3. Reconocimiento y síntesis de habla

Conversión de texto a voz en castellano aplicando el algoritmo PSOLA.

José Luis Bullón y Juan Carlos Pérez.
Dpto. de Sistemas Informáticos y Computación (DSIC).
Universidad Politécnica de Valencia.

Resumen

Se presenta un conversor de texto a voz para el castellano basado en la concatenación de unidades fonéticas previamente codificadas, que puede funcionar en un ordenador compatible PC sin más material adicional que un conversor digital-analógico. La principal innovación introducida en este sintetizador es la utilización del método **PSOLA** (de *Pitch Syncronous Overlap-Add*) para controlar el tono y la duración de las unidades fonéticas en el momento de su emisión. Este método, desarrollado recientemente, permite modificar la frecuencia fundamental de una señal de un modo sencillo actuando directamente sobre la representación de la misma en el dominio del tiempo, pudiendo utilizarse por tanto para formar la curva prosódica de los mensajes a emitir. Esta técnica ya había sido aplicada en la síntesis de voz para otros idiomas como el francés o el italiano, obteniéndose una naturalidad de la voz superior a la de otros sistemas similares. El sintetizador aquí desarrollado proporciona una primera idea de los resultados que se pueden obtener con el PSOLA en la síntesis de voz en nuestro idioma.

1. Introducción

Los métodos de síntesis de voz utilizados actualmente se pueden dividir principalmente en dos grupos [Moulines,90a]: la síntesis por reglas y la basada en la concatenación de unidades previamente almacenadas. En el primer método se intenta conseguir una reproducción de la voz humana a partir de un estudio de determinadas características de ésta y mediante una serie de reglas que se aplican a ciertas fuentes de señal sonora a lo largo del tiempo [Llisterri,89]. En cuanto a las técnicas de síntesis por concatenación de unidades fonéticas, surgen históricamente como un intento de reducir la complejidad (sobre todo en cuanto a los cálculos necesarios) que presentan los sistemas de síntesis por reglas. Se basan en el almacenamiento de segmentos de voz que posteriormente son concatenados para producir frases de cualquier longitud. Los factores a tener en cuenta con este tipo de síntesis son los siguientes:

-Elección de unidades adecuadas (segmentos de voz) a ser almacenadas.

-Selección de una técnica de codificación de esas unidades.

-Empleo de un método que permita modificar los parámetros prosódicos de los segmentos sin que se degrade la calidad de la voz.

A la hora de seleccionar el tipo de unidad habrá que buscar un compromiso entre

minimizar la cantidad de memoria necesaria para su almacenamiento, por un lado, o bien reducir al máximo el problema de la coarticulación por otro. Las posibilidades existentes abarcan desde el empleo de los fonemas, en cuyo caso es necesario disponer de un completo conjunto de reglas, al de las frases completas, cuya utilidad se reduce a aquellos casos en que el conjunto de mensajes a emitir es limitado. Unidades muy utilizadas son los difonemas y las semisílabas [Llisterri,89].

Otro aspecto a considerar es la elección de una técnica de codificación que nos permita la reconstrucción de la onda sonora a partir de las unidades almacenadas. Suelen emplearse, entre otras, técnicas en el dominio de la frecuencia, técnicas en el dominio del tiempo y codificación predictiva lineal (LPC) [Martí,86].

Por último, como hemos comentado, es necesario disponer de un mecanismo que permita modificar la frecuencia fundamental para poder dar al mensaje la curva prosódica adecuada. Este proceso resulta sencillo si la técnica de codificación empleada es LPC o basada en el dominio de la frecuencia. La modificación de la frecuencia fundamental de una señal en el dominio del tiempo, sin embargo, presenta dificultades si se pretende conservar su envolvente espectral.

Una alternativa, introducida en [Pérez,89] y [Pérez,91], es la de tomar muestras de voz a diferentes frecuencias y emitir una u otra según el tono deseado de esa muestra dentro de la frase. Este método presenta el inconveniente de que, al ser mayor el número de unidades a almacenar, se requiere una cantidad de memoria más elevada. En el citado trabajo, esto se resuelve tomando a distintas frecuencias únicamente las muestras correspondientes a las vocales y controlando con ellas el tono de toda la frase. Aunque el resultado obtenido en cuanto a la relación entre calidad y memoria requerida es más que aceptable, el sistema construido presenta ciertas deficiencias en cuanto a su naturalidad que convendría mejorar.

En los últimos años se ha introducido un nuevo método que actúa en el dominio del tiempo y que combina la rapidez propia de este tipo de técnicas con la calidad de otras más sofisticadas. Este algoritmo se conoce con el nombre de TD-PSOLA [Moulines,90b], y en el sistema descrito en el presente artículo se ha podido comprobar que efectivamente proporciona una calidad elevada con unos requerimientos de "hardware" muy bajos.

2. Descripción general del sistema

El sistema de conversión de texto a voz desarrollado se basa en la concatenación de unidades previamente almacenadas. El tipo de unidades empleadas se encuadra dentro de lo que algunos autores denominan **semisílabas**, que tienen un tamaño intermedio entre los fonemas y los difonemas. En nuestro caso concreto, se almacena cada fonema consonántico con un pequeño fragmento de transición a cada una de las vocales. Con esto se resuelve parcialmente el problema de la coarticulación, puesto que la concatenación se produce en un punto suficientemente estable de la señal. Además de

éstas, como veremos a continuación, se almacenan otros dos tipos de muestras, codificadas todas ellas mediante la técnica PCM (*Pulse Coded Modulation*).

