REALIZACIÓN DE UN SINTENTIZADORMULTIPULSO PARA UN CONVERSOR TEXTO-VOZ EN CASTELLANO*

A. J. Rubio Ayuso
J. C. Segura Luna
V. E. Sánchez Calle
J. M. López Soler
J. L. Pérez Córdoba

Departamento de Electrónica y Tecnología de los Computadores
Universidad de Granada

Resumen

Se describe ~~en este~~ trabajo la realización de un sintetizador multipulso para un conversor texto-voz en castellano. Las unidades sonoras con las que se compone el mensaje oral son difonemas, ya que éstos permiten la correcta resolución de los problemas que plantea la concatenación entre dos unidades. El tipo de codificación utilizada es la de LPC Multipulso con análisis síncrono, que permite de forma sencilla el control del *pitch* y de la duración de cada uno de los sonidos.

1 Introducción

La utilización cada vez más frecuente de diversas máquinas para la realización de tareas con interacción hombre-máquina está requiriendo el desarrollo de la capacidad de reproducir mensajes en forma oral por parte de éstas, ya que es la forma más sencilla y eficiente de comunicación humana [3].

Cuando la variedad de mensajes es suficientemente amplia y si su contenido debe adaptarse con facilidad a cambios producidos en una base de información, no es adecuada la simple reproducción concatenada de mensajes pregrabados. Además, los mensajes pregrabados no posibilitan el control prosódico de las frases pronunciadas.

En estos casos es necesaria la introducción de sistemas de conversión de textos escritos en voz humana sintética. Así, los mensajes escritos se leen de un fichero o se generan a partir de los datos proporcionados por un módulo especializado, y el Conversor Texto-Voz produce la señal sonora necesaria, incorporando la información contextual disponible para el control automático de la prosodia.

En este trabajo se describe la realización de un Sintetizador de Voz, que toma como entrada los resultados de un Módulo de Control Prosódico y la transcripción escrita de la frase a pronunciar.

En la siguiente sección se presenta un visión general del problema de la conversión texto a voz, junto con un esquema general de módulos de la solución adoptada.

En la sección 3 se plantean los problemas prácticos principales que surjen a la hora de establecer un método sistemático para el análisis de la voz y su síntesis posterior. La solución a estos problemas adoptada en este trabajo se presenta en las secciones 4 y 5.

*Este trabajo es resultado de un contrato entre la Universidad de Granada y ENA Telecomunicaciones SA

Las secciones 6 y 7 explican los detalles del análisis de la voz para la creación de la base de datos de difonemas y de la síntesis de voz a partir de dicha base de datos.

El trabajo finaliza con la presentación de los resultados experimentales y el comentario de las conclusiones obtenidas.

2 Esquema general de un conversor texto-voz

Como puede verse en la figura 1, el conversor texto-voz, tal como está concebido en este trabajo, consta de cuatro módulos fundamentales.

El *Transcriptor* convierte el texto escrito en una cadena de unidades sonoras adecuadas para el correcto funcionamiento del sistema. Como se indicará posteriormente, las unidades sonoras elegidas en el presente trabajo son los difonemas, ya que permiten resolver de forma adecuada y sencilla los principales problemas relacionados con la conexión entre distintas unidades sonoras.

Otro de los módulos del conversor es el *Control Prosódico*. Este módulo establece los perfiles de amplitud y *pitch* que han de ser utilizados en la síntesis de la señal, así como la duración que debe tener cada uno de las unidades.

