ISSN (electrónico): 2014-2242 / www.cac.cat

QUADERNS
DEL CAC

Aplicación de técnicas de generación automática del habla en producción audiovisual

Francesc Alías
Miembro del Grupo de Investigación en Tecnologías Media de La Salle - Universidad Ramon Llull

falias@salle.url.edu

Joan Claudi Socoró
Miembro del Grupo de Investigación en Tecnologías Media de La Salle - Universidad Ramon Llull

jclaudi@salle.url.edu

Ignasi Iriondo
Miembro del Grupo de Investigación en Tecnologías Media de La Salle - Universidad Ramon Llull

iriondo@salle.url.edu

Artículo recibido el 01/06/2011 y aceptado el 13/12/2011

Resumen
En este artículo se presenta un resumen del trabajo de investigación que lleva el mismo título, realizado gracias a la ayuda concedida por el CAC en la VII convocatoria de Ayudas a proyectos de investigación sobre comunicación audiovisual. Tras estudiar el grado de implantación de los sistemas de síntesis de voz en Cataluña, se analiza la viabilidad de su uso en el ámbito de la creación de producciones audiovisuales. En este artículo se presentan las conclusiones del estudio de campo realizado y los experimentos desarrollados a partir del sistema de síntesis del habla de La Salle (Universidad Ramon Llull) adaptado al catalán.

Palabras clave
síntesi de veu, produccions audiovisuals, audiodescripció, valoració subjectiva de qualitat.

Abstract
This article presents a summary of the research work of the same title, developed thanks to the grant awarded by the CAC in the VII call of research projects on audiovisual communication. After studying the degree of implementation of speech synthesis systems in Catalonia, we analyze the feasibility of its use for the creation of audiovisual productions. This article presents the findings of the field study and the experiments developed after adapting the speech synthesis system of La Salle (Universitat Ramon Llull) to the Catalan language.

Keywords
speech synthesis, audiovisual productions, audio description, subjective assessment of quality.

1. Introducción

La síntesis de voz o del habla es la técnica que permite generar automáticamente una locución con características similares a las de una voz humana a partir de un texto de entrada. Los sistemas de síntesis de voz pueden llegar a confundirse con los sistemas que hacen un uso de voz grabada para la reproducción de mensajes de voz, pero hay que tener claro que, en general, la síntesis de voz se refiere a las técnicas que permiten generar cualquier mensaje oral.

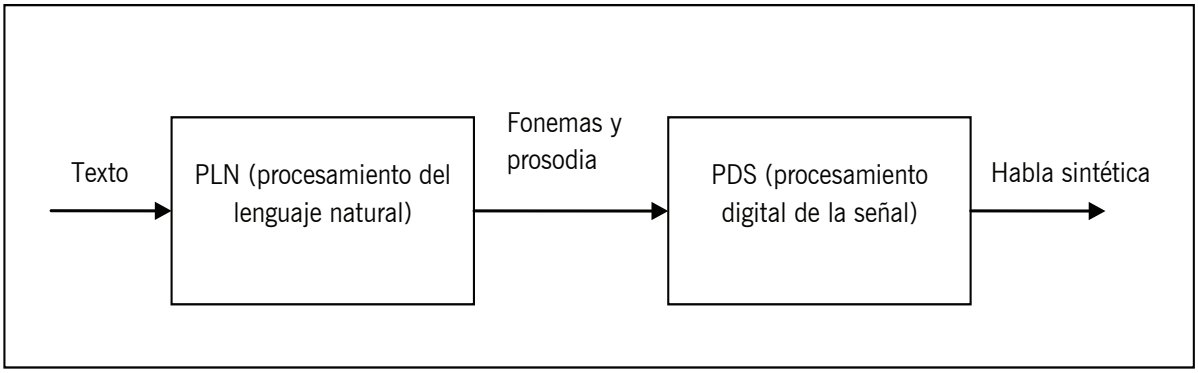
El texto de entrada puede provenir de un correo electrónico, de una web o bien puede escribirse directamente desde un teclado. Algunas de las aplicaciones típicas de este tipo de sistemas son la ayuda a personas con una determinada discapacidad (por ejemplo, visual), el apoyo para el aprendizaje de lenguas, las aplicaciones telefónicas, las aplicaciones multimedia y las interfaces persona-máquina en general.

Lejos de querer imitar el proceso real con el que los humanos generamos el habla, existe un modelo funcional que permite abordar, con los recursos disponibles hoy en día, la construcción de un sistema que convierta un texto de entrada cualquiera en su correspondiente voz sintética. Este modelo funcional, extendido y ampliamente aceptado por la comunidad dedicada a las tecnologías del habla, es lo que se describe en el diagrama de bloques de la figura 1.

Como puede observarse en la figura 1, en primer lugar tenemos el bloque de procesamiento del lenguaje natural del habla (PLN), que es el encargado de encontrar, a partir del texto de entrada que se quiere "leer", cuál es la transcripción fonética del texto (es decir, cuáles son los sonidos que se producirán a lo largo de la locución de salida) y también cuál debe ser la prosodia asociada (cómo deben sonar cada uno de estos sonidos, específicamente sobre sus características tanto de entonación como de ritmo). En segundo lugar, aparece el bloque de procesamiento digital de la señal (PDS), que se encarga de generar, a partir de los requerimientos dados por el módulo anterior, la señal de habla sintética de salida.

El resto de este artículo está estructurado de la siguiente

Figura 1. Modelo funcional de un sistema de conversión de texto en habla (CTP)



Fuente: Elaboración propia.

forma: la sección 2 presenta una recopilación de entidades que destacan por su aportación en el mundo de la síntesis de voz. La sección 3 muestra los resultados más representativos de un estudio de campo sobre la síntesis de voz en el entorno del audiovisual en Cataluña, así como en colectivos de personas con capacidades visuales reducidas. Mediante un conjunto de entrevistas personales, se han recogido las opiniones más relevantes en relación con el grado de madurez alcanzado por esta tecnología, las limitaciones que más destacan de su uso y los retos de futuro que deben permitir en un futuro un grado más elevado de penetración de la conversión de texto a habla en dichos sectores. La sección 4 describe el proceso de adaptación del sintetizador de voz de La Salle (URL) al catalán, siguiendo uno de los objetivos fijados en el proyecto de investigación financiado. Utilizando este sintetizador, se han realizado las pruebas que describe la sección 5, que han permitido validar de modo subjetivo la viabilidad del uso de la síntesis de voz como herramienta para generar material audiovisual (concretamente, con ejemplos de anuncios y noticias). Finalmente, la sección 6 incluye las principales conclusiones de este trabajo y las líneas de investigación que pueden permitir avanzar hacia un mayor grado de implantación de la síntesis de voz en los medios audiovisuales.