2.1. UNIDADES UTILIZADAS

El conjunto de unidades seleccionadas se puede dividir en los siguientes tres grupos:

- **Unidades consonánticas**: constan de cada fonema consonántico con un pequeño fragmento de vocal. El fragmento de fonema vocálico debe ir tanto antes como después del consonántico, dando lugar a una unidad distinta en cada caso. Los fonemas para los que se han obtenido estas unidades son los siguientes: /b/, /d/, /g/, /k/, /l/, //, /m/, /n/, //, /p/, /r/, /t/, /y/.

- **Unidades vocálicas**: este segundo grupo de unidades lo constituyen, por una parte, las cinco vocales del castellano aisladas, y, por otra, unos fragmentos correspondientes a las transiciones que se pueden dar entre las mismas. Éstas unidades permiten suavizar los puntos de concatenación de las vocales. El conjunto completo de unidades dentro de este grupo es, pues, el siguiente:

* a, e, i, o, u.	* ia, ie, ii, io, iu.
* aa, ae, ai, ao, au.	* oa, oe, oi, oo, ou.
* ea, ee, ei, eo, eu.	* ua, ue, ui, uo, uu.

- **Consonantes aisladas**: el último grupo de unidades lo constituyen aquellos sonidos consonánticos que se pueden pronunciar aceptablemente sin necesidad de ir acompañados de una vocal. Se incluyen dentro de este grupo los siguientes fonemas: //, /f/, /x/, //, /s/, /_/.

2.2. OBTENCIÓN DE LAS UNIDADES

A la hora de obtener las unidades de voz es importante que cada una se halle lo más aislada posible de la influencia de otros fonemas o sonidos. Además de esto, todas las muestras deben estar almacenadas con un tono similar (a ser posible idéntico), puesto que será el sistema el que genere posteriormente la prosodia adecuada modificando ese tono para cada muestra. Con el fin de conseguir este segundo objetivo, se han tomado todas las muestras en una única sesión y con la ayuda de un generador de frecuencia como referencia. Para obtener las muestras lo más aisladas posible, se han obtenido a partir de logotomas en los que se encontraban en posición inicial o seguidas por consonantes oclusivas, puesto que éstas tienen un silencio previo que facilita la detección del final del sonido anterior. Cada una de las palabras se pronunció tres veces consecutivas con el fin de seleccionar posteriormente la de mayor calidad [Bullón,92].

Para la extracción de las unidades de voz definitivas se utilizó un editor de

señales que permitía realizar de una forma cómoda los tratamientos necesarios de la onda sonora. El proceso consistió en introducir en el computador las muestras, refinándose cada unidad individualmente e intentando ante todo que su longitud fuera la adecuada, teniendo en cuenta tanto los estudios referentes a la duración de cada fonema [Navarro,16], [Quilis,81], como resultados experimentales preliminares.

2.3. CONSTRUCCIÓN DE MENSAJES A PARTIR DE LAS UNIDADES

La forma en que se seleccionan las unidades definidas para formar frases viene explicada en [Pérez,89] y [Bullón,92]. A modo de resumen se exponen a continuación una serie de ejemplos ilustrativos.

ventana ---> be - e - en - ta - a - na - a.

derecho ---> de - e - re - e - ch - o.

casta ---> ka - a - s - ta - a.

trampa ---> ta - ra - a - am - pa - a.

fresa ---> f - re - e - s - a.

El tratamiento de los casos en que aparecen varias vocales consecutivas ha sido ligeramente modificado respecto al utilizado en [Pérez,89]. Lo que se hace en este caso es simplemente separar cada vocal en una unidad y posteriormente intercalar entre cada par de ellas una unidad con el fragmento de transición que corresponda. Esto se hace así tanto si los fonemas se encuentran dentro de una misma palabra como si pertenecen a palabras distintas. Mostremos unos ejemplos ilustrativos de este proceso:

veo hielo ---> be - e - eo - o - oi - i - ie - e - lo - o.

odio a Eusebio ---> o - di - i - io - o - oa - a - ae - e - eu - u - s - e - bi - i - io - o.

2.4. FORMACIÓN DE LA CURVA PROSÓDICA

Para dotar de prosodia a los mensajes se ha seguido la aproximación adoptada en [Pérez,89], según la cual se definen un total de ocho posibles *prosodemas* [Navarro,16] que serán aplicados a cada grupo fónico según su tipo. La forma en que este proceso se lleva a cabo puede consultarse en [Pérez,89].

3. Descripción del algoritmo PSOLA

El esquema general de funcionamiento del PSOLA se puede resumir en la ejecución de tres etapas [Moulines,90b][Valbret,91]: un análisis de la onda original para conseguir una representación no paramétrica de la misma, la modificación prosódica a

parti de esta representación, y finalmente la producción de la señal sintética construida a partir de la representación intermedia modificada.

En el primero de estos pasos, la señal original se descompone en una serie de unidades de corta duración superpuestas denominadas señales ST de análisis (de *Short Term*). En la segunda etapa, estas señales son modificadas (mediante cierto procedimiento que posteriormente describiremos) con lo que se convierten en lo que denominamos señales ST de síntesis. Mediante la superposición y suma de estas últimas se genera la onda sintetizada, siendo este proceso el que da lugar a la denominación PSOLA, que proviene de *Pitch Synchronous Overlap-Add* (solapamiento y suma sincronizada con la frecuencia fundamental). La forma en que se opere sobre las señales ST de análisis para obtener las de síntesis ha dado lugar a distintas variantes del método general.

