 1998.

[9] R. Guaus and I. Iriondo, "Diphone based Unit Selection for Catalan Text-to-Speech Synthesis," in *Workshop on Text, Speech and Dialogue (TSD2000)*, Brno, República Checa, 2000.

[10] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Joint Meeting of ASA, EAA, and DAGA2*, Berlín, Alemania, 1999, pp. 18–24.

[11] G. Bailly, "Audiovisual speech synthesis," in *ETRW on Speech Synthesis*, Perthshire, Escocia, 2001.

[12] "Multilingual text-to-speech systems," http://www.bell-labs.com/project/tts/.

[13] A.W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *EuroSpeech*, Rodes, Grecia, 1997, pp. 601–604.

[14] G. Coorman, J. Fackrell, P. Rutten, and B. Van Coile, "Segment selection in the L&H RealSpeak laboratory TTS system," in *ICSLP*, Beijing, China, 2000, vol. 2, pp. 395–398.

[15] A. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP*, Atlanta, Estados Unidos, 1996, vol. 1, pp. 373–376.

[16] A.W. Black, "Perfect Synthesis for all of the people all of the time," in *IEEE TTS Workshop 2002 (Keynote)*, Santa Monica, Estados Unidos, 2002.

[17] "HTK ver. 3.1.1," http://htk.eng.cam.ac.uk/index.shtml.

[18] F. Alías and I. Iriondo, "Asignación automática de marcas de pitch basada en programación dinámica," in *Procesamiento del Lenguaje Natural*, Jaén, España, 2001, number 27, pp. 225–231.

[19] A. Rodríguez, P. Lázaro, N. Montoya, J. Blanco, D. Bernadas, J. Oliver, and L. Longhi, "Modelización acústica de la expresión emocional en el español," in *Procesamiento del Lenguaje Natural*, Lleida, España, 1999, number 25, pp. 159–166.

[20] I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. Blanco, D. Bernadas, J. Oliver, D. Tena, and L. Longhi, "Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques," in *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, Irlanda del Norte, 2000, pp. 161–166.

[21] Y. Stylianou, *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modifications*, Ph.D. thesis, École Nationale des Télécommunications, París (Francia), 1996.

[22] I. Iriondo, F. Alías, and J. Melenchón, "Un modelo híbrido orientado a la síntesis multimodal del habla," in *Procesamiento del Lenguaje Natural*, Valladolid, España, 2002, number 29, pp. 159–163.

[23] J. Melenchon, F. Alías, and I. Iriondo, "Previs: a person-specific realistic virtual speaker," in *ICME 2002 Proceedings (CD-ROM)*, Lausanne, Suiza, 2002.

[24] J. Melenchon, I. Iriondo, and F. Alías, "Modelo 2d parametrizado basado en imágenes reales orientado a síntesis de cabezas parlantes," in *Proceedings del XVII Simposium Nacional de la Unión Científica Internacional de Radio (URSI2002)*, Alcalá de Henares , España, 2002, pp. 383–384.

[25] "SAPI," http://www.microsoft.com/speech/.

[26] J.A. Montero, E. Martínez, J.A. Moran, F. Alías, and J. Rodríguez, "ALGTEC: un complemento a la enseñanza del álgebra lineal en carreras de ingeniería de telecomunicaciones," in *Actas de la Conferencia Internacional sobre Educación, Formación y Nuevas Tecnologías (CD-ROM)*, Virtual Educa, Valencia, España, 2002.

[27] "Microsoft Agents," http://microsoft.com/msagent/.

[28] "Festival Speech Synthesis System," http://festvox.org.