2. Implantación de la síntesis de voz en el mundo audiovisual en Cataluña

Para estudiar el grado de implantación real de las tecnologías de síntesis de voz en Cataluña en el mundo del audiovisual, se ha realizado un trabajo de campo extenso para recoger las opiniones de sus principales actores ante la implantación actual y la posible introducción futura de los sistemas de síntesis de voz en los medios de comunicación audiovisual. Además, durante este proceso se ha podido constatar que hay una parte de la población, las personas con discapacidad visual, que son grandes consumidoras de los sistemas de síntesis de voz. Por este motivo, este grupo de usuarios también se ha incluido en el estudio realizado para conocer su opinión respecto al uso de las tecnologías de síntesis del habla en el contexto de las producciones audiovisuales (Torrens 2010).

A continuación, se presenta una recopilación representativa de las empresas, centros de investigación y productos más relevantes en el contexto de la generación de voz sintética en catalán. En este contexto, se recogen tanto empresas de ámbito catalán como internacional, así como productos que se encuentran en la red.

2.1. Universidades y centros de investigación

1) TALP (Tecnologies i Aplicacions del Llenguatge i de la Parla) de la Universitat Politècnica de Catalunya

En cuanto a la síntesis de voz en catalán cabe destacar, por un lado, que el grupo de investigación TALP dispone de un sistema propio de conversión de texto a habla, llamado OGMIOS (*http://www.talp.cat/ttsdemo/index.php*), y, por otro, que trabajaron en la incorporación del catalán en la plataforma Festival, desarrollada por el sistema operativo Linux (*http://www.cstr.ed.ac.uk/projects/festival/*), y el resultado es Festcat, que se incluyó en la distribución Linkat de la Generalitat de Cataluña. Todas estas aplicaciones pueden descargarse gratuitamente desde la web de Festcat y se publican bajo los términos de la licencia LGPL. Para más información, consultar la web <http://gps-tsc.upc.es/veu/festcat/>.

Parte de este trabajo se desarrolló en el marco del proyecto Tecnoparla: Tecnologies de la parla en català, enfocado a estudiar la viabilidad de la traducción de voz aplicada a la traducción de noticias audiovisuales. El proyecto estudió las diferentes tecnologías clave que intervienen en un sistema de traducción de voz (reconocimiento, traducción y síntesis de voz), se centró en la incorporación del catalán y abordó el progreso en las tres tecnologías implicadas y su integración. Concretamente, con relación a la síntesis de voz se utilizó el sistema de software abierto Festival (Linux) adaptado al catalán (Festcat). Puede encontrarse más información en la siguiente web: <http://www.talp.cat/tecnoparla/>

2) GTM (Grup de Recerca en Tecnologies Mèdia), La Salle - Universitat Ramon Llull

Este grupo tiene una amplia experiencia en el mundo de la generación del habla sintética. Desde sus inicios, a finales de los años ochenta, ya se centró en la investigación en síntesis del habla en catalán, a través de trabajos de Martí (1985) y Campos (1992), posteriormente continuados por Guaus e Iriondo (2000) e Iriondo *et al.* (2004), este último trabajo enfocado a la síntesis expresiva (emotiva) en catalán. Para más información, puede consultarse la siguiente web:

<http://www.salle.url.edu/portal/departaments/home-depts-DTM-projectes-PM?cf_seccio=PM&pag=1>

3) Barcelona Media - Centre d'Innovació de la Fundació Barcelona Media

Barcelona Media incorpora una línea de investigación en voz y lenguaje que centra la investigación en el procesamiento del lenguaje, tanto escrito como oral, y desarrolla aplicaciones en corrección y traducción automáticas, análisis y procesamiento de la información, generación automática de textos a partir de bases de datos, y síntesis de voz, para disponer de herramientas para el procesamiento automatizado de contenidos lingüísticos en entornos multilingües o en los que el lenguaje humano se convierte en la modalidad de interacción prioritaria.

En el ámbito de la síntesis de voz trabajan con el objetivo de crear una voz sintética catalana, una castellana y una bilingüe (catalana y castellana), así como de introducir naturalidad expresiva y entonativa (prosodia) y facilitar la creación de locutores especializados. Puede encontrarse más información en la siguiente web: < http://www.barcelonamedia.org/linies/7/ca>.

2.2. Empresas

1) Verbio

Empresa dedicada a vender productos relacionados con las tecnologías del habla ubicada en Barcelona.

- En cuanto a la síntesis del habla, ofrecen conversión de texto a habla en diferentes idiomas.
  <http://www.verbio.com/webverbio3/html/productes.php?id=1>
- Demostraciones de las voces en catalán: Meritxell y Oriol.
  <http://www.verbio.com/webverbio3/html/demos_ttsonline.php>
- Demostraciones de noticias:
  <http://www.verbio.com/webverbio3/html/demos_news.php>
  Enlaza a Vilaweb.cat, pero indica que no hay noticias disponibles.

2) Loquendo

Empresa dedicada a la venta de productos relacionados con las tecnologías del habla.

- En cuanto a la síntesis del habla, ofrecen conversión de texto a habla en diferentes idiomas. Se trata de un sistema de síntesis de voz basado en selección de unidades.
  <http://www.loquendo.com/es/technology/tts.htm>
- Demostraciones de las voces en catalán: Montserrat y Jordi.
  <http://www.loquendo.com/es/demos/demo_tts.htm>

3) CereProc

La empresa CereProc, en colaboración con Barcelona Media, ha desarrollado un sistema de síntesis de voz femenina bilingüe en catalán y en castellano. Ofrecen una voz femenina sintética, bilingüe, en catalán y en castellano, con entonación natural, disponible para múltiples aplicaciones. El proyecto ha contado con el apoyo de la Generalitat de Cataluña.

