Asignación automática de marcas de *pitch* basada en programación dinámica

Francesc Alías Pujol y **Ignasi Iriondo Sanz**
{falias, iriondo}@salleURL.edu
Enginyeria i Arquitectura La Salle, Universitat Ramon Llull
Pg. Bonanova 8, 08022 Barcelona, España

Resumen En este artículo se presenta la implementación y evaluación de un sistema de generación automática de marcas de *pitch*, para el etiquetado de un corpus de voz. El sistema está basado en dos conceptos: la energía de la señal de voz y la programación dinámica [10].

La evaluación es doble: respecto al etiquetado de un corpus de habla continua en catalán y respecto al funcionamiento de la utilidad de Entropic equivalente [12]. Además se ha desarrollado un sistema híbrido (PDEnt), combinando el sistema de Entropic y los bloques de programación dinámica del sistema que se describe en el artículo. Los resultados que se obtienen para los dos sistemas implementados son muy satisfactorios.

1 Introducción

El proceso de generación de voz en sistemas de síntesis concatenativa parte de fragmentos o unidades de voz humana real procedentes de un corpus de voz. Este corpus, además de contener las muestras de voz, debe etiquetarse [1][2] con precisión para obtener la información acústica que permita elegir y procesar las unidades adecuadas para la síntesis del habla.

Una de estas etiquetas son las marcas de *pitch*, que se sitúan siguiendo el período fundamental (T_0) de la señal temporal y son necesarias en las aplicaciones *pitch*-síncronas de procesado de la señal de voz, como por ejemplo *PSOLA*.

Aunque el etiquetado podría realizarse manualmente [8], es imprescindible disponer de un sistema automático que permita obtener las marcas de *pitch* del corpus, sobretodo en el contexto de los grandes corpus de voz que se utilizan en los sistemas de conversión texto voz basados en selección de unidades [3].

2 Descripción del sistema

El sistema automático desarrollado está basado en los conceptos expuestos por V. Goncharoff y P. Gries en la conferencia SIP'98 [7]. Para determinar la posición de las marcas de *pitch*, el sistema desarrollado sólo necesita las muestras de la señal de voz analizada. Existen otros sistemas [1][2] que parten de la señal de excitación glotal, grabada simultáneamente con la voz, procedente de un electroglotógrafo (EGG). En un futuro resultaría interesante evaluar el funcionamiento del sistema desarrollado utilizando, como dato de entrada, la señal glotal que sigue la excitación de la señal emitida.

2.1 Criterios escogidos

Existen distintas posibilidades para determinar dónde deben depositarse las marcas de *pitch*. Dentro del período de señal de voz: algunas técnicas la colocan en el instante de cierre glotal (*CGI*) [4][9][5], otras en la posición de máxima energía positiva [1], otras en el primer paso por cero, etc. A lo largo de toda la señal: algunos sistemas [12] sólo colocan marcas en las zonas sonoras, otros [1][2], en las zonas sordas distribuyen las marcas con una separación arbitraria constante, por ejemplo cada 10 ms, etc.

En el sistema que se ha desarrollado, las marcas *pitch* siguen los siguientes criterios:

- Se colocan tanto en las zonas sonoras como en las sordas o silencios. Así no es necesario un proceso [11][6] que deba distinguir la sonoridad de la señal.
- En las zonas periódicas, se sitúan sobre el pico de amplitud máxima (positiva o negativa) de cada uno de los períodos.
- En las zonas aperiódicas (silencios o tramos sordos), se distribuyen casi-equiespaciadamente (áreas de transición).

2.2 Estructura del sistema

El sistema se desglosa en los tres procesos habituales [6] de un sistema de estas características (figura 1): a)Preprocesado de la señal, que consiste en extraer el contorno de energía de la señal de voz. b)Generación de las marcas iniciales, definido en el bloque de estimación de la periodicidad. c)Postprocesado, que mediante el bloque de estimación de la 'fase', ajusta la posición final de las marcas de *pitch*.

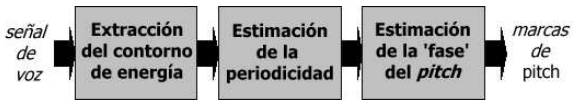

Figura 1: *Diagrama de bloques del sistema de generación automática de marcas de* pitch.

2.2.1 Extracción del contorno de energía

Se aplica un filtrado paso bajo a la señal analizada para poder obtener un pico de energía sobre cada uno de los períodos de la señal. El contorno de energía (figura 2) se consigue convolucionando las muestras de la señal de voz ($\underline{x}[n]$) al cuadrado con una ventana ($\underline{w}[n]$) de 12.8 ms, obtenida de la convolución de dos ventanas *hanning* de 6.4 ms, para mejorar su selectividad:

$$\underline{c_e}[n] = \underline{x}^2[n] * \underline{w}[n] \qquad (1)$$

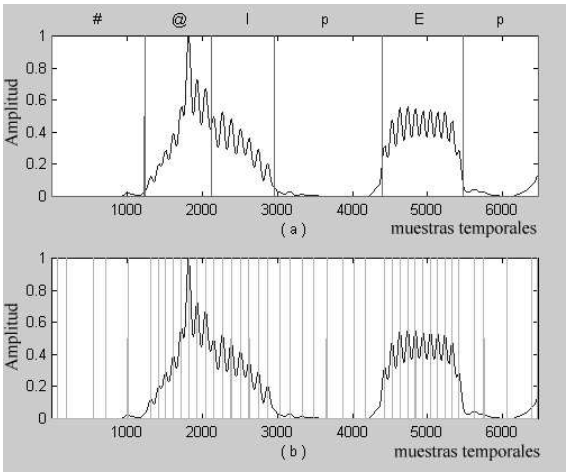

Figura 2: *(a) Contorno de energía de la señal /@lpEp/ (notación SAMPA). (b) