Modelado lingüístico y acústico para un sistema de conversión de texto a habla

E. López Gonzalo**, E. Rodríguez Banga*,
C. García Mateo*, Luis A. Hernández Gómez**

(*)DTC - ETSI Telecomunicación. Universidad de Vigo. As Lagoas-Marcosende.
36200 VIGO. e-mail : erbanga@dtc.uvigo.es , carmen@dtc.uvigo.es

(**)DSSR - ETSI Telecomunicación. UPM. Ciudad Universitaria.
28040 MADRID. e-mail: eduardo@gaps.ssr.upm.es , luis@gaps.ssr.upm.es

Resumen

En esta comunicación se presenta un sistema basado en la concatenación de difonos que utiliza nuevos algoritmos de procesado lingüístico y acústico con objeto de mejorar la inteligibilidad y naturalidad de sistemas precedentes. A partir del texto de entrada, un módulo lingüístico-prosódico obtiene la transcripción fonética y un conjunto de marcas prosódicas que reflejan su estructura sintáctica y rítmica. El procesado acústico está basado en la concatenación de difonos, utilizando para ello un codificador armónico multibanda que permite las modificaciones prosódicas de manera sencilla y eficiente, a la vez que proporciona una señal sintética de buena calidad.

Tema: Síntesis del habla

1 Introducción

Actualmente, el objetivo central en los sistemas de conversión de texto a habla, es generar habla con un grado de inteligibilidad y naturalidad lo más elevado posible. Para ello, es clave un buen modelado tanto lingüístico (en sus aspectos de procesado del texto y generación prosódica) como acústico. En la figura 1, se muestra la división en dos bloques que se suele realizar en estos sistemas. El primer bloque, trata de generar a partir del texto una representación fonética-prosódica del mismo. La representación fonética hace referencia a aquellos elementos segmentales que permiten al procesador acústico concatenar las unidades acústicas adecuadas para producir los sonidos que configuran un mensaje, mientras que la representación prosódica configura los elementos suprasegmentales: duración de cada sonido y evolución temporal de la frecuencia fundamental y energía a lo largo de la elocución. El procesador acústico transforma esta representación en la señal representativa del habla. En este artículo se van a presentar los esfuerzos conjuntos que el Dpto. de Señales Sistemas y Radiocomunicaciones de Madrid y el Dpto. de Tecnologías de las Comunicaciones de Vigo para el desarrollo de un sistema de conversión de texto a habla en castellano.

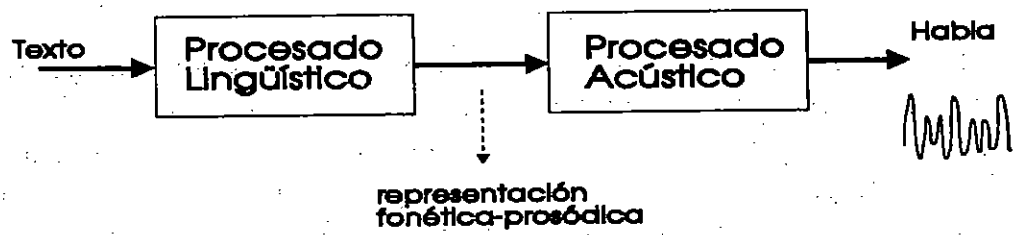

Figura 1: Sistema de conversión de texto a habla

A continuación se describen las distintas partes de nuestro sistema. En la sección 2) se describe el procesado lingüístico, describiéndose las etapas de análisis de texto, transcripción fonético-prosódica y acceso a la base de datos de parámetros prosódicos. Posteriormente, en la seccion 3, se trata el procesado acústico, partiendo de las bases del modelado armónico para llegar, finalmente, a su aplicación en un sistema de conversión de texto a habla. Finalmente, se presentan unas breves conclusiones.

2 Procesado lingüístico

El procesado lingüístico es clave para generar voz de alta calidad. Dos son los objetivos, como dijimos, para este módulo. Mientras que el cómputo de la representación fonética

partir del texto, es comparativamente más fácil para el castellano que para otros idiomas, el éxito en el modelado prosódico influirá directamente en la naturalidad de la voz sintética [1].

Nuestro módulo de procesado lingüístico se puede dividir en tres partes diferenciadas:

- Análisis del texto.
- Resolución de la transcripción fonético-prosódica.
- Acceso a la Base de Datos de parámetros prosódicos.

2.1 Análisis del texto

El análisis del texto tiene como entrada el texto a sintetizar. Su primera tarea es delimitar dentro de él una unidad de síntesis para ser procesada por el resto de los módulos. Aunque generalmente se admite que puede haber factores dentro de un párrafo que afecten la prosodia de una frase del mismo, éstos no son considerados en nuestro sistema, por lo que la unidad de síntesis elegida se corresponde por lo que habitualmente se entiende por frase. Nuestro módulo detecta una frase, detectando los signos ortográficos "?", "!", "...", ";" y el punto final de frase. Este último es el más difícil de reconocer debido a que se usa también para marcar abreviaturas y cifras numéricas. En nuestro sistema se detectan basándonos en un autómata finito reconocedor de una secuencia regular descrita en [2] que presenta un buen comportamiento en los casos sin errores tipográficos.