<http://www.cereproc.com/products/voices>

4) Nuance

Nuance Vocalizer (antes RealSpeak) dispone de una voz femenina en catalán (Núria). Sin embargo, no se puede encontrar demasiada información en la web de la empresa.

<http://www.nuance.es/realspeak/>

<http://www.nuance.com/for-business/by-solution/contact-center-customer-care/cccc-solutions>

<services/vocalizer/vocalizer-languages/index.htm>

5) Telefónica I+D

Dispone de un sistema de conversión de texto en habla multilingüe (Rodríguez *et al.* 2008). No se ha encontrado información que nos permita afirmar que se trata de un producto independiente que ofrece la empresa (ver *http://www.tid.es*). Sin embargo, es una tecnología que la empresa ha incorporado a alguno de sus productos, como el lector de mensajes cortos (*http://saladeprensa.telefonica.es/documentos/24moviles.pdf*) o para ayudar a personas con discapacidad (*http://saladeprensa.telefonica.es/documentos/22comunicador.pdf*).

2.3. Otros productos

1) eSpeak

eSpeak es un sistema de síntesis basado en formatos que trabaja bajo las plataformas Linux y Windows, y que puede utilizarse bajo la licencia GNU *General Public License* (software libre).

<http://espeak.sourceforge.net/>

2) JAWS (*Job Access With Speech*)

Está dirigido a personas ciegas o de baja visión.

- Lee el contenido de la pantalla mediante voz sintética.
  <http://www.freedomscientific.com/products/fs/jaws-product-page.asp>
- Incorpora la voz en catalán debido a que incorpora sistemas de síntesis de otras empresas, como puede ser Nuance (Núria).
  <http://www.freedomscientific.com/downloads/jaws/JAWS10-whats-new.asp>

3. Implantación de la síntesis de voz en el mundo audiovisual en Cataluña

Para estudiar el grado de implantación real de las tecnologías de síntesis de voz en Cataluña en el mundo del audiovisual, se ha realizado un extenso trabajo de campo para recoger las

opiniones de sus actores principales ante la implantación actual y la posible introducción futura de los sistemas de síntesis de voz en los medios de comunicación audiovisual. Además, durante este proceso se ha podido constatar que hay una parte de la población, las personas con discapacidad visual, que son grandes consumidores de los sistemas de síntesis de voz. Es por ello que este grupo de usuarios también se ha incluido en el estudio realizado para conocer su opinión respecto al uso de las tecnologías de síntesis del habla en el contexto de las producciones audiovisuales. Los detalles del trabajo de campo pueden encontrarse en Torrens (2010).

A continuación, se analizan los resultados obtenidos del trabajo de campo a partir de las diferentes respuestas recogidas de las entrevistas que se han realizado a los principales actores del sector (emisoras de radio, televisión, productoras y estudios de sonido y doblaje) mediante entrevistas realizadas a personas que trabajan en este sector, tanto desde la vertiente técnica como desde la no técnica.

Además, también se ha entrevistado a un grupo de usuarios potencialmente muy interesado en la inclusión de la síntesis de voz en el mundo de la comunicación audiovisual, como es el de las personas con discapacidad visual. A continuación, se presentan las conclusiones del estudio contextualizadas para este sector de la sociedad.

3.1. Medios de comunicación

Las entrevistas realizadas a los medios de comunicación audiovisual se han desglosado en tres grandes grupos: 1) radios, 2) televisiones y productoras de televisión, y 3) estudios de audio, de doblaje y de postproducción. Se contactó con la gran mayoría de entidades líderes del sector dentro del territorio catalán, ya que el estudio se centraba en la aplicación de la síntesis de voz en catalán. De todas las entidades mencionadas, atendieron la entrevista las siguientes 19 entidades mediante un representante de sus departamentos técnicos / de emisiones (ver Torrens 2010 para más detalles):

1. Radios: Catalunya Ràdio, 40 Principales Barcelona, COMRàdio, RAC 1 y Onda Rambla - Punto Radio.
2. Televisiones y productoras: TV3, 8tv, RAC105tv y Gestmusic.
3. Estudios de audio, de doblaje y de postproducción: Oido (*http://www.oido.net/*), INFINIA (*http://www.infinia.es/*), Onda Estudios (*http://www.ondaestudios.com/*), Cyo Studios (*http://www.cyostudios.com/*), Dubbing Films (*http://www.dubbingfilms.com/*), Tatudec (*http://www.tatudec.com/*), Dvmusic (*http://www.dv-music.es/*), Seimar RLM Estudios, Soundub (*http://www.abaira.es/*) y Sonygraf (*http://www.sonygraf.com*).

De estas entrevistas, se puede concluir lo siguiente:

- Tanto las radios y las televisiones como los estudios de sonido son conocedores de la tecnología de los sistemas de síntesis de voz.
- Analizando el primero de los grupos, ninguna de las emisoras de radio con las que se ha contactado utiliza los sistemas de síntesis de voz, salvo un par que lo han usado sólo para generar voz robótica o para crear algún efecto concreto, y lo han hecho utilizando software libre.

Hay diversas opiniones respecto al uso de las tecnologías de síntesis del habla en un futuro: dos de las personas representantes de los departamentos técnicos de las emisoras creen que podrían ser útiles, pero sólo de modo complementario, es decir, para la creación de efectos o para emisoras automatizadas. Otra expone que se perdería el encanto y la magia que da un medio como la radio; las dos restantes piensan que los sintetizadores aún se encuentran lejos de ser utilizados por la falta de expresión y de entonación natural en la voz.