En el llamado PSOLA/FFT, las unidades ST de síntesis se obtienen a partir de modificaciones en el dominio de la frecuencia de las de análisis. La complejidad de los cálculos que requiere este método sigue siendo demasiado elevada para poder implementarse directamente en sistemas sencillos. Más recientemente se ha introducido el PSOLA/MPLPC, que, aunque actúa en el dominio de la excitación mediante impulsos y es por tanto computacionalmente más simple que el anterior, sigue requiriendo la descomposición de la señal sonora. En [Hamon,89] pueden obtenerse más detalles acerca de estos dos métodos.

La variante que requiere menor esfuerzo computacional se conoce con el nombre de TD-PSOLA (TD por *Time Domain* o dominio del tiempo). La idea básica de funcionamiento consiste en variar el grado de solapamiento de las señales ST de síntesis con lo que se consigue modificar la frecuencia fundamental de la onda resultante sin cambiar su envolvente espectral. Paralelamente a este ajuste del solapamiento entre las señales, resulta sumamente sencillo controlar la longitud o duración de la onda producida mediante la duplicación o eliminación de alguna de las señales ST. Vamos a describir más formalmente el proceso completo.

a) Análisis y síntesis mediante PSOLA.

Como ya se ha comentado, la señal de voz digitalizada $s(n)$ se descompone en una serie de unidades superpuestas $s_m(n)$ denominadas señales ST de análisis. Éstas se obtienen multiplicando la señal por una secuencia de "ventanas" $h_m(n)$, según la expresión:

$$S_m(n) = h_m(t_m - n)\, s(n)$$

En esta ecuación, $h_m(n)$ representa una ventana de la que de momento sólo diremos que es simétrica y centrada en n=0 (posteriormente especificaremos más detalles sobre ella). Los sucesivos instantes t_m se seleccionan síncronamente con la frecuencia fundamental de la señal. En la figura 1 se puede ver una ilustración de este proceso.

sucesivos instantes t_m se seleccionan síncronamente con la frecuencia fundamental de la señal. En la figura 1 se puede ver una ilustración de este proceso.

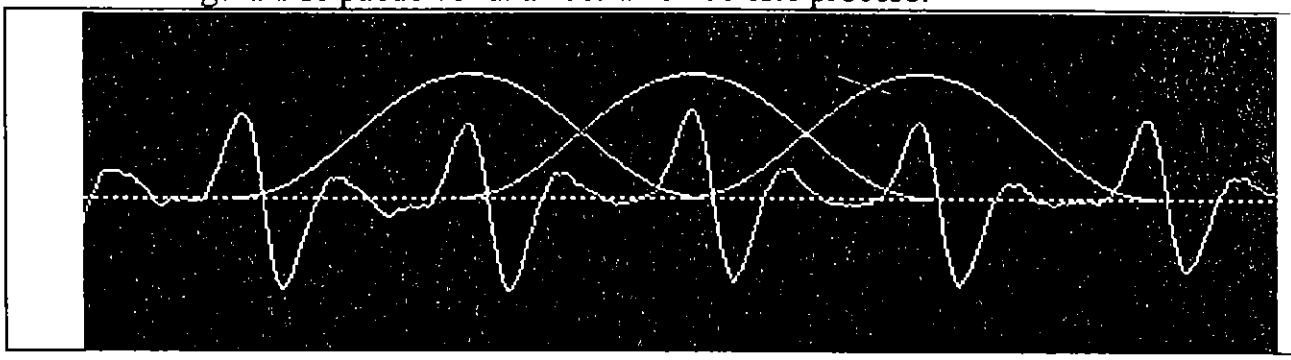


Figura 1. *Ventanas de análisis sobre la onda sonora. Las señales ST de análisis se obtendrán mediante el produ[illegible] de las ventanas por la señal, tal como se expresa en (3.1).*

La secuencia de señales ST que se obtiene es procesada según veremos para producir otro conjunto de señales ST, $s'_q(n)$, que denominamos de síntesis y que se sincronizan con un nuevo conjunto de marcas temporales t_q. La relación entre las ST de análisis y las de síntesis viene de este modo determinada por una función de alineamiento temporal: t_q ---> t_m, que implícitamente expresa la frecuencia de la onda sintetizada respecto a la de la original (figura 2).

Una vez realizados estos cálculos, la señal de voz sintética $s'(n)$ puede obtenerse mediante un proceso de solapamiento y suma de las señales ST de síntesis. Esto se puede hacer según la expresión:

$$s'(n) = \frac{\sum_q \alpha_q \, s'_q(n) \, h'_q(t'_q - n)}{\sum_q h'^2_q(t'_q - n)} \qquad (3.2)$$

en la que h representa las ventanas de síntesis, y α_q un factor de compensación debido a las variaciones de energía que se producen.