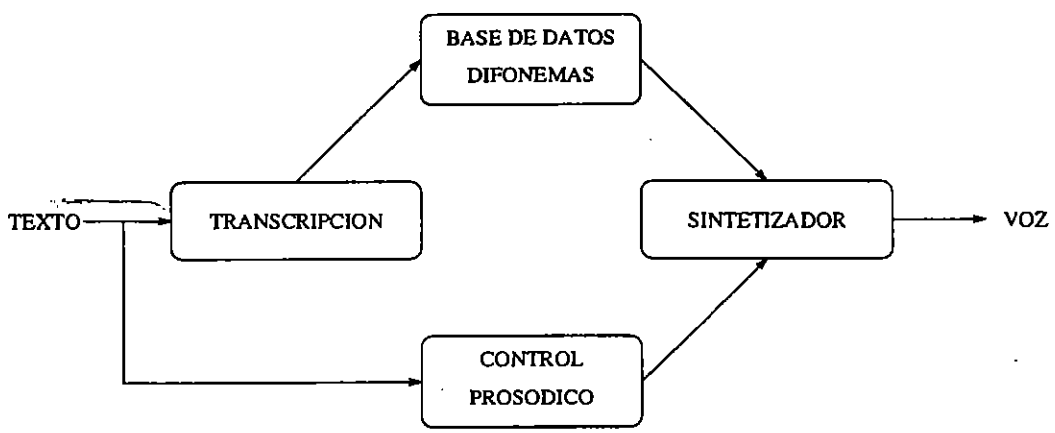
La síntesis propiamente dicha se hace utilizando los datos almacenados en una *base de datos de unidades* (difonemas en este caso), correspondientes a las unidades procedentes de la transcripción del texto. Esta base contiene toda la información necesaria para la generación de los distintos sonidos, debiendo permitir una correcta concatenación de sonidos, junto con un perfecto control de la amplitud, de la frecuencia fundamental (pitch), y de la duración de cada uno de los sonidos elementales.

3 Problemas relacionados con la síntesis de las unidades sonoras

Al decidir la forma general en que el conversor va a obtener voz a partir de las unidades almacenadas en la memoria, hay que tener presentes los problemas que han de resolverse para conseguir que la síntesis sea agradable al oído, y de calidad suficientemente buena.

En primer lugar, es obvio que la única solución medianamente sencilla y poco costosa en volumen de memoria necesaria es la síntesis de la voz a partir de una serie de unidades sonoras elementales que representen de la forma más adecuada y completa al idioma castellano. En realidad, esta es la única opción razonable, razón por la que su elección se ha dado por supuesta en lo que antecede [7]. El problema real estriba en que las unidades deben ser tales que

Figura 1: Esquema general del conversor



permitan ser representadas de forma que faciliten su concatenación. Se comprueba fácilmente que la concatenación apropiada de las unidades sonoras es uno de los caballos de batalla de un convertidor texto-voz. No debe haber brusquedades de ningún tipo en las fronteras entre segmentos concatenados, a la vez que debe permitirse un perfecto control independiente de la amplitud, del pitch de la señal y de la duración de cada uno de los fonemas. Esto implica la elección cuidadosa de tanto el tipo de codificación utilizada como el tipo de unidades seleccionadas.

Estos tres parámetros, amplitud, pitch y duración son el resultado del módulo de control prosódico y determinan la entonación de la frase.

4 Elección del tipo de unidades sonoras

El tipo de las unidades sonoras debe elegirse de modo que permitan resolver los problemas planteados en la sección anterior.

Deben cumplir también simultáneamente los requisitos de representar suficientemente bien todas las posibilidades sonoras del idioma castellano y de constituir un conjunto de cardinal no excesivamente elevado, ya que la base de datos que las contenga deberá ser albergada en la memoria del sistema de conversión completo.

Entre las distintas posibilidades encontradas en la literatura [1], y atendiendo a los criterios anteriormente mencionados, en este trabajo se han utilizado los difonemas como unidades elementales para la síntesis de la voz. Se entiende por difonema el segmento de señal sonora que va desde la parte estable de un fonema (generalmente en su centro temporal) hasta la parte estable del siguiente fonema.

Efectivamente, los difonemas cumplen suficientemente bien las condiciones impuestas a priori. Representan razonablemente bien al conjunto de sonidos del idioma, ya que están incluidas todas las coarticulaciones posibles. Además, estas coarticulaciones están bien representadas ya que los fonemas se parten siempre por su parte más estable, supuestamente muy independiente del contexto.

Una posible alternativa, más segura desde el punto de vista de las coarticulaciones, sería la consideración de los llamados trifonemas, pero tiene el grave inconveniente del elevado número de tales unidades. El número de difonemas es algo menor que el cuadrado del número de fonemas, ya que hay algunas combinaciones imposibles en castellano. En este trabajo se utiliza un número de difonemas inferior a 500.