- En ninguna de las televisiones ni en la productora con las que se ha podido contactar se utilizan los sistemas de síntesis de voz para generar productos audiovisuales. Sin embargo, la opinión de los técnicos consultados es bastante variada. En un caso, se indica que no interesan porque lo que gusta es la voz humana. Por el contrario, se expone que podrían utilizarse en programas automáticos que den información sobre la bolsa o el tiempo y, también, en anuncios publicitarios, documentales y promociones por el gran ahorro económico que supondría en la generación de estos productos. Esta última indicación se ha extraído de la entrevista realizada al representante técnico del Departamento de Audio de la productora de televisión Gestmusic. Aunque algunos técnicos vean viable aplicar voz sintética para diversas aplicaciones, también indican que los sistemas de síntesis de voz deberían madurar a nivel de naturalidad para poder producir diversas entonaciones (voces agudas, graves, juveniles, serias,...) .
- Sólo dos de los departamentos técnicos del último grupo (estudios de sonido, de doblaje y de postproducción) han utilizado alguna vez un sintetizador de voz, pero sólo para crear efectos en el ámbito musical o para manipular voces. La opinión general respecto a la implantación de estos sistemas de comunicación en un futuro es muy similar en todos los estudios consultados. La gran mayoría de las personas entrevistadas destaca que hasta que los sistemas de síntesis de voz no estén más perfeccionados –en el sentido de aumentar la naturalidad de la voz sintética generada para transmitir emociones de forma realista, tal y como lo hace una persona–, la voz sintética no se podrá utilizar ni en el sector de la televisión ni en el de la radio.

Como valoración global de la idea de la introducción de los sistemas de síntesis de voz en los medios de comunicación audiovisual, puede decirse que las opiniones de los 19 técnicos entrevistados, en principio contrarios a integrarlos en el proceso de creación de contenidos audiovisuales, podrían cambiar si se llegaran a sintetizar de forma natural las emociones en la voz y se consiguieran voces sintéticas menos robóticas y, por lo tanto, más parecidas a la voz natural producida por el ser humano.

3.2. Usuarios potenciales

En cuanto a las entrevistas realizadas en el contexto de las tecnologías para las personas con discapacidad visual, las entrevistas se han realizado a dos perfiles diferentes: 1) los técnicos que trabajan en los medios de comunicación recogidos en el apartado anterior, para conocer su opinión respecto al uso de voz sintética para la audiodescripción (tecnología que ellos ya conocen), y 2) el sector de la población que sufre algún tipo de discapacidad visual, ya que es esencial considerar su opinión para conocer la viabilidad de la introducción de voz artificial en estos medios.

La mayoría de personas dedicadas a las tecnologías del sonido (englobando a los técnicos de la radio, de la televisión y de estudios de audio, de doblaje y de postproducción entrevistados) cree que podría aplicarse voz sintética en la audiodescripción si fuera más natural y "creíble", aunque, en algunos casos, se piensa que tampoco supone un gran ahorro de tiempo y que no vale la pena sustituir la voz natural. Concretamente, se han recogido opiniones en el sentido de que los sistemas de síntesis de voz deberían mejorar mucho en cuanto a calidad sintética e, incluso, se afirma que es más rápido grabar con una persona. Sin embargo, en el conjunto de las entrevistas, ha habido dos que destacan especialmente debido a que son claramente diferentes de las demás. Concretamente, en ellas se indica que:

- Antes de incorporar las tecnologías de síntesis de voz a la producción audiovisual, debería preguntarse a las personas con discapacidad visual, que realmente son sus usuarios finales, sobre la viabilidad de usar voz sintética para la audiodescripción y si no les gusta, debería descartarse esta opción.
- Siempre es mejor si la entonación y la naturalidad del mensaje son buenas, pero los costes pueden ser un factor clave. En este sentido, aunque la voz sintética no sea del todo natural, puede permitir abaratar los costes de la creación del audio y, por lo tanto, puede ser más rentable que contratar a un locutor o locutora.

3.3. Usuarios con discapacidad visual

Mediante la colaboración con la ONCE, se han podido entrevistar a 51 personas con discapacidad visual de todo el territorio español. La distribución por territorios de las personas consultadas es la siguiente: 16 personas de Madrid, 8 de Andalucía, 5 de la Comunidad Valenciana, 3 de Cataluña, 3 de Galicia, 3 de las Islas Baleares, 3 de Asturias, 2 de Canarias y un solo representante en el resto de territorios (Cantabria, País Vasco, Castilla y León, etc.). La distribución por profesiones es la siguiente: 10 jubilados, 9 agentes vendedores de cupones, 7 fisioterapeutas, 2 estudiantes, 2 maestros y un único representante en las demás profesiones (programador, músico, logopeda, periodista, abogado, administrativo, telefonista, trabajador de banca, etc.). Por lo tanto, en las entrevistas se ha querido recoger un amplio espectro de perfiles de usuarios potenciales con discapacidad visual (ver Torrens 2010 para una enumeración detallada).

De las entrevistas realizadas a las personas con discapacidad visual, ya sea total o parcial, se extraen dos ideas muy interesantes relacionadas con los medios de comunicación:

- Casi todas las personas que han colaborado respondiendo al cuestionario creen que en un futuro se podría utilizar voz sintética para la audiodescripción en la televisión y en el cine. Indican que sería muy interesante que una voz les explicara todo lo que no pueden ver en programas de televisión, documentales, películas... Todo lo que les permita una normalización y una integración en el consumo de productos audiovisuales es bienvenido.
- Respecto a la introducción de los sistemas de síntesis de voz en la radio, las opiniones son diversas. Más de la mitad creen que es innecesario y prefieren la voz humana. Del resto de entrevistas, algunas consideran que puede ser útil, dependiendo de la calidad de las voces sintéticas, y otras, aunque lo aceptan, no creen que sea imprescindible.

Finalmente, podemos concluir que el día que se consiga naturalidad y emotividad en las voces sintéticas, la audiodescripción puede ser una buena vía para introducir progresivamente los sistemas de síntesis de voz en el mundo de las producciones audiovisuales, ya que casi todas las personas con discapacidad visual utilizan estos sistemas. Mientras se espera este avance en las voces, la viabilidad de introducir los sistemas de síntesis de voz en la radio o la televisión parece difícil, pero existe la opción de utilizarlos en sectores o en aplicaciones en los que no sea necesario la expresividad o se quiera modelar una voz robótica.

4. Adaptación del sistema de síntesis de La Salle al catalán

En el ámbito técnico del proyecto, una de las fases clave ha sido la encargada de desarrollar los recursos lingüísticos y de procesamiento de la señal para la creación de las voces en catalán. Los recursos lingüísticos, como el sistema de transcripción fonética, el analizador morfosintáctico, etc., que forman parte del módulo de procesamiento del lenguaje natural (PLN) del sistema de síntesis son propios y han sido desarrollados en el marco del grupo de investigación durante los últimos años de investigación. En cambio, las bases de datos de síntesis de voz en catalán son públicas y han sido desarrolladas por el grupo de investigación TALP de la Universitat Politècnica de Catalunya, con financiación de la Generalitat de Cataluña, en el marco del proyecto Festcat (*http:// gps-tsc.upc.es/veu/festcat*).