En nuestra aplicación es posible, entre otras simplificaciones, utilizar una ventana de síntesis constante sin una pérdida significativa de prestaciones [Hamon,89], reduciéndose la expresión anterior a:

$$s'(n) = \sum_q \alpha_q \, s'_q(n) \qquad (3.3)$$

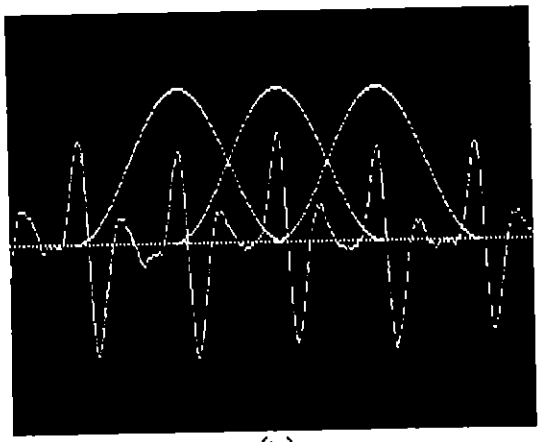
El factor de normalización _ se mantiene, en este caso, para compensar las modificaciones de energía que se pueden producir debido a la suma de los valores de las ventanas en las zonas de solapamiento de las mismas.

b) Algoritmo de modificación prosódica.

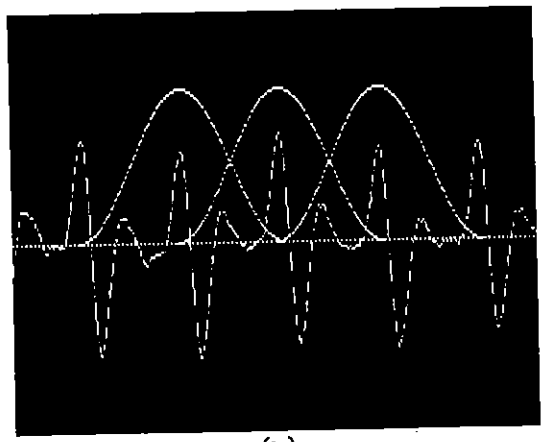
Como ya se ha comentado, el control de la frecuencia y de la duración de la señal sintética se lleva a cabo mediante la selección de las marcas t_q y la definición de la función de alineamiento temporal t_q ---> t_m, que relaciona las marcas de síntesis con las de análisis. Realmente, esta función lo que hace es asociar cada unidad ST de síntesis, $s(n)$, con la de análisis que debe ser copiada en su lugar, y los valores de t_q determinan los retardos que deben ser introducidos entre unidades sucesivas. Esto puede representarse mediante la siguiente expresión, que define las señales ST de síntesis a partir de las de análisis:

$$s'_q(n) = S_m(n - t_m + t'_q) \qquad (3.4)$$

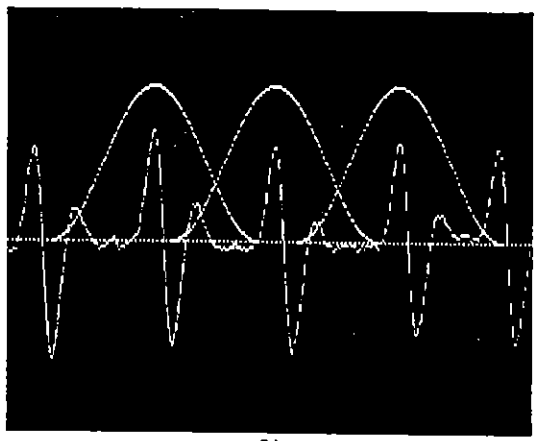
Si la duración y la frecuencia de la señal deben ser modificadas por un mismo factor β, la relación entre las señales ST de análisis y las de síntesis será de uno a uno. En este caso el algoritmo simplemente debe copiar las unidades de análisis en el eje de tiempo de las de síntesis, ajustando el retardo entre ellas según el factor β. En el caso general en que la duración y la frecuencia requieran factores de ajuste diferentes, la relación ya no será de uno a uno, pero lo único que deberá realizar el algoritmo será ajustar el retardo y eliminar o duplicar algunas de las unidades de análisis (figura 2).



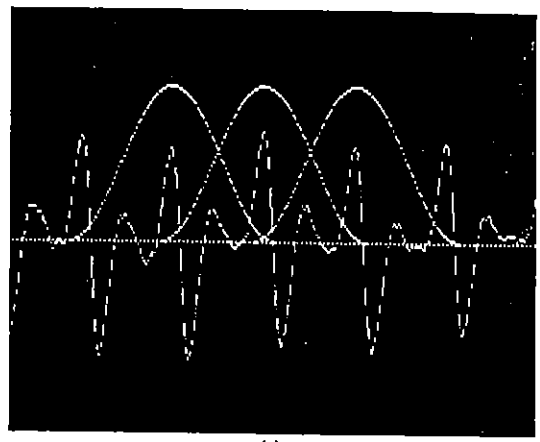
(a_1)



(a_2)



(b)



(c)

Figura 2. *En estas gráficas queda representado el efecto que produce en la onda sintetizada la variación del grado de solapamiento de las señales ST de análisis. En (a_1) (que aparece duplicada en (a_2) por mayor claridad) se puede ver la onda original. En (b) se ha producido una separación de las ventanas (que no varían individualmente de tamaño), originando una disminución de la frecuencia en la señal sintetizada. Obsérvese cómo de este modo se modifica igualmente la duración, por lo que, si se quiere mantener constante, deberá eliminarse alguna señal ST. En (c) se representa el proceso inverso, es decir, se aproximan las ventanas, con lo que se aumenta la frecuencia pero se disminuye la duración.*