Una vez determinada la unidad de síntesis, los sistemas de síntesis proceden a realizar un preprocesado de la misma. El preprocesador reconoce y aisla las diferentes palabras que integran la unidad; los signos ortográficos se consideran palabras ya que juegan un papel importante en el marcado prosódico. Se normaliza también la escritura de las palabras para reducir la variabilidad tipográfica que se presenta al resto de los módulos. También se encarga de poner en letra los guarismos numéricos y de transformar ciertas palabras contenidas en una tabla, como por ejemplo algunas abreviaturas. Además, ayuda al resto de los módulos en los casos detectables a este nivel para determinar la categoría gramatical de cada palabra. No nos vamos a detener en la descripción detallada del mismo ya que es bastante parecido al descrito en [3] y el lector puede encontrar allí más detalles.

El siguiente módulo realiza para cada palabra determinada anteriormente un análisis morfológico. Para ello, se dispone de un léxico de palabras pertenecientes a series cerradas (preposiciones, conjunciones,...), y de una tabla de sufijos y raíces útiles para determinar la categoría gramatical de las palabras pertenencientes a series abiertas. La salida de este módulo es una lista de categorías para cada palabra, con una descripción de su género,

número y persona. Las categorías producidas por el módulo, no se corresponden exáctamente con las categorías que se encuentran en las gramáticas tradicionales, sino que se eligen atendiendo a la utilidad en el marcado prosódico. Por ejemplo, la preposición "de" no tiene un tratamiento idéntico al resto de las preposiciones.

Una misma palabra puede pertenecer a varias categorías, por lo que el siguiente módulo trata de determinar la categoría léxica de cada palabra mediante reglas de contexto gramatical donde se analizan secuencias de categorías gramaticales. El funcionamiento de este módulo es importante para, por ejemplo, determinar la correcta acentuación de algunas palabras [4]. Este módulo, en definitiva, trata de modelar de alguna manera las restricciones sintácticas entre palabras al constituir una frase gramatical. A la salida del módulo, cada palabra tiene asignada una categoría única. Aunque el funcionamiento del módulo es bastante bueno (más del 90% de los casos se detectan correctamente), se hace patente el hecho de mejorar este módulo en base a restricciones sintácticas más complejas.

Una vez realizado el análisis contextual, se trata de realizar un análisis de la estructura de la oración. Este módulo actualmente es relativamente simple, comparado con lo que representaría un análisis sintáctico formal y está inspirado en el trabajo presentado en [5]. Se consideran tanto factores rítmicos como sintácticos al establecer un coeficiente entre dos palabras sucesivas. Este coeficiente representa el grado de dependencia o coherencia entre esas dos palabras. A alteraciones en el orden normal de las palabras se les asigna un coeficiente alto y a palabras conectadas íntimimamente un coeficiente bajo. Estos coeficientes se han diseñado manualmente atendiendo a nuestras exigencias de marcado prosódico. Aunque este módulo suele recoger las propiedades sintácticas más relevantes, no deja de ser una aproximación heurística al análisis sintáctico, y por tanto, presenta dificultades para reconocer en la frase estructuras sintácticas más complejas, que son importantes para la determinación de la prosodia. Así, las estructuras enumerativas (por ejemplo, la frase que está leyendo), las estructuras comparativas (como por ejemplo: "Hay que cuidar, tanto los aspectos prácticos, como los teóricos"), la inversión en el orden normal de la frase y la detección de elipsis, son difíciles de detectar partiendo de esta aproximación local.

Por todo ello, actualmente estamos abordando una solución más general basada en un análisis sintáctico más riguroso de la frase de entrada. El modelo se basa en la teoría sintáctica ideada por Vergne [6]. Esta teoría, gira en torno al concepto de bloque y la construcción gramatical por concatenación e inserción de bloques. Fruto de una colaboración, se ha determinado una relación bastante estrecha entre el marcado prosódico y las dependencias entre bloques, en base a cuatro reglas sencillas [7]. Esto parece poner de manifiesto que la idea de bloque juega un papel fundamental en el habla. Actualmente, estamos realizando una versión del analizador sintáctico-prosódico basado en esta teoría

ptándola al caso del español. Por otro lado, este análisis sintáctico no es costoso com-acionalmente y por tanto se adapta bastante bien a la problemática de la conversión to a habla.

2 Resolución de la transcripción fonético-prosódica

este apartado se analizan los módulos, que partiendo de la representación morfo-ntáctica determinan la transcripción fonético-prosódica de cada unidad de síntesis. Pa-e que en la estructura rítmica del español entra en juego de una manera determinante unidad denominada sílaba. Así lo demuestran, por poner un ejemplo sencillo, las reglas acentuación del castellano. Por tanto, el primer módulo realiza una descomposición lábica de la frase de entrada, basándose en el análisis dado en [8].