De este proyecto se han escogido las dos voces –Ona y Pau– que tienen más extensión, ya que el sistema de síntesis de voz del Grup de Recerca en Tecnologies Mèdia de La Salle (URL) está basado en la técnica de selección de unidades en función de los parámetros predichos por el modelo prosódico.

Una vez se dispone de los ficheros de voz, hay que "crear una nueva voz" por el sistema de síntesis, es decir, hay que procesar las muestras de voz para que sean útiles para generar voz sintética. La creación de una nueva voz consta de tres partes principales:

1. La segmentación de la base de datos en unidades de síntesis, que se encargan de determinar el inicio y el final de cada una de las unidades acústicas (difonemas, en este caso) que integran los mensajes grabados en los archivos de voz.
2. La indexación y la parametrización de la base de datos, que se encargan de generar el conjunto de ficheros en formato XML que contienen los parámetros que describen el contenido acústico de la base de datos (duración, energía, frecuencia fundamental de las unidades). Al mismo tiempo, hay que ajustar la función de coste de selección, cuestión que implica, por una parte, precalcular todos los costes de las unidades de la base de datos y, por otra, ajustar los pesos de la función de coste (Alías *et al.* 2011).
3. El entrenamiento del modelo prosódico, que es el encargado de determinar la pronunciación más adecuada de un texto de entrada a sintetizar a partir de la extracción de patrones prosódicos que se extraen las muestras de voz disponibles (Iriondo *et al.* 2007).

Una vez finalizadas estas tres fases, ya se dispone de las voces Ona y Pau integradas en el sistema de síntesis de voz de La Salle para realizar los experimentos que tienen el objetivo de analizar la viabilidad del uso de la síntesis de voz en producciones audiovisuales y que se describen a continuación.

5. Experimentos y resultados

En el ámbito de la síntesis del habla pueden evaluarse diferentes características, como la inteligibilidad, la naturalidad y la expresividad. En algunas aplicaciones, como por ejemplo, en las máquinas parlantes para personas invidentes, la inteligibilidad del habla a alta velocidad es más importante que la naturalidad (Llisterri *et al.* 1993). En cambio, una correcta prosodia y una elevada naturalidad son esenciales en la mayoría de aplicaciones multimedia. La evaluación puede realizarse a diferentes niveles (segmento, palabra, frase o párrafo) y con diferentes tipos de pruebas (Campbell 2007).

A fin de disponer de una evaluación subjetiva de la viabilidad del uso de la síntesis de voz en el momento de generar material audiovisual, se han preparado dos tests perceptivos: uno de anuncios publicitarios y otro de noticias. Por cada test, se prepararon un conjunto de parejas de estímulos. Cada pareja tenía el mismo contenido verbal, pero una estaba generada con el sistema de síntesis y la otra era leída por una persona. Una vez preparados los estímulos, se decidió el tipo de prueba más adecuada para presentar a los oyentes y la metodología de evaluación de los mismos. En el caso de los anuncios, sólo llevaban el canal de audio, mientras que en el caso de las noticias eran vídeos con imágenes relacionadas con la noticia y el canal de audio formado por la pista de sonido de fondo (música, ruido de calle, voces, etc.) superpuesta a la pista de voz en *off.*

Como ya se ha señalado, el objetivo de la prueba ha consistido en evaluar la síntesis del habla en anuncios o en noticias. Se disponía de una pareja de ficheros de audio (anuncios) o vídeo (noticias) por cada elemento que había que evaluar. Se plantearon diferentes posibilidades de presentación de los estímulos (de modo individual o por parejas) y de escalas de puntuación. A partir de la recomendación P.800 de la Unión Internacional de Telecomunicaciones (UIT) (UIT-T 1996), se escogió el índice de evaluación comparativa *Comparison Mean Opinion Score* (CMOS), que permite comparar dos estímulos, *A* y *B,* como:

- A mucho mejor que B
- A mejor que B
- A ligeramente mejor que B
- Ninguna preferencia
- B ligeramente mejor que A
- B mejor que A
- B mucho mejor que A

Con esta escala, los oyentes pudieron evaluar comparativamente los dos estímulos presentados escuchándolos tantas veces como era necesario.

5.1. Anuncios publicitarios

Para evaluar el uso de la síntesis del habla en situaciones reales, se elaboró un test con siete anuncios publicitarios. Por cada anuncio, se generaron dos ficheros de sonido, uno a partir de la lectura del anuncio por parte de una locutora *amateur* y el otro, utilizando nuestro sintetizador de habla en catalán.

El test se realizó con la plataforma *on-line* TRUE (Testing platfoRm for mUltimedia Evaluation) (Planet *et al.* 2008), que permite diseñar y realizar el test de forma remota.

Por cada pareja de audios asociados al mismo anuncio, al participante del test se le formularon dos preguntas:

1. "Los audios siguientes *(A* el de arriba, *B* el de abajo) corresponden a dos lecturas de anuncios publicitarios. No se trata de evaluar si te gusta más la voz de una mujer u otra, sino, para un uso en publicidad, indica tu preferencia, fijándote en la NATURALIDAD de la pronunciación y la entonación:"
2. "En cuanto a la INTELIGIBILIDAD, ¿qué te parece?"