El control de la duración de los distintos fonemas queda facilitado enormemente con la elección de los difonemas como unidades elementales de síntesis. Esto es debido también a que los difonemas comienzan y terminan en las partes estables de los dos fonemas a los que pertenece. Esto permite su fácil prolongación temporal al ser estables las fronteras entre difonemas.

En el caso de los sonidos sonoros, la prolongación consiste en la repetición periódica del último periodo almacenado en la base de datos y correspondiente al difonema en cuestión. De la misma, forma el acortamiento temporal de un sonido consiste en la supresión de los periodos necesarios. Esta repetición o supresión de periodos da buenos resultados precisamente gracias al hecho de que los extremos de los difonemas corresponden a zonas estables de los fonemas. En el caso de los sonidos sordos se alarga o acorta el tiempo necesario, utilizando sencillamente la información contenida en los límites del difonema.

5 Elección del tipo de codificación

La información contenida en la base de datos, acerca de los distintos difonemas, debe permitir la síntesis correcta de los mismos, teniendo presente que debe admitir una concatenación suave de distintos difonemas, permitiendo además el control de amplitud, pitch y duración.

La opción más obvia es la codificación LPC (Linear Predictive Coding) de cada uno de los difonemas. En este supuesto, cada uno de los difonemas vendría descrito mediante una serie

de ventanas (*frames*) de análisis, en las que se supone que la señal es suficientemente estable para ser representada por un conjunto de coeficientes de predicción, junto con la decisión de sonoridad/no sonoridad.

Pero la codificación LPC implica que cada uno de los segmentos sonoros se generará en el sintetizador a partir de un solo pulso de excitación por ventana, lo que proporciona una calidad aceptable para ciertas aplicaciones, pero claramente mejorable.

En este trabajo, atendiendo a la calidad deseable para la voz sintetizada finalmente, se ha decidido utilizar la codificación LPC Multipulso (MPLPC). En esta técnica, la excitación para la síntesis de la voz consiste en una serie de pulsos de posiciones y amplitudes variables de una ventana a otra (ver siguiente sección). La información disponible para cada una de las ventanas es claramente mayor, aumentando sensiblemente la calidad y naturalidad de la voz generada [9].

El problema que plantea la codificación MPLPC es que no permite el control del pitch de forma tan sencilla como en el caso de LPC normal, en el que el control consiste sencillamente en la separación temporal adecuada de los pulsos de excitación.

Esta dificultad ha impuesto la necesidad de realizar el análisis MPLPC síncrono con el periodo del pitch, que en principio no había que calcular. De nuevo, el control del pitch consiste simplemente en separar o aproximar en el tiempo los conjuntos de pulsos correspondientes a cada una de las ventanas que componen uno de los difonemas.

El cambio de un difonema a otro no supone más que el cambio de los coeficientes de predicción (que de todos modos cambian con cada ventana) y del conjunto de pulsos de excitación, que igualmente cambia en cada ventana. No hay diferencia apreciable entre el cambio producido entre dos ventanas consecutivas del mismo difonema y el cambio entre la última ventana de un difonema y la primera del siguiente.

En la sección siguiente se describen con más detalle los aspectos matemáticos de la codificación MPLPC.

6 Análisis síncrono con LPC Multipulso

El sistema diseñado trabaja con una representación paramétrica de la voz basada en el modelo de Predicción Lineal [4]. En este modelo la señal de voz $s(n)$ viene dada como una combinación lineal de los valores anteriores y de una entrada $u(n)$

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + Gu(n) \qquad (1)$$

donde G es un factor de ganancia y a_k los denominados coeficientes de predicción. A estas técnicas de análisis predictivas se les denomina a menudo técnicas LPC (Linear Predictive Coding). Por este motivo, dicho modelo aplicado a la voz recibe el nombre de modelo LPC de producción de voz. En este modelo se representa el tracto vocal por un filtro lineal todo polos variable con el tiempo $H(z)$, caracterizado por los coeficientes de predicción lineal a_k, y se separa la contribución al espectro de la señal por parte de la excitación y por parte del filtro lineal. El filtro determina la envolvente espectral mientras que la excitación crea la estructura fina.