A continuación, se determina la posición (en caso de que exista) del acento de cada palabra (o de los dos acentos, en el caso de los adverbios derivados de adjetivos por el sufijo "-mente"), basándonos en las conocidas reglas de acentuación del español.

Con la información disponible en este momento ya podemos determinar las posiciones que van a delimitar grupos melódicos [8] [9]. Estos grupos melódicos suelen coincidir con los grupos fónicos (el grupo fónico es la unidad delimitada por dos pausas consecutivas en la elocución). El grupo melódico, además de por pausas, puede ir marcado por depresiones de intensidad, retraso de la articulación y cambio brusco en la altura musical. Observamos que un concepto abstracto como grupo melódico, puede ser realizado acústicamente por varias variables interdependientes, lo que complica su estudio. En nuestro sintetizador los grupos melódicos son siempre marcados por pausa (aunque puede ser muy breve). La determinación de grupos melódicos se basa sobre todo en los coeficientes de relación sintáctica asignados anteriormente. Aquellas palabras representadas por coeficientes que superan un determinado umbral, son consideradas palabras frontera de grupos melódicos. Después se analiza el número de sílabas de cada grupo melódico, insertando un nuevo grupo, si alguno sobrepasa una longitud de 16 sílabas. La marca delimitadora de grupo en este caso se introduce donde se encuentre el coeficiente más alto marcando aquellas palabras más independientes, dentro de este grupo melódico largo. Este proceso se repite hasta que todos los grupos melódicos tengan una longitud inferior o igual a 16 sílabas.

La realización acústica de cada grupo melódico mediante las variables mencionadas anteriormente depende del tipo de grupo melódico [9]. Nosotros hemos definido un catálogo de nueve tipos de grupos melódicos:

1. Grupo terminativo: Aparece al final de las oraciones enunciativas con un tonema descendente. Por ejemplo: "Cogió el tren de las cinco y media".

2. Grupo continuativo: Aparece en la o las proposiciones no finales de las oraciones compuestas, ya sea por coordinación o subordinación. Su tonema es ascendente. Por ejemplo: "Dijo que le dolía la cabeza y le llevamos al médico", "Si quieres comer, ven antes de las dos".

3. Grupo contrastivo. Igual que la anterior pero con una fución de contraste con respecto a ella, de manera que el tonema es ligeramente descendente.

4. Grupo incidental: Aparece cuando dentro de una oración afirmativa se realiza un inciso que amplía el conocimiento de un hecho sin ser completamente necesario para la determinación del mismo (no aumenta su comprensión). Ejemplo "Llegaron las mujeres, que estaban cansadas".

5. Grupo vocativo: Es un medio que sirve para llamar a una persona. Expresar el fin elocutivo sin el uso del verbo. Ejemplo "Juan, ven aquí".

6. Grupo parentético: La entonación de los paréntesis sirve para desligarlo del resto del discurso. Ejemplo "Quiero más (dijo Juan en voz baja) y María le oyó".

7. Grupo yuxtapuesto: Aparece en aposiciones, oraciones compuestas por yuxtaposición y en algunas ocasiones por coordinación adversativa.

8. Grupo apelativo: Aparece en la proposición final de las interrogativas, ó en la inicial de las interrogativas optativas.

9. Grupo apelativo pronominal: Aparece en las preguntas con pronombre interrogativo.

La determinación del tipo de grupo melódico se realiza mediante un análisis contextual de los coeficientes de relación, que codifican en parte la relación sintáctica entre dos palabras sucesivas.

Además del grupo melódico, se considera en nuestro sintetizador otra unidad prosódica más pequeña que es el grupo acentual, es decir aquella secuencia de palabras sin acento que finaliza en una palabra con acento léxico. Los grupos melódicos se subdividen en grupos de acento inicial, medio y final. Esta división tiene como objetivo motivos prácticos de implementación del procesado prosódico como veremos después.

Una vez realizado el marcado prosódico, realizamos la transcripción fonética. Esta se lleva a cabo en dos fases, primero se realiza una transcripción de letras a fonemas y luego de éstos a los alófonos considerados. Estas dos tareas se realizan después de la determinación de los grupos melódicos (que en nuestro caso serán siempre grupos fónicos) ya que necesitamos saber dónde están situadas las pausas.

2.3 Acceso a la base de datos de parámetros prosódicos

La información obtenida en todos los módulos anteriores se usa para acceder a una base de datos prosódica que modela la duración y el contorno de frecuencia fundamental de cada vocal (con unos valores normalizados), además de la duración de la pausa (al final de los grupos melódicos). Después se modelan las variables prosódicas de las consonantes y se desnormalizan los valores anteriores teniendo en cuenta el contexto segmental. Se realiza así un modelado conjunto de la duración y la frecuencia fundamental para modelar sus posibles interacciones, la energía por el momento no es modelada, por ser la menos importante desde un punto de vista perceptual. Los factores considerados para acceder a la base de datos son:

1. El tipo de grupo melódico (nueve casos posibles).
2. La posición del grupo acentual dentro del grupo melódico (inicial, medio o final).
3. El tipo de grupo acentual según la posición del acento.
4. La situación de la vocal dentro del grupo acentual.