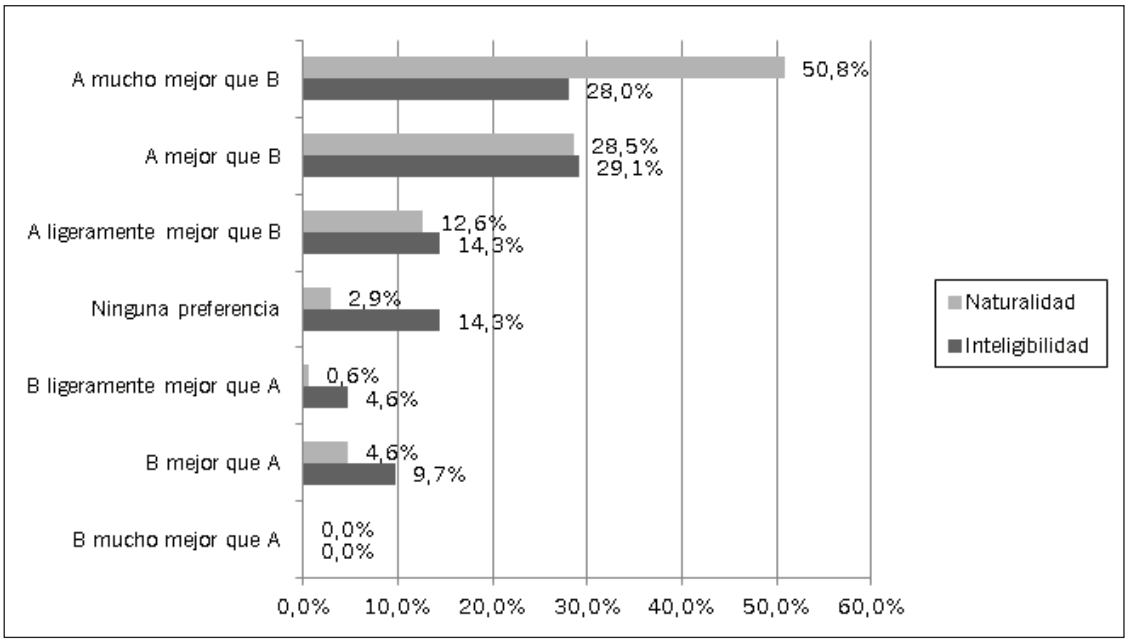
El test lo realizaron 25 oyentes (12 mujeres y 13 hombres) de edades comprendidas entre los 18 y los 66 años.

Los resultados de preferencia obtenidos con este test se muestran en la figura 2, donde *A* representa la voz natural y *B,* la voz generada con el sintetizador. Los resultados, como es de esperar, muestran una preferencia clara por la voz natural, especialmente en cuanto a naturalidad, aunque en inteligibilidad la diferencia no es tan grande.

5.2. Vídeos de noticias

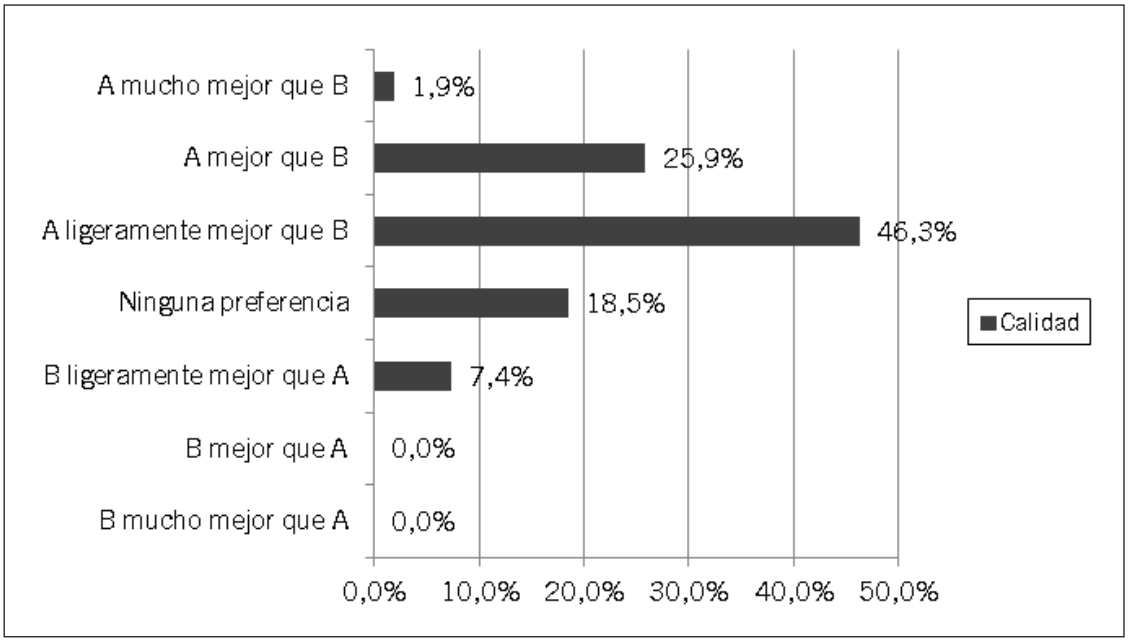
En este experimento se ha querido añadir a la voz dos componentes habituales en el material audiovisual: la imagen y una pista de sonido adicional a la de voz. Se preparó un test con tres parejas de noticias. A partir de material extraído de YouTube y de la voz generada con nuestro sintetizador, se generaron

Figura 2. Resultados del test de anuncios publicitarios en cuanto a inteligibilidad y naturalidad. A se corresponde a la voz natural y B, a la voz sintetizada



Fuente: Elaboración propia.

Figura 3. Resultados del test de vídeos de noticias en cuanto a calidad de la voz en off. A se corresponde a la voz natural y B, en la voz sintetizada



Fuente: Elaboración propia.

vídeos de noticias que contenían tres pistas: la de vídeo propiamente y dos de audio (sonido de fondo y voz).

El test también se realizó con la plataforma TRUE y se trataba de un CMOS de siete categorías. Participaron 20 personas (17 hombres y 3 mujeres) de edades comprendidas entre los 24 y los 41 años. A los usuarios no se les informó del origen de las dos voces. Al final del test se preguntó el sexo y la edad del participante y si era experto en tecnologías del habla, y se le formuló dos preguntas de respuesta abierta:

1. "La voz del vídeo de abajo ha sido generada por ordenador, ¿qué te ha parecido?"
2. "¿Crees que es factible el uso de síntesis de voz para explicar noticias en programas que se generen automáticamente?"

Tabla 1. Selección de respuestas a la pregunta "La voz del vídeo de abajo ha sido generada por ordenador, ¿qué te ha parecido?"