Partiendo de ese modelo, el primer objetivo es determinar el valor de los coeficientes de predicción a_k para una ventana de análisis formada por N muestras. Partiendo de la base de que la excitación $u(n)$ es totalmente desconocida lo más inmediato es expresar $s(n)$ de una forma aproximada como una combinación lineal de las muestras anteriores [4]. A esta aproximación de $s(n)$ la vamos a llamar $\bar{s}(n)$

$$\bar{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \qquad (2)$$

y el error $e(n)$ entre el valor actual de $s(n)$ y el predicho $\bar{s}(n)$ es

$$e(n) = s(n) - \bar{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \quad (3)$$

A continuación se calculan los parámetros a_k que hacen mínimo el error, en concreto aquellos que minimizan el error cuadrático medio E

$$E = \sum_n e^2(n) = \sum_n (s(n) + \sum_{k=1}^{p} a_k s(n-k))^2 \quad (4)$$

E se minimiza haciendo

$$\frac{\partial E}{\partial a_k} = 0, \quad 1 \leq i \leq p \quad (5)$$

Resolviendo el sistema de ecuaciones a que da lugar (5) se obtienen los coeficientes de predicción a_k.

Hay diversas técnicas para llevar a cabo dicho proceso de minimización y de resolución del sistema de ecuaciones resultante [6]. En este trabajo se va a utilizar el denominado método de Burg, el cual garantiza la estabilidad del filtro resultante y es apropiado para segmentos cortos de voz.

6.1 EL modelo de excitación multipulso

En la sección anterior se ha estudiado como caracterizar el filtro de predicción lineal con el que se modela el tracto vocal. En esta sección se va a estudiar como caracterizar la excitación de dicho filtro.

Efectivamente, para calcular los coeficientes del filtro todo polos se hacía la aproximación

$$\bar{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \quad (6)$$

de donde teniendo en cuenta que

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + Gu(n) \quad (7)$$

se tiene para el error $e(n)$

$$e(n) = Gu(n) \quad (8)$$

donde $Gu(n)$ es denominada señal residuo [6] y constituye la excitación del filtro todo polos normalizado en ganancia.

Denominando $x(n)$ a la excitación, se puede expresar $s(n)$ también de la siguiente forma

$$s(n) = \sum_{k=0}^{n} h(n-k)x(k) + r(n), \quad 0 \leq n \leq N-1 \quad (9)$$

donde el primer sumando representa la convolución entre la excitación del filtro todo polos $x(n)$ y la respuesta impulso de dicho filtro $h(n)$ y el segundo constituye la contribución de la memoria del filtro. Pasando este segundo sumando al primer término de (9),

$$y(n) = s(n) - r(n) = \sum_{k=0}^{n} h(n-k)x(k), \quad 0 \leq n \leq N-1 \quad (10)$$

se obtiene una nueva señal $y(n)$ a la que solo contribuyen la excitación y la respuesta impulso del filtro de la ventana actual.

En el modelo de excitación multipulso se expresa la señal residuo o excitación $x(n)$ como una combinación lineal de l vectores unidad o, lo que es lo mismo, va a estar representada por l pulsos. Se tiene, pues, la siguiente expresión para $x(n)$

$$x(n) = \sum_{k=0}^{l-1} b_k \delta(n - n_k) \tag{11}$$

donde b_k son las amplitudes de los pulsos y n_k las correspondientes posiciones.

La expresión del error cuadrático medio que se comete al utilizar esta representación de la excitación viene dada por

$$E = \sum_{n=0}^{N-1} (y(n) - x(n) * h(n))^2 = \sum_{n=0}^{N-1} (y(n) - \sum_{k=0}^{l-1} b_k h(n - n_k))^2 \tag{12}$$

Hay que minimizar E, tanto con respecto a las amplitudes de los pulsos b_k, como con respecto a las posiciones n_k. Minimizando sólo con respecto a las amplitudes b_k, $0 \leq k \leq l-1$ se obtiene el siguiente sistema de ecuaciones:

$$\sum_{k=0}^{l-1} b_k \psi_{n_k n_j} = c_{n_j}, \quad 0 \leq j \leq l-1 \tag{13}$$

donde ψ es la autocorrelación de la respuesta impulso $h(n)$

$$\psi_{ij} = \sum_{n=0}^{N-1} h(n-i)h(n-j), \quad 0 \leq i,j \leq N-1 \tag{14}$$