En cuanto a la ventana a emplear, aquí, como en [Hamon,89] y [Moulines,90b], se hace uso de la denominada ventana de Hamming, definida según la expresión:

$$h(t) = 0.54 - 0.46 \cos\left(\frac{2t}{L}\right) \qquad (3.5)$$

siendo L la longitud de la ventana. La función se define para valores de t entre 0 y L, siendo igual a 0 en otro caso. En cuanto a la longitud que debe tener, en [Hamon,89] se comenta que es conveniente que sea proporcional al periodo de la señal en ese punto, según una expresión como la siguiente:

$$h'_q(n) = h\left(\frac{n}{\mu P}\right) \qquad (3.6)$$

En esta ecuación, $h(t)$ es la ventana definida en el intervalo unidad, P es el periodo de la onda en ese punto, y μ un factor de proporcionalidad que indica el número de periodos que abarca la ventana.

abarca la ventana. Más concretamente, en la referencia citada se sugiere un tamaño de ventana de dos periodos completos de onda ($\mu = 2$), es decir, definirla de una longitud

doble al periodo de la señal en ese punto. Otro factor importante a tener en cuanta a la hora de definir las ventanas [Moulines,90b] es el hecho de que deben estar rigurosamente sincronizadas con los instantes de mayor amplitud de la onda dentro de cada periodo (como se muestra en las figuras 1 y 2), ya que en caso contrario la calidad de la voz se ve sensiblemente afectada.

Cabe comentar por último una variante de esta ventana que ha sido utilizada con buenos resultados [Moulines,90b], y que consiste en intercalar un tramo horizontal en el máximo relativo de la función. Permitiendo que este tramo tenga una longitud variable, se puede obtener un control más flexible de la longitud total de la ventana, con lo que se simplifican los cálculos de obtención de la señal sintética. En nuestro caso, como veremos, esto no es necesario, ya que gran parte de los cálculos se hacen anticipadamente y las unidades se almacenan ya multiplicadas por la ventana.

4. Generación de la prosodia

La parte inicial del tratamiento que recibe el texto introducido en el sistema es muy similar a la utilizada en [Pérez,89], y consiste básicamente en descomponer cada grupo fónico en una serie de unidades fonéticas pertenecientes al conjunto definido. Dentro de este proceso se determina la acentuación y la duración de cada vocal, de modo que, según se describe en [Bullón,92], la serie de unidades fonéticas se almacena en una estructura que contiene, para cada unidad, su nombre, su tono y su duración. El tono se desdobla en dos componentes: inicial y final, que serán utilizadas posteriormente y que de momento reciben el mismo valor. Además, en este nivel todas las unidades adoptan únicamente dos posibles tonos, según se trate de una vocal acentuada o no. En el cuadro 1 se muestra un ejemplo de este proceso.

ENTRADA	: 'Me voy a casa.'
SALIDA	: (('ME',50,50,6), ('E',50,50,4), ('BO',50,50,6), ('O',56,56,8), ('OI',50,50,6), ('I',50,50,4), ('IA',50,50,6), ('A',50,50,4), ('KA',50,50,6), ('A'56,56,6), ('S',50,50,6), ('A',50,50,6)

Nota: El valor 50 corresponde al tono medio normal de una vocal átona, y el 56 al de una tónica. Por su parte, el 6 representa la duración media de una vocal semilarga, el 4, la de una breve, y el 8 la de una larga.**Cuadro 1**. *Ejemplo de descomposición en unidades fonéticas.*

Cuadro 1. *Ejemplo de descomposición en unidades fonéticas*

Para formar la entonación de la frase se asigna un tono a cada unidad fonética, en función de su posición relativa dentro del grupo fónico y del tipo de curva prosódica que tenga éste asociado. El procedimiento consiste en asignar valores en primer lugar a las cuatro unidades fonéticas que definen la curva melódica, a saber: la primera y última unidad acentuada del grupo fónico, y las que se encuentran en la primera y última posiciones absolutas del mismo. Al resto de las unidades se le asigna un tono calculado mediante una interpolación lineal entre estas cuatro referencias. Esto se hace de forma que el tono final de una unidad coincide siempre con el inicial de la siguiente, con el fin

de que no haya variaciones bruscas de frecuencia a lo largo de la frase. El resultado de este proceso se ilustra en el cuadro 2.

ENTRADA	: Descomposición mostrada en el cuadro 1.
SALIDA	: (('ME',32,40,6), ('E',40,48,4), ('BO',48,56,6), ('O',56,56,8), ('OI',56,50,6), ('I',50,50,4), ('IA',50,50,6), ('A',50,50,4), ('KA',50,56,6), ('A'56,56,6), ('S',56,41,6), ('A',41,26,6).**Cuadro 2.** *Ejemplo de codificación de la curva prosódica*

Cuadro 2. *Ejemplo de codificación de la curva prosódica.*

Con esto llegamos a una serie de unidades fonéticas pertenecientes a nuestro conjunto, y con tres valores numéricos asociados a cada una de ellas: uno representa su tono inicial, otro su tono final y el tercero su duración. Antes de comentar la manera en que se lleva a cabo la emisión de las muestras correspondientes a este conjunto de unidades, debe hacerse una descripción del método de proceso y organización de las estructuras de datos utilizadas.