"Bastante aceptable, aunque un poco lenta y con algunos errores en sonidos concretos."
"Buena calidad en general, aunque hay algunas discontinuidades y saltos en la entonación."
"Bastante lograda, pero en ciertos momentos se nota que no es humana."
"A veces muy bien (incluso mejor que la original), otras no. Los "gallos" puntuales hacen bajar su calidad global."
"Poco natural, aunque se notaba un poco de expresividad y la calidad del audio estaba muy bien conseguida. Tal vez problemas en el hecho de mantener un ritmo constante, se notan saltos de ritmo."
"La voz es un poco metálica. La entonación no es suficientemente natural. En todo momento notas, sin duda, que te está hablando una máquina. Pese a todo, el mensaje se entiende correctamente."
"Bastante bien, sobre todo en el primero. El sonido de fondo disimula los errores. En función de la temática, el estilo de locución debería variar (por ej., en ambiente festivo, habla más ágil). "
"Se nota que es una voz sintética, pero no es molesto porque se integra bien con la música y las imágenes, y su calidad permite que se entienda bien todo lo que dice, incluso mejor, a veces, que la real."
"Bastante buena en cuanto a verosimilitud de voz humana y de entonación. El hecho que la convierte de menos calidad que la humana son unos ruidos, "clics", que aparecen de vez en cuando."
"En el primer test la calidad era bastante buena, mientras que en el resto la calidad ha decaído. Se nota bastante la concatenación entre unidades."
"Basta buena, el principal problema son los artefactos de coarticulación, que restan naturalidad a la voz."
"Bastante buena teniendo en cuenta que es audio sintético. De todos modos, se nota bastante que no es una voz humana natural."
"Calidad aceptable. El único problema que detecto que se repite a menudo es el alargamiento/arrastre de algunas vocales y consonantes."
"La voz es correcta y clara, pero de vez en cuando hace sonidos extraños y suena como distorsionada."

Fuente: Elaboración propia.

Los resultados obtenidos se muestran en la figura 3, donde puede observarse cómo la respuesta mayoritaria es que la voz natural es ligeramente mejor que la sintética (46,3%). Cabe destacar que prácticamente un 26% de las respuestas (18,5% de *ninguna preferencia* más un 7,4% de *la voz sintética es ligeramente mejor que la natural*) indican que la voz sintética es aceptable en este contexto.

Si analizamos las respuestas de los participantes que han manifestado, tras realizar el test, su opinión respecto al uso de la síntesis del habla para generar noticias, podemos destacar dos ideas generales. En primer lugar, que los oyentes son muy sensibles a errores puntuales en una determinada parte del texto y que falta mejorar la expresividad y el ritmo. En segundo lugar, la opinión mayoritaria es que el uso de esta tecnología lo ven factible para generar noticias de última hora, por ejemplo para la web o en programas de generación semiautomática. Para ilustrar ambas conclusiones, a continuación se reproduce un amplio conjunto de las respuestas obtenidas en las tablas 1 y 2.

Si comparamos los resultados con el test de anuncios publicitarios podemos comprobar que el hecho de añadir vídeo y sonido de fondo ayuda a disimular los errores de síntesis y a desviar la atención, con lo cual mejora la aceptabilidad de utilizar voz sintética.

Los archivos de audio y de vídeo generados por los experimentos pueden encontrarse en la siguiente web: <http://www.salleurl.edu/portal/departaments/home-depts-DTM-projectes-info?id_projecte=67>

6. Conclusiones y líneas de futuro

En este trabajo, tras revisar el estado de la cuestión en el ámbito de la síntesis de voz (también conocido como *sistemas de conversión de texto en habla*), se ha estudiado la situación

Tabla 2. Selección de respuestas a la pregunta "¿Crees que es factible el uso de síntesis de voz para explicar noticias en programas que se generen automáticamente?"

"Sí, lo veo factible e interesante."
"Sí, especialmente si se trata de noticias cortas y de última hora, de modo que sea más adecuada una producción semiautomatizada que haga posible disponer con mayor celeridad de los contenidos."
"En un futuro deberá ser más que viable."
"No sería factible para un telediario para televisión, por ejemplo, pero quizá sí para contenido en la web, donde la calidad del contenido no es lo que prima, sino el contenido en sí mismo."
"Le falta naturalidad y expresividad, que ayudan a hacer una noticia más atractiva. Sin embargo, la inteligibilidad es muy buena y el mensaje puede transmitirse perfectamente. Sería factible."
"Sí. Pese a la falta de naturalidad, que es mejorable. El resultado es bastante satisfactorio."
"Sí. Los pequeños problemas con la síntesis quedan bajo la pista sonora de la noticia y no suponen un problema para entenderla. Además, formalmente la locución es correcta (tono neutro)."
"Sí. Es igual de inteligible que la voz humana."
"Sí, pero dependiendo del ámbito en el que se aplique. Si es en plataformas web, creo que a nivel de usuario puede aceptarse esta calidad. "
"Sí, siempre que se eviten los artefactos antes mencionados."
"Sí me parece factible, pero no tal y como está ahora el TTS. Aún le falta más naturalidad. La voz que genera ahora resulta demasiado desagradable para un locutor al que tienes que escuchar habitualmente."
"La comprensión es perfecta. Si se pudiera mejorar el tema de las pequeñas distorsiones haría el seguimiento de las noticias más agradable. "

Font: Elaboración propia.

de esta tecnología en Cataluña y, concretamente, en el ámbito de las producciones audiovisuales. En la actualidad hay varios centros de investigación y empresas que trabajan en el desarrollo y mejora de los sistemas de síntesis del habla en catalán. Sin embargo, la implantación de estos sistemas en el contexto de la generación de producciones audiovisuales todavía es muy reducida. Dada esta situación, se ha evaluado la viabilidad de la implantación de esta tecnología en el mundo de las producciones audiovisuales, a partir de un trabajo de campo que ha consistido en varias entrevistas tanto a personal técnico como a usuarios potenciales, así como un conjunto de experimentos diseñados para estudiar el grado de aceptación de la síntesis en ejemplos reales.

Tanto de las entrevistas como de los experimentos realizados, puede concluirse que el uso de voz sintética en contenido *broadcast* puede ser una realidad en los próximos años si se mejoran ciertos aspectos relacionados con el hecho de conseguir la expresividad propia del contenido. Otro aspecto importante es el número de modos que forman parte del contenido. Si la voz va acompañada de otros elementos de audio superpuestos así como del canal de vídeo, entonces el uso de voz sintética se prevé más factible. En cambio, en contenidos donde sólo hay voz (p. ej. un anuncio publicitario para radio), la exigencia de los oyentes sobre la calidad de esta voz es mucho mayor.