y c es la correlación cruzada entre la respuesta impulso $h(n)$ y la señal $y(n) = s(n) - r(n)$

$$c_i = \sum_{n=0}^{N-1} y(n)h(n-i), \quad 0 \leq i \leq N-1 \tag{15}$$

Sustituyendo en la expresión de E se obtiene el error mínimo para un conjunto de posiciones dadas:

$$E^m = \sum_{n=0}^{N-1} y^2(n) - \sum_{k=0}^{l-1} b_k c_{n_k} \tag{16}$$

6.1.1 Determinación de las posiciones y las amplitudes de los pulsos

Para resolver óptimamente el problema de determinar las posiciones y amplitudes de los pulsos que minimizan E habría que tomar todas las combinaciones de l posiciones de entre las N posibles posiciones y resolver el sistema de ecuaciones (13) para todas ellas. Una vez hecho esto, habría que sustituir los valores de b_k calculados para cada combinación de posiciones en (16) y ver qué combinación hace mínimo E^m. Lógicamente esto es imposible de llevar a la práctica ya que hay $\binom{N}{l} = \frac{N!}{l!(N-l)!}$ combinaciones posibles para las posiciones y este número se hace rápidamente intratable a medida que el número de pulsos l aumenta. Por consiguiente es preciso recurrir a algun procedimiento sub-óptimo [8] que no requiera tanto cálculo.

El procedimiento va a consistir en calcular las amplitudes y posiciones por pasos, determinando la posición y la amplitud de un solo pulso en cada paso. Esto reduce el cálculo necesario a l búsquedas de orden N. El procedimiento comienza calculando la posición y la amplitud de un solo pulso, a continuación se calcula el error restando la contribución del pulso que se acaba de calcular y así se continua el proceso de buscar nuevos pulsos, un pulso cada

vez, para reducir el error cuadrático medio pesado hasta que tenga un valor lo suficientemente pequeño.

Supóngase que la excitación consistiera sólo en un pulso situado en la posición i, en este caso las ecuaciones (13) y (16) se reducen a

$$b_0 = c_i/\psi_{ii} \tag{17}$$

y

$$E^m = \sum_{n=0}^{N-1} y^2(n) - c_i^2/\psi_{ii} \tag{18}$$

Como puede verse en la ecuación (18), E^m queda como una función de i y además se ve que para minimizar E^m hay que maximizar la expresión c_i^2/ψ_{ii} que depende exclusivamente de i. Luego lo que se debe hacer es ir probando los N valores que puede tomar i y ver cuál de ellos maximiza c_i^2/ψ_{ii}. Ese valor de i, al que se llamará n_0, será la posición del primer pulso. La amplitud se calcula simplemente sustituyendo en (17). Una vez hecho esto puede considerarse que b_0 y n_0 son constantes conocidas en la ecuación (12) y se añade un pulso más b_1. Minimizando E con respecto a b_1 queda la ecuación

$$b_1\psi_{n_1n_1} = c_{n_1} - b_0\psi_{n_0n_1} \tag{19}$$

Como puede verse, esta ecuación sólo tiene dos incógnitas b_1 y n_1. Si la posición de este segundo pulso es i, despejando b_1 queda

$$b_1 = [c_i - b_0\psi_{n_0i}]/\psi_{ii} \tag{20}$$

y

$$E^m = \sum_{n=0}^{N-1} y^2(n) - [c_i - b_0\psi_{n_0i}]^2/\psi_{ii} \tag{21}$$

La posición del pulso será aquel valor de i que maximize la expresión

$$[c_i - b_0\psi_{n_0i}]^2/\psi_{ii} \tag{22}$$

Una vez determinado el valor de n_1 se calcula b_1 sustituyendo en (20).

Este procedimiento se repite para cada pulso, es decir el pulso j se encuentra en el máximo de la función

$$\frac{[c_i - \sum_{k=0}^{j-1} b_k\psi_{n_ki}]^2}{\psi_{ii}}, \quad 0 \leq i \leq N-1 \tag{23}$$

y su amplitud viene dada por

$$b_j = \frac{[c_i - \sum_{k=0}^{j-1} b_k\psi_{n_ki}]}{\psi n_j} n_j \tag{24}$$

Como se ve, el procedimiento consta de 3 pasos para cada pulso:

1. Construcción de la función dada por la ecuación (23).
2. Cálculo del máximo de esa función.
3. Cálculo de la amplitud del pulso mediante la ecuación (24).