Cada muestra de voz fue almacenada en un fichero, codificada mediante técnica PCM a una frecuencia de muestreo de 11 Khz. Se definió, para cada unidad, un conjunto de marcas sincronizadas con la frecuencia fundamental, y coincidentes con máximos relativos de la misma. Pese a que existen multitud de métodos automáticos de detección de la frecuencia fundamental, se desechó su utilización puesto que, al ser el tono de las muestras conocido y estable, se podía implementar un programa relativamente simple que permitiera realizarlo manualmente de forma cómoda, obteniendo una mayor precisión. Un ejemplo de este marcado se puede observar en la figura 3, que corresponde a un fragmento de la unidad 'A'.

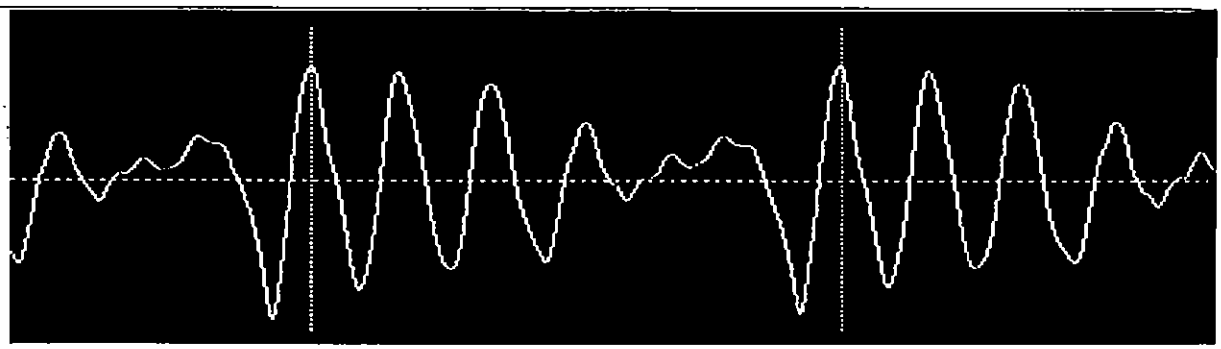


Figura 3. *Ejemplo del proceso de marcado del fichero de muestras correspondiente a la unidad 'A'.*

Una vez definido el conjunto completo de marcas para la totalidad de los ficheros, se calcula la ventana de Hamming centrada en cada marca y se multiplica por los valores del fragmento de señal que comprende. Pese a que las ventanas se solapan entre ellas, son almacenadas de forma contigua en una estructura tipo matriz en la que cada fila contiene una señal ST, es decir, el producto de una ventana por un fragmento de señal, habiendo tantas filas como señales ST definidas para esa muestra.

Las modificaciones del tono y la duración en la señal de voz sintetizada se producen variando el grado de solapamiento de las señales ST de la muestra (figura 2). Para llevar a cabo un control simultáneo de estos dos factores se ha adoptado la siguiente

simplificación: siempre que tanto el tono inicial como el final de la muestra a emitir sean superiores al tono definido como medio o normal, se duplicará una señal ST con el fin de compensar la variación en la longitud; si, por el contrario, ambos tonos son inferiores al normal, se eliminará una ventana. Si cada uno de los dos tonos queda a un lado del tono medio, asumiremos que se compensan el efecto alargador que se produce por un extremo de la muestra y el recortador que se produce por el opuesto, por lo que no se efectuará ningún control de tiempo. Hay que decir que, en lugar de seguir esta simplificación, se puede calcular de una manera también sencilla la variación exacta que se produce en la duración de la muestra debido a la modificación de su tono. Sin embargo, la unidad temporal mínima de que disponemos para compensar estas diferencias es la mitad de la longitud de una ventana, y, dentro del rango de variaciones de tono en que nos movemos, en ningún caso excederá de ese tiempo, por lo que el resultado obtenido es el mismo.

De un modo independiente a este proceso, la duración de cada unidad viene representada por un valor numérico que indica la cantidad de ventanas a eliminar o duplicar, en función de su diferencia respecto a un valor constante considerado como duración normal. La selección de las filas de la matriz a eliminar o replicar se hace de forma lineal, de modo que se hallen equidistantes entre sí y respecto a las filas de los extremos.

La modificación de la frecuencia, por su parte, se lleva a cabo paralelamente a la emisión de los valores de la muestra. Para ello se calculan, a partir de los tonos inicial y final asignados a la unidad, un punto de solapamiento asociado a cada fila de la matriz. La emisión propiamente dicha consiste en recorrer cada ventana, teniendo en cuenta que, si se alcanza el punto de solapamiento, debe sumarse el valor que corresponda de la ventana siguiente.

Como vimos en la expresión (3.3), el algoritmo PSOLA incluye un factor _ que compensa las modificaciones de amplitud que se producen en las zonas de solapamiento de las ventanas. En el presente sistema se ha optado por darle el valor constante 1 a dicho factor, con lo que en las zonas de mayor frecuencia de la señal resulta asimismo ligeramente aumentada su amplitud. Esto da lugar a un ligero acento de intensidad, que se superpone al acento melódico, mejorando la naturalidad de la voz producida. Este fenómeno se puede observar en las figuras 4 y 5, que representan, respectivamente, un fragmento de la primera y segunda vocal de la palabra *cazar*.