Esta aproximación contrasta con otras basadas en reglas, pero nosotros somos de la opinión de que la nuestra permite capturar implícitamente rasgos prosódicos de un locutor. Para ello nuestro objetivo fue modelar la prosodia de un locutor para transplantarla al sistema de síntesis. Este objetivo es mucho menos ambicioso que modelar la prosodia del castellano que requeriría otra metodología, pero es suficiente para la aplicación que nos ocupa.

Para esta tarea, diseñamos un corpus especial teniendo en cuenta los tipos de grupos melódicos y acentuales que se consideran. Este corpus es muy amplio y para realizar un modelado del mismo se debe ajustar a la aplicación del conversor . Para limitar más el número de oraciones, nosotros tomamos en cuenta fundamentalmente la posición del primer y último acento dentro del grupo melódico y estudiamos principalmente relaciones entre grupos melódicos en oraciones con dos grupos melódicos. La metodología seguida está inspirada en el trabajo desarrollado en [10].

Después de grabar el corpus se analizan con herramientas diseñadas a tal efecto, el contorno de frecuencia fundamental (para las vocales) y la duración de cada sonido. Los contornos de frecuencia fundamental se deben "estilizar" [11] para conservar únicamente aquellos movimientos melódicos que son percibidos por un oyente medio. En vista de las curvas decidimos describir estos movimientos con los siguientes atributos:

1. La dirección del movimiento de frecuencia fundamental (subida o bajada).

2. La pendiente del movimiento en semitonos por milisegundo.

3. Extensión en el tiempo y alineado con respecto a las vocales tónicas.

Se desarrolló una herramienta gráfica para la estilización de contornos. La herramienta muestra la evolución temporal de la frecuencia fundamental en escala semilogarítmica, permitiendo su aproximación por tramos rectos. El módulo acústico descrito en este mismo artículo se usó como codificador de voz para producir el corpus prosódico imponiendo el nuevo contorno estilizado. Si se percibía alguna diferencia prosódica se usaban más puntos de interpolación.

Los resultados del proceso de estilización se analizan estadísticamente para almacenarlos en la base de datos. Este proceso se valida por la consistencia presentada en los contornos de una misma clase.

3 Procesado acústico

La etapa de procesado acústico es la encargada de generar la señal de voz sintética a partir de la información segmental (fonemas) y suprasegmental (frecuencia fundamental, duración y energía). Su función es, por tanto, seleccionar las diferentes unidades de la base de datos que se van a concatenar y ajustar su prosodia a la obtenida por el módulo de procesado lingüístico.

En los últimos años, se han desarrollado varios algoritmos de modificación prosódica que tratan de mejorar la calidad de la voz sintética. La mayor parte de estos algoritmos están basados en técnicas de "Overlap-Add" síncronas con la frecuencia fundamental (PSOLA), existiendo diversas variantes según se trabaje en el domino del tiempo (TD-PSOLA) o en el de la frecuencia (FD-PSOLA) [12]. Esta última variante proporciona una forma más flexible de modificar las características espectrales de la señal de voz que su equivalente versión temporal, evitando, asimismo, gran parte de la distorsión introducida por ésta. Sin embargo, hasta la fecha, ningún algoritmo de modificación prosódica en el dominio de la frecuencia ha sido propuesto como una solución eficiente a la hora de realizar un sistema de síntesis en tiempo real.

Junto con el algoritmo de modificación prosódica, un esquema de síntesis basado en la concatenación de unidades de voz requiere una extensa base de datos. En aplicaciones con restricciones de espacio de memoria es preciso codificar las unidades antes de almacenarlas. En este caso, el módulo de procesado acústico realiza tres tareas:

- Decodificación. Las unidades de la base de datos son reconstruidas a partir de sus versiones parametrizadas.

- Modificación prosódica. La prosodia de cada segmento de voz se ajusta a los valores determinados por el módulo de procesado lingüístico.
- Concatenación de los diferentes segmentos de voz para obtener la señal sintética.

Algunos algoritmos, como el LP-PSOLA, llevan a término estas tres tareas en tres etapas diferenciadas. Cabe, pues, plantearse la posibilidad de desarrollar nuevos algoritmos que permitan realizar la decodificación, las modificaciones prosódicas y la concatenación de unidades a un mismo tiempo. El tradicional vocoder LPC proporciona una forma flexible y eficiente para integrar estas tres tareas en una sola etapa, pero a expensas de una calidad de voz muy limitada. Por el contrario, los codificadores de voz basados en un modelado armónico proporcionan una voz sintética de muy buena calidad, especialmente si entre los armónicos de la frecuencia fundamental se mantienen las diferencias relativas de fase. Además, este esquema permite la utilización de múltiples decisiones sordas/sonoras, lo que conduce a una mejor representación de la señal de excitación, con la consiguiente mejora en la calidad de la voz sintética.