Finalmente, una vez que las posiciones de todos los pulsos han sido determinadas, se recalculan las amplitudes usando la ecuación (13). También se pueden recalcular las amplitudes después de determinar la posición de cada pulso [2]. en vez de sólo una vez al final del proceso.

6.2 Análisis síncrono

En esta sección se va a describir el procedimiento de análisis de la señal de voz desarrollado para el conversor texto a voz. En las secciones anteriores se ha presentado un modelo matemático de producción de de la señal de voz que consiste en un filtro todo polos que modela el tracto vocal y la excitación a dicho filtro. De acuerdo con este modelo un segmento de voz de longitud N va a estar representado por k coeficientes de predicción a_k y las amplitudes y las posiciones de los l pulsos de la excitación multipulso.

Se seguirá un procedimiento de análisis distinto para los sonidos sonoros y los sordos. Como es bien conocido los sonidos de voz pueden clasificarse básicamente en dos clases, sonoros y sordos, dependiendo de su modo de excitación. Durante la producción de voz sonora el tracto vocal es excitado por una serie de pulsos glotales casi periódicos generados por las cuerdas vocales. Al periodo con que se generan los pulsos glotales se le denomina periodo de pitch. En la excitación para señal sorda no intervienen las cuerdas vocales y consiste simplemente en las turbulencias del aire que proviene de los pulmones. Si se observa la señal residuo $Gu(n)$ para distintos segmentos de voz se ve que presenta unos picos periódicos correspondientes a los pulsos glotales para unos y sin embargo para otros la excitación tiene un aspecto se señal ruido. Los primeros corresponderían a señal sonora y los segundos a sorda.

Se va a realizar un análisis síncrono para los segmentos sonoros y un análisis asíncrono para los sordos (en realidad en estos segmentos, el sincronismo no tiene sentido). Por síncrono se entiende que se va a utilizar una ventana de análisis de longitud igual a un periodo de pitch, mientras que en el análisis asíncrono el tamaño de la ventana es fijo y determinado de antemano.

Inicialmente se toma la señal de voz y utilizando el conocido algoritmo SIFT [5] se determina qué segmentos son sonoros o sordos y para aquellos que son sonoros se calcula el periodo de pitch.

A continuación para los segmentos sonoros se realiza un analisis síncrono utilizando ventanas de análisis de longitud igual al periodo de pitch y centradas sobre los picos de la señal residuo correspondientes a pulsos glotales. Se realiza un análisis LPC de orden 14 de cada una de estas ventanas de señal y se calcula la excitación multipulso, utilizandose 6 pulsos cada 4 ms. Por consiguiente cada ventana de señal sonora quedará caracterizada por el periodo de pitch, los catorce coeficientes de predicción y las posiciones y amplitudes de los pulsos normalizados en energía.

Para los segmentos sordos se utiliza una ventana de longitud 8 ms, siendo el orden de predicción para el filtro LPC tambien 14 y utilizándose asimismo 6 pulsos cada 4 ms para la excitación multipulso. Por tanto cada ventana de senal sorda quedará caracterizada por los catorce coeficientes de predicción y las posiciones y amplitudes de los pulsos normalizados en energía.

De acuerdo con las expresiones vistas en esta y anteriores secciones, la síntesis de la señal sonora, una vez conocidos los coeficientes de predicción y la amplitud y posición de los pulsos, es similar a la correspondiente a LPC, con la diferencia de que en lugar de un pulso de excitación por periodo se dispone de varios, cada uno de ellos en un instante de tiempo determinado y de un valor de amplitud distinto.

7 Creación de la base de difonemas

Una vez seleccionado el locutor apropiado para la creación de la base de datos de difonemas, lo que se hace atendiendo a las posibilidades que muestra cada uno de los locutores disponibles en cuanto a buen control del pitch y otros parámetros, se utiliza un conjunto de frases especialmente diseñadas para contener todos los fonemas en la mayor parte de los contextos. El conjunto de frases utilizadas en este trabajo resulta ser de 170 frases.

A continuación se ha hecho una segmentación de las frases en difonemas, utilizando para ello los contornos de pitch y de amplitud como información adicional a la propia señal.