El uso de la síntesis de voz (no sólo en catalán) como medio para disponer de sistemas más automatizados y capaces de servir contenidos en un formato más natural y que permita también más capacidad de incluir a todo el mundo, es uno de los retos en el que ya se está trabajando. En este contexto, hay estudios (*http://www.daisy.org/benefits-text-speech-technology-and-audio-visual-presentation*) que afirman que la inclusión de la modalidad acústica como forma alternativa de presentar contenidos puramente en un formato textual, por ejemplo, permiten aumentar la capacidad de retención, siendo por tanto una forma apropiada de presentación para actividades de aprendizaje en entornos de un carácter más audiovisual. Existen ya empresas que basan su negocio en dar servicios de voz automatizada a partir de información textual, como IVO Software (*http://www.ivona.com*), Odiogo (*http://www.odiogo.com/*) o NextUp.com (*http://www.nextup.com*), que permiten, por ejemplo, incorporar información oral a una web o dar soluciones para generar voz de forma automática a partir de documentos de texto. Aunque soluciones como estas nos permitirán cada vez más disponer de sistemas con un mayor grado de

adaptación a la persona usuaria, aún estamos lejos de ver sistemas que actúen como lo hacemos las personas y eviten cualquier mínimo artefacto sonoro o acentúen los rangos de expresividad propios de una voz humana. En todo caso, las soluciones que hoy en día podemos encontrar son soluciones que aún no nos permiten encontrar un mensaje de calidad equiparable a una locución hablada por una persona real en una conversación real, pero nos vamos acercando, y los nuevos paradigmas de interacción y de intercambio con los proveedores de contenidos que el futuro nos depara seguro que tendrán en cuenta el uso de la tecnología de la síntesis de voz como herramienta muy válida para enfatizar o redundar en un mensaje más cercano al humano.

Para posibilitar la utilización de la síntesis del habla en contenidos audiovisuales hay que seguir avanzando en las siguientes líneas de investigación:

- Mejorar la expresividad del habla generada para adaptar los rasgos suprasegmentales (ritmo, entonación, intensidad, énfasis, etc.) a las características propias del modo de locución de cada tipo de contenido. Esta mejora puede conseguirse si se cuenta con la aportación de los conocimientos de expertos en el campo de la comunicación audiovisual.
- Mejorar la calidad segmental de la síntesis para evitar artefactos sonoros, ya que hay que tener en cuenta que el oído humano es muy sensible a estos pequeños errores. En este aspecto, influyen errores relacionados con la fonética y con el procesamiento de la señal. Por tanto, sería deseable contar con expertos en fonética que aportaran conocimiento para mejorar, por ejemplo, las reglas de transcripción fonética, especialmente las que hacen referencia a la coarticulación. En cuanto al procesamiento de la señal, hay camino por recorrer en la parametrización y el modelado de la voz para poder llevar a cabo modificaciones de sus características sin distorsionarla.
- Conseguir nuevos métodos para generar nuevas voces mediante técnicas de transformación de voz que permitan aumentar el número de voces de alta calidad disponibles en un idioma determinado.

Referencias

Alías, F.; Formiga, L.; Llorà, X. “Efficient and reliable perceptual weight tuning for unit-selection Text-to-Speech synthesis based on active interactive genetic algorithms: a proof-of-concept”. *Speech Communication,* vol. 53 (5), p. 786-800, mayo-junio, 2011.

Campbell, N. “Evaluation of Text and Speech Systems”. *Text, Speech and Language Technology*, vol. 37 (2007), p. 29-64, Springer, Dordrecht.

Camps, J.; Bailly, G.; Martí, J. “Synthèse à partir du texte pour le catalan”. *Proceedings of 19èmes Journeés d’Études sur la Parole* (1992), p. 329–333, Bruselas, Bélgica.

Guaus, R.; Iriondo, I. “Diphone based Unit Selection for Catalan Text-to-Speech Synthesis”. *Proceedings of Workshop on Text, Speech and Dialogue* (2000), Brno, República Checa.

Iriondo,I.; Alías, F.; Melenchón, J.; Llorca, M.A. “Modeling and Synthesizing Emotional Speech for Catalan Text-to-Speech Synthesis”. *Tutorial and Research Workshop on Affective Dialog Systems, Lecture Notes in Artificial Intelligence*, núm. 3068 (2004), Springer Verlag, p. 197-208, Kloster Irsee, Alemania.

Iriondo,I.; Socoró, J. C.; Alías, F. “Prosody Modelling of Spanish for Expressive Speech Synthesis”. *International Conference on Acoustics, Speech and Signal Processing*, vol. IV (2007), p. 821-824, Hawaii, Estados Unidos.

Llisterri, J.; Fernández, N.; Gudayol, F.; Poyatos, J. J.; Martí, J. “Testing user’s acceptance of Ciber232, a text to speech system used by blind persons”. *Proceedings of the ESCA Workshop on Speech and Language Technology for Disabled Persons* (1993), p. 203–206, Estocolmo, Suecia.

Planet, S.; Iriondo, I.;Martínez, E.; Montero, J.A. “TRUE: an online testing platform for multimedia evaluation”. Proceedings of the Second *International Workshop on Emotion: Corpora for Research on Emotion and Affect at the 6th Conference on Language Resources & Evaluation* (2008), Marrakech, Marruecos.

Rodríguez, M.A.; Escalada, J. G.; Armenta, A.; Garrido, J.M. “Nuevo módulo de análisis prosódico del conversor texto-voz multilingüe de Telefónica I+D”. *Actas de las V Jornadas en Tecnología del Habla* (2008), p. 157-160.

Torrens, A.”Estudi sobre la utilització de les tecnologies de síntesi de veu en els mitjans audiovisuals de Catalunya”. *Treball final de carrera* (2010). Barcelona: La Salle - Universitat Ramon Llull.

UIT-T (1996). “Recomendación P.800: Métodos de determinación subjetiva de la calidad de transmisión*”. Sector de Normalización de las Telecomunicaciones de Unión Internacional de Telecomunicaciones*.
<http://www.itu.int/rec/T-REC-P.800-199608-I/es>