Basándonos en un modelado armónico de la señal de voz, proponemos una nueva familia de algoritmos de modificación prosódica en el dominio de la frecuencia. En la actualidad, hemos desarrollado dos algoritmos diferentes. El primero está orientado a aplicaciones que requieren un almacenamiento muy eficiente de la base de datos y proporciona una voz sintética de buena calidad. El segundo requiere alguna memoria adicional pero, como contrapartida, proporciona una señal sintética de alta calidad.

3.1 Bases del modelado armónico

El modelado armónico no es más que un caso particular del modelado sinusoidal [13]. Este último asume que un tramo estacionario de una señal de voz se puede representar como una suma de cosenos cuya frecuencias, amplitudes y fases vienen determinadas por el espectro del tramo de voz original, es decir

$$\hat{s}(n) = \sum_{l=1}^{L} A_l cos(\omega_l n + \phi_l) \tag{1}$$

donde ω_l representa a las frecuencias de los picos espectrales del tramo de voz original, A_l y ϕ_l sus correspondientes amplitudes y fases respectivamente, y L el número de picos considerados.

Debido a que este método es poco eficiente (ya que necesita hallar todos los picos espectrales y almacenar el valor de sus correspondientes frecuencias), se comienzan a

introducir restricciones. Basándose en que en un tramo sonoro los picos espectrales están armónicamente relacionados (ver figura 2), se introduce la llamada restricción armónica. El modelo se reduce entonces a:

$$\hat{s}(n) = \sum_{l=1}^{L} A_l cos(l\omega_0 n + \phi_l) \quad (2)$$

donde ω_0 representa a la denominada frecuencia fundamental.

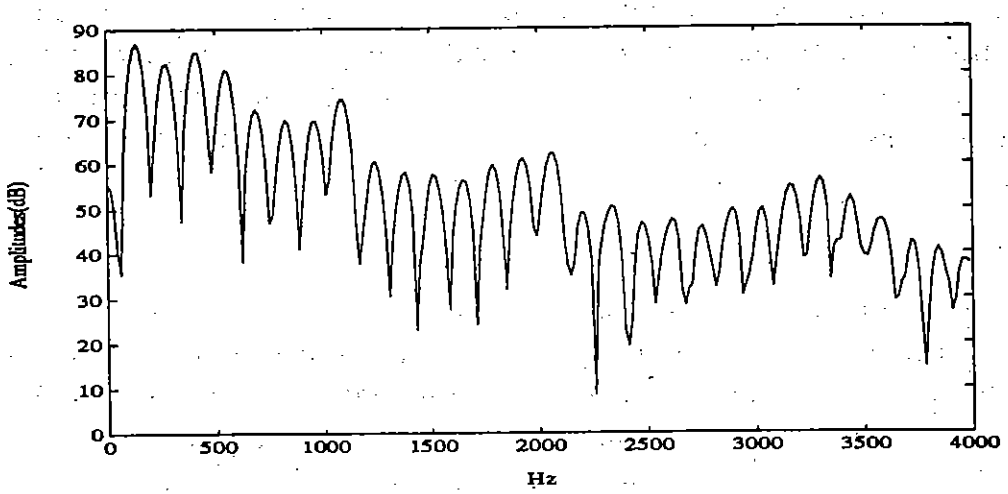

Figura 2: Espectro de un tramo de voz sonoro.

Este modelo que, en principio, sólo parece válido para segmentos sonoros de la señal de voz, tambien se puede aplicar a segmentos sordos si las frecuencias de las sinusoides están lo bastante próximas entre sí. No obstante, este esquema adolece de una limitación importante: clasifica cada tramo de voz como sordo o sonoro. Como consecuencia, algunos sonidos y, en especial, los tramos de transición no son bien representados. Para solucionar este problema aparecen nuevos modelos que consideran a la señal de voz como suma de una componente sonora y de una componente sorda.

El codificador MBE (Multiband Excitation) realiza múltiples decisiones sordas/sonoras por tramo de voz [14]. Para ello realiza una aproximación periódica del espectro de la señal de voz (que es prácticamente idéntica al espectro de $\hat{s}(n)$ en (2)) y la compara con el espectro original. Las zonas (bandas) de frecuencia donde la aproximación se ajusta bien al espectro original son declaradas sonoras y las restantes, sordas (ver figura 3).