Hecha la segmentación, se ha escrito en un fichero la lista de los distintos difonemas, junto

con su localización en el conjunto de frases. Esta localización se elige entre todas las posibles para un difonema dado, en función de la calidad auditiva de los resultados, aunque hay algunas características que se consideran a priori, como pueden ser la posible saturación de la señal en el momento de la grabación, su correcta segmentación, etc.

Un programa de ordenador utiliza este fichero para la creación de la base de datos. Esta consiste de una cabecera seguida de la información correspondiente a cada uno de los difonemas.

La cabecera de la base de datos contiene el número de difonemas, junto con la lista de éstos y su localización en la base de datos (de frases). La localización está expresada mediante el fichero original del cual se toman las muestras para el análisis, los instantes de tiempo inicial y final del correspondiente difonema y el *offset* en la base de datos de difonemas.

La información correspondiente a cada uno de los difonemas, que aparece en la base de datos es la siguiente:

- Representación ASCII del difonema (2 caracteres).
- Número de segmentos de análisis correspondientes al difonema.
- Para cada uno de los segmentos:
  - Longitud del segmento (número de muestras).
  - Clasificación sonoro/no sonoro.
  - Ganancia de predicción.
  - Amplitud cuadrática media.
  - Número de coeficientes de predicción lineal.
  - Vector de coeficientes de predicción lineal.
  - Número de pulsos para la excitación.
  - Vector de amplitud de los pulsos de excitación.
  - Vector de posición de los pulsos de excitación.

Aunque alguno de los parámetros (concretamente pitch y amplitud) podrían haberse suprimido, ya que el sintetizador ha de utilizar los proporcionados por el Control Prosódico, se han incluido en la base de datos para permitir comprobaciones, al disponerse en todo momento de los valores originales a cada uno de los segmentos que componen los distintos difonemas.

8 Conclusión

Se ha presentado en este trabajo la realización de un sintetizador LPC multipluso síncrono que ha mostrado ser de gran valor en su aplicación a un conversor texto-voz en castellano.

Las características de la síntesis LPC multipluso son tales que proporcionan una adecuada calidad de la voz sintetizada. El hecho de que el análisis realizado sea síncrono permite el sencillo control de los parámetros de pitch y duración de los distintos sonidos, fundamentales junto con la amplitud, de la naturalidad de la voz generada.

El sistema completo funciona por concatenación de difonemas, para lo que se ha construido una base de datos con todos los difonemas correspondientes al idioma castellano, para un locutor masculino.

Referencias

1. Bagger-Sorensen, B; Bertelsen, O; Domler, P "A text-to-speech system for Danish", EUSIPCO-90 (Barcelona), pp. 1119-1122.
2. Berouti, M.; Garten, H.; Kabal, P.; Mermelstein, P. "Efficient Compatation and Encoding of the Multipulse Excitation for LPC", Proc. ICASSP-84, pp. 10.1.1-10.1.4, 1984.
3. Klatt, D "Review of text-to-speech conversion for english", Journal of the Acoustical Society of America, vol 82, n. 3, Sept 1987, pp. 737-792.

4. Makhoul, J "Linear Prediction: a tutorial review", Proc. IEEE, vol. 63, pp. 561-580, 1975.

5. Markel, J. D. "The SIFT algorithm for Fundamental Frequncy Estimation", IEEE Trans. on Audio and Electroacoustics, Dec. 1972, vol. AU-20, pp. 367-377.

6. Rabiner, L. R.; Schafer, R. W. "Digital Processing of Speech Signals", Prentice-Hall, 1978.

7. Rodríguez-Crespo, M. A.; Escalada, R.; Colás, J "Introducción a la síntesis del habla a partir de texto", Revista Española de Electrónica, Nov 1990, pp. 42-47.

8. Singhal, S; Atal, B. S. "Amplitude-optimization and Pitch Prediction in Multipluse Coders", IEEE Trans. on ASSP, vol. 37, pp. 317-327, March 1989.

9. Varga, A; Fallside, F. "A technique for using multipulse linear predictive speech synthesis in text-to-speech type systems", IEEE Trans. on ASSP, vol. ASSP-35, n. 4, April 1987, pp. 586-587.