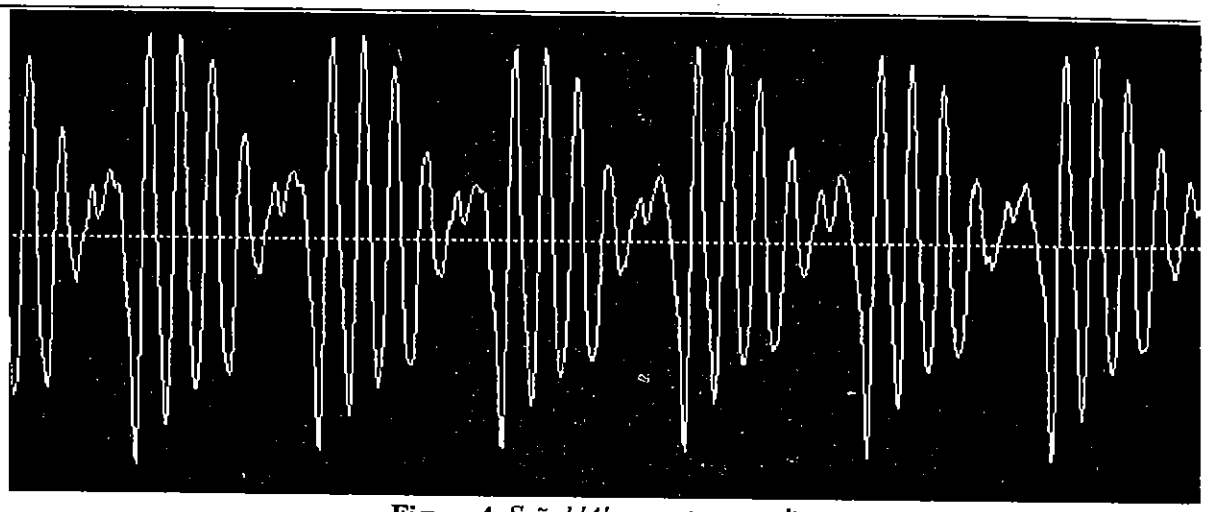


Figura 4. *Señal 'A' en un tono medio.*

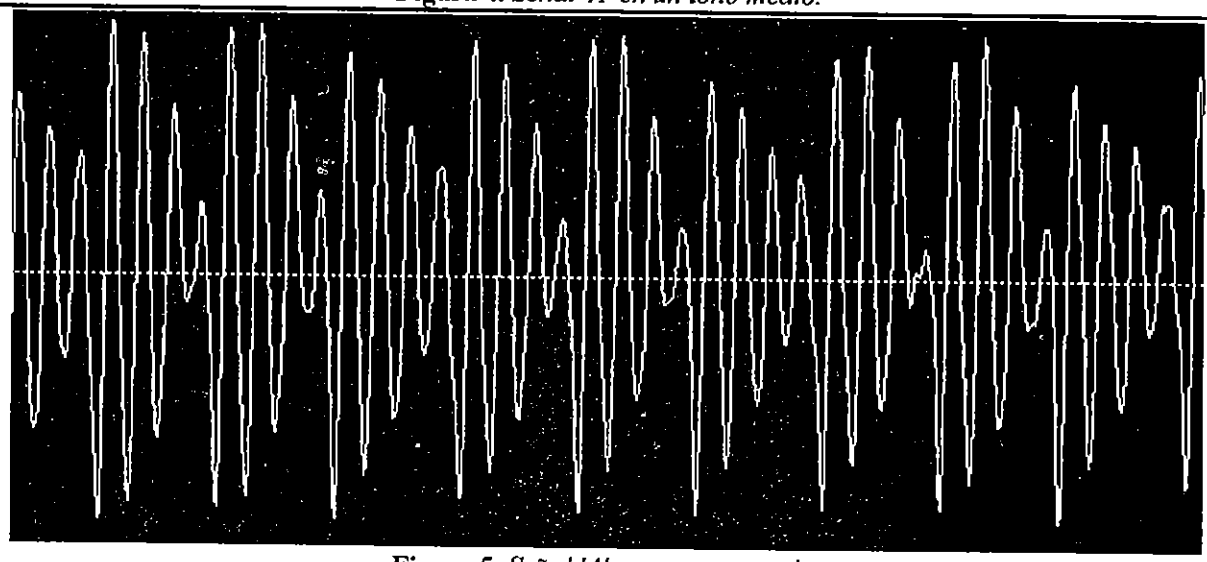


Figura 5. *Señal 'A' en un tono agudo.*

Por otra parte, las figuras 6 y 7 muestran los sonogramas de dos frases emitidas por el sintetizador; en ellos puede observarse cómo evolucionan los armónicos y la frecuencia fundamental a lo largo de las mismas.

5. Conclusiones

Se presentan los primeros resultados obtenidos al aplicar el método PSOLA en la síntesis de voz en castellano. Este método ha sido desarrollado recientemente y permite un control de la prosodia relativamente sencillo actuando sobre la representación de la señal en el dominio del tiempo. El hecho de que la señal se ha almacenado *ya multiplicada por la ventana* supone un planteamiento original que aporta una sensible disminución de las necesidades de cómputo en tiempo real. El método de síntesis empleado se ha basado en la concatenación de unidades previamente almacenadas, codificadas mediante la técnica PCM a una frecuencia de muestreo de 11 khz. El tipo de unidades seleccionadas han sido las introducidas en [Pérez,89], por la buena relación entre calidad y consumo de memoria demostrada en dicho trabajo. Todo esto ha permitido la implementación del sistema en un computador de tipo compatible PC sin material adicional (a excepción de un sencillo conversor D/A).