En la práctica, se define una banda de frecuencia como un conjunto de tres armónicos consecutivos. El número máximo de armónicos considerado en [15] es 36, por tanto el

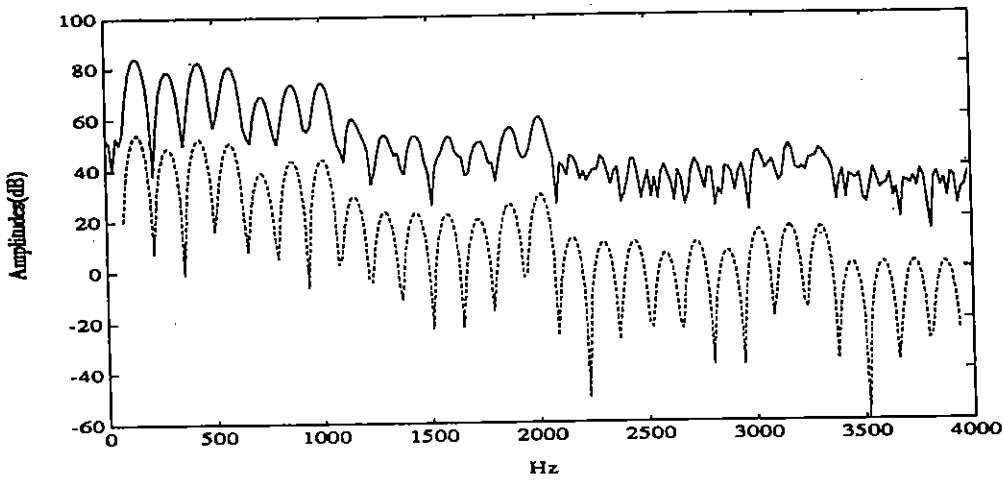

Figura 3: Espectro de un tramo de voz y la aproximación periódica. Esta última ha sido desplazada para permitir una mejor comparación.

número máximo de bandas es 12. Para cada segmento de 20 ms de la señal de voz se obtienen la frecuencia fundamental, un conjunto de decisiones sordas/sonoras y las amplitudes de los armónicos considerados.

Para reconstruir la señal de voz, la componente sorda se genera como transformada inversa de Fourier de un espectro que es nulo en las bandas sonoras y ruidoso en las bandas sordas. La generación de la componente sonora se realiza mediante:

$$s_{son}(n) = \sum_{l=1}^{L} A_l(n)\ cos(\theta_l(n)) \tag{3}$$

donde L es el número de armónicos declarados sonoros, $A_l(n)$ representa la evolución de las amplitudes de los armónicos y $\theta_l(n)$ es la función de fase instantánea.

Este modelado mediante suma de armónicos de diferentes amplitudes y fases plantea un importante problema: mantener la continuidad de fases entre tramas adyacentes. Las componentes sonoras de dos tramas sintéticas consecutivas presentan, en general, una discontinuidad en la frontera entre ambas. Estas discontinuidades, por pequeñas que sean, son percibidas por el oído en forma de "clicks" muy molestos. Para evitar este problema se debe asegurar un cambio gradual de las amplitudes y las fases entre tramas adyacentes, utilizándose técnicas de interpolación para determinar la evolución de $A_l(n)$ y de $\theta_l(n)$.

3.2 Modificación prosódica mediante modelado armónico

Utilizando un modelo armónico se pueden variar fácilmente los parámetros suprasegmentales de la señal de voz (frecuencia fundamental, duración, energía). Así, modificar la frecuencia fundamental se reduce a estimar las amplitudes espectrales de los nuevos armónicos. La duración es modificada mediante un escalado temporal de la fase instantánea y de la amplitud de los diferentes armónicos, lo que resulta en que la longitud de los tramos sintetizados es variable. Por otro lado, la energía se puede controlar mediante una ganancia variable actuando sobre la señal de voz sintética.

3.3 Procesado de la base de datos

La base de datos de unidades elementales de voz (difonos en nuestro caso) es parametrizada utlizando un codificador MBE o una variante que hemos desarrollado para conseguir un almacenamiento más eficiente (MBE-LPC) [16]. Este último obtiene para cada tramo de la señal de voz la frecuencia fundamental, el conjunto de decisiones sordas/sonoras y los coeficientes de predicción lineal de un modelo autorregresivo. Como el número de coeficientes utilizados es 10 y estos pueden ser cuantificados vectorialmente de forma muy eficiente, este método representa un gran ahorro de memoria en comparación con el algoritmo original. Por supuesto, la calidad de la voz sintetizada sufre una cierta degradación, pero sigue siendo muy aceptable.

Una vez que toda la base de datos ha sido parametrizada, a cada unidad elemental se le asigna un conjunto de parámetros (que se corresponden con varios segmentos de la señal de voz original) y a cada sonido perteneciente a la unidad elemental un subconjunto diferente. De esta forma es posible variar la duración de los diferentes sonidos que componen la unidad elemental de manera prácticamente independiente.

Es de destacar que, a diferencia de las técnicas PSOLA, el procedimiento de análisis utilizado no es síncrono con la frecuencia fundamental y, por tanto, no es necesario etiquetar todas las unidades de la base de datos con las denominadas "pitch marks", lo que conlleva un importante ahorro en el procesado de la base de datos.

3.4 Síntesis y algoritmos de modificación prosódica

Se han desarrollado dos algoritmos, basados en el modelo armónico, que realizan la decodificación, modificación prosódica y concatenación en una única etapa. La elección entre uno u otro depende del codificador utilizado en el procesado de la base de datos (MBE o MBE-LPC). Sus diferencias radican en la estimación de las amplitudes espectrales de los armónicos de la nueva frecuencia fundamental y en el algoritmo de fase utilizado para

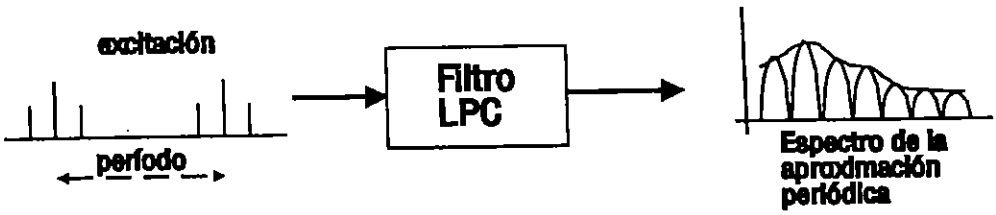

Figura 4: Aproximación periódica de la envolvente espectral

garantizar que no haya discontinuidades entre tramas.

Si la base de datos ha sido procesada utilizando un codificador MBE, las amplitudes espectrales de los nuevos armónicos son calculadas, mediante interpolación lineal, a partir de las amplitudes de los armónicos de la frecuencia fundamental original. El algoritmo de fase utilizado en este caso es el descrito en [15]. La generación de la componentes sonora y sorda es, básicamente, idéntica al algoritmo original del MBE (descrito anteriormente). No obstante, se debe ser cuidadoso al determinar sin un armónico pertenece a una banda sorda o sonora, puesto que, al variar la frecuencia fundamental, el número de armónicos en una determinada banda de frecuencia puede ser distinto de tres.

En caso de que la base de datos haya sido procesada utilizando el codificador MBE-LPC, el algoritmo de síntesis se vuelve un poco más complejo [17], debido a que es necesario estimar las amplitudes de los armónicos de la frecuencia fundamental a partir de la envolvente espectral (determinada por los coeficientes de predicción lineal). Para ello, se genera una señal de excitación periódica (con período el correspondiente a la nueva frecuencia fundamental) que sirve de entrada al filtro LPC. A la salida se obtiene una secuencia cuyo espectro es muy parecido a la aproximación periódica utilizada en la etapa de análisis del codificador MBE, con la diferencia de que los picos espectrales están situados en los armónicos de la nueva frecuencia fundamental (figura 4). A partir de este espectro se pueden estimar la amplitudes espectrales de los nuevos armónicos. Como algoritmo de fase se utiliza una interpolación polinómica de orden tres, de forma que la evolución de la fase instantánea sea lo más suave posible.

Los algoritmos de modificación prosódica presentados proporcionan una gran flexibilidad con una baja carga computacional, lo que los hace ideales para la realización de sistemas de conversión de texto a voz en tiempo real. La utilización de algoritmos de interpolación para determinar la evolución de la amplitud y la fase instantánea de los distintos armónicos, hace prácticamente innecesaria la utilización de otras técnicas para suavizar las transiciones entre diferentes unidades de la base de datos. Por otro lado, el