El resultado obtenido se puede calificar, en general, como notablemente bueno, y en cualquier caso sensiblemente mejor que el conseguido en [Pérez,89]. Se han llevado a cabo unas pruebas orientativas consistentes en la lectura de textos extraídos aleatoriamente de novelas y libros científicos, obteniéndose un elevado grado de comprensión. No obstante, todos los oyentes coincidieron en señalar que la voz obtenida presenta un aspecto claramente artificial. Esto es debido en parte al problema de la coarticulación, que, si bien es parcialmente resuelto con el tipo de unidades empleadas, no se puede soslayar completamente sin un completo conjunto de reglas a aplicar en el proceso de síntesis. Sin embargo, para ello sería preciso disponer de un computador sensiblemente más potente que el escogido en nuestro trabajo.

Una posible mejora del sistema pasaría por la inclusión de más de un alófono por cada fonema en el conjunto de unidades. Esto implicaría un mayor consumo de memoria y de tiempo, aunque, teniendo en cuenta la evolución de las prestaciones de los computadores tipo compatible PC, no supondría una limitación importante.

Como conclusión de todo lo expuesto, se puede comentar que:

- La síntesis de voz mediante concatenación de unidades previamente almacenadas puede ser empleada, con un resultado aceptable, en computadores compatibles PC con una configuración mínima basada en el microprocesador Intel 386 y una memoria RAM de 640 kbytes, bajo sistema operativo MS-DOS.

- La aplicación del esquema PSOLA para el control de la prosodia, incluyendo la acentuación, la cantidad y la entonación, parece una opción válida para un sistema de síntesis de esas características. Ello se debe a la simplicidad de los cálculos que requiere, derivada fundamentalmente del hecho de que se aplica directamente a la representación temporal de la señal.

- La determinación de las características de las unidades a ser concatenadas juega un papel muy importante en el resultado final. En esto se incluye, en primer lugar, tanto la obtención de unas unidades de una buena calidad como la determinación de su longitud; por otra parte, es igualmente importante el cálculo de sus peculiaridades a la hora de ser emitidas, que dependerán de su contexto en el mensaje al que pertencen. Dentro de estas peculiaridades nos referimos a la determinación de su tono y duración, calculados a partir de estudios fonéticos previos.

- Sin embargo, la voz obtenida mediante un sistema de este tipo presentará inevitablemente un aspecto artificial, derivado de la imposibilidad de resolver completamente el problema de la coarticulación entre fonemas. Por este motivo, si se pretende eliminar este problema se deberá recurrir a un método de síntesis por reglas o basado en algún tipo de aprendizaje inductivo, llevado a cabo por

un computador de mayores prestaciones. No menospreciamos pese a ello la gran utilidad que está demostrando el presente sistema, principalmente en la comunicación de personas con algún tipo de minusvalía, y más si tenemos en cuenta la enorme implantación que tiene en el mercado el tipo de computador para el que se ha diseñado.

Referencias

Bullón, J.L. (1992). *Conversión de texto a voz en castellano mediante técnicas en el dominio del tiempo*. Proyecto Fin de Carrera. Facultad de Informática. Universidad Politécnica de Valencia.

Garrido, J.M. (1991a). *Modelización de patrones melódicos del español para la síntesis y el reconocimiento del habla*. Universidad Autónoma de Barcelona.

Garrido, J.M. (1991b). *Estilización de patrones melódicos del español para sistemas de conversión texto-habla*. Boletín de la Sociedad Española para el procesamiento del Lenguaje Natural, nº 11. Dic. 1991.

Hamon, C. ; Moulines, E. ; Charpentier, F. (1989). *A Diphone Synthesis System based on Time-Domain Prosodic Modifications of Speech*. Proc. of IEEE Intern. Conf. ICASSP 89; pp.238-241.

Llisterri, J. (1989). *La síntesis del habla: Estado de la cuestión*. Boletín de la Sociedad Española para el procesamiento del Lenguaje Natural, núm 6; pp.19-41.

Martí, J.M. (1986). *Estudi acústic del català i síntesi automàtica per ordinador*. Tesis Doctoral. Facultad de Ciencias Físicas. Universidad de Valencia.

Moulines, E.; Sorin, C.; Charpentier, F. (1990a). *New approaches for improving the quality of text-to-speech systems*. Verba 90. Int. Conference on Speech Technology. Roma.

Moulines, E.; Emerad, F.; Larreur, D.; Charpentier, F. (1990b). *A real-time french text-to-speech system generating high-quality synthetic speech*. Proc. of IEEE Intern. Conf. ICASSP 90; pp.309-312.

Navarro, T. (1916). *Manual de pronunciación española*. Madrid, CSIC. Decimoctava edición, 1974.

Pérez, J.C. (1989). *Sistema de Conversión de Texto a Voz para Castellano*. Proyecto Fin de Carrera. Facultad de Informática. Universidad Politécnica de Valencia.

Pérez, J.C.; Vidal, E. (1991). *Un Sistema de Conversión de Texto a Voz para Castellano*. Boletín de la Sociedad Española para el procesamiento del Lenguaje Natural. Dic. 1991.

Quilis, A. (1981). *Fonética Acústica de la Lengua Española*. Editorial Gredos. Madrid.

Valbret, H.; Moulines, E.; Tubach, J.P. (1991). *Voice transformation using PSOLA technique*. Proc. Eurospeech 91, Génova, sep 91; pp. 345-348.