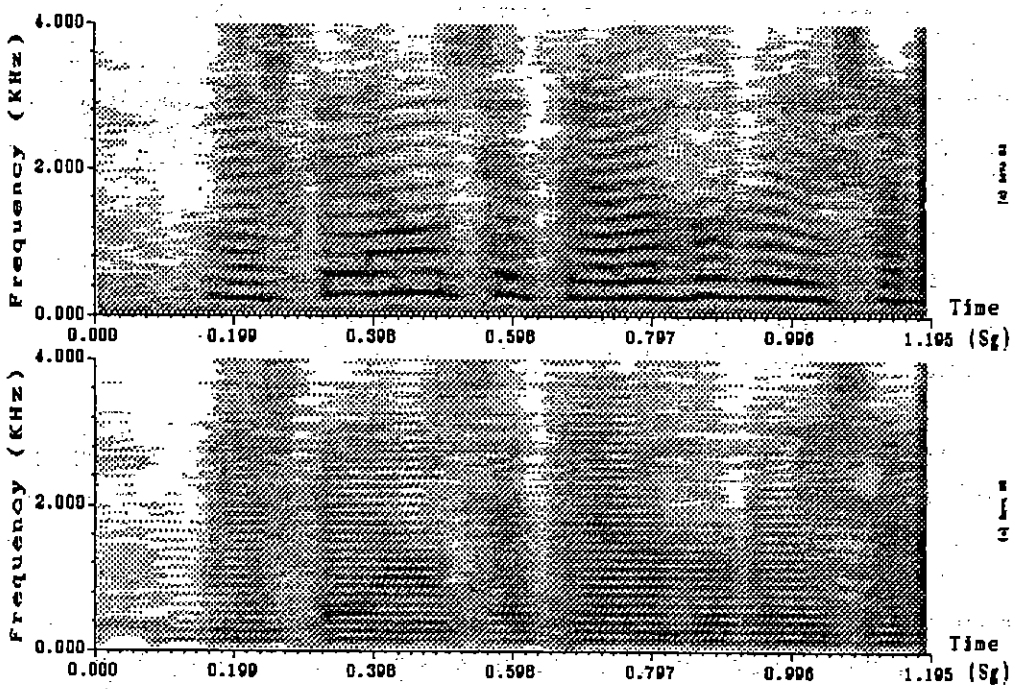

Figura 5: Espectrogramas de la señal original y de la sintética

algoritmo utilizado para la modificación de la duración permite acortar o alargar los diferentes sonidos sin introducir ninguna distorsión [18]. Además, la utilización de múltiples decisiones sordo/sonoro permite aumentar la frecuencia fundamental sin reducir el ancho de banda y sin trasladar propiedades locales del espectro de la señal de voz (tales como amplitudes de los armónicos y la energía de las diferentes bandas de frecuencia), como así ocurre en las técnicas FD-PSOLA. A modo de ejemplo, en la figura 5 se representa, arriba, el espectrograma de una señal de voz y, abajo, el correspondiente a la misma frase pero con una frecuencia fundamental muy baja. Se puede observar que, incluso en este caso extremo, el ancho de banda de la señal de voz no se reduce y no hay un desplazamiento de energías de alta a baja frecuencia.

4 Conclusiones

En este artículo hemos descrito un sistema de conversión de texto a habla basado en nuevos algoritmos de procesado lingüístico y de modificación prosódica. La calidad de habla obtenida con este modelo es buena, resultando muy inteligible. No obstante, estamos introduciendo nuevas modificaciones en los algoritmos que permitirán mejorar su naturalidad.

Referencias

[1] F. Emerard, L. Mortamet, y A. Cozannet, "Prosodic processing in a text-to-speech synthesis system using a database and learning procedures," en *Talking Machines: Theories, Models and Applications* (G. Bailly y E. Benoit, eds.), Elsevier, 1992.

[2] M. Y. Liberman y K. W. Church en *Advances in Speech Signal Processing* (S. Furui y M. M. Sondhi, eds.), cap. Text Analysis and Word Pronuntiation, New York: Marcel Dekker, Inc., 1992.

[3] M. A. Rodríguez, J. G. Escalada, y A. Macarrón, "Amigo: Un conversor texto-voz para español," en *Jornadas de la Sociedad Española para el Procesado del Lenguaje Natural*, (Granada), 1992.

[4] A. Quilis, "Fonética acústica de la lengua española," en *Biblioteca románica hispánica, Manuales, 49*, Gredos, 1991.

[5] G. Bailly, "Integration of rhythmic and syntatic constraints in a model of generation of french prosody," *Speech Communication*, vol. 8, pag. 137–146, 1989.

[6] J. Vergne, "Syntax as clipping blocks: structures, algorithms and rules," en *Jornadas de la Sociedad Española para el Procesado del Lenguaje Natural (SEPLN)*, (Granada), 1992.

[7] J. Vergne y E. López-Gonzalo, "Comunicaciones privadas," Octubre 1992-Marzo 1993.

[8] R. A. Española, ed., *Esbozo de una nueva gramática de la lengua española.* Espasa Calpe, 1973.

[9] T.Ñavarro-Tomás, *Manual de entonación española.* New York: Hispanic Institute. Madrid. Guadarrama, 4 ed., 1974.

[10] V. Auberge, "Semi-automatic constitution of a prosodic contour lexicons for the text-to-speech synthesis," en *Proceedings ESCA Workshop on Speech Synthesis*, (Autrans), 1990.

[11] R. Collier, "Multi-lingual intonation synthesis: principles and applications," en *Proceedings ESCA Workshop on Speech Synthesis*, (Autrans), 1990.

[12] F. Charpentier y E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *2nd European Conference ond Speech Communication and Technology. EUROSPEECH'89*, pag. 13–19, 1989.

[13] R. J. McAulay y T. F. Quatieri, "Low-rate speech coding based on the sinusoidal model," en *Advances in Speech Signal Processing* (S. Furui y M. M. Sondhi, eds.), pag. 165–208, N.Y.: Marcel Dekker, 1992.

[14] D. W. Griffin y J. S. Lim, "Multiband excitation vocoder," en *IEEE Trans. Acoust., Speech, Signal Process. ASSP-36*, pag. 1223–1235.

[15] "INMARSAT-M voice coding system description," *DRAFT version 1.3*, 1991.

[16] C. Garcia, E. R. Banga, J. L. Alba, y L. Hernandez, "Analysis, synthesis and quantization procedures for a 2.5 kbps voice coder obtained by combining lp coding and harmonic coding," *European Signal Processing Conference. EUSIPCO'92*, vol. 1, pag. 471–474, 1992.

[17] E. R. Banga, E. López-Gonzalo, y C. García-Mateo, "A text-to-speech system for Spanish with a frequency domain based prosodic modification algorithm," en *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, (Minnesota(USA)), pag. 183–186, 1993.

[18] E. R. Banga y C. Garcia-Mateo, "New frequency domain prosodic modification techniques," en *3rd European Conference on Speech Communication and Technology. EUROSPEECH'93*, (Berlin), 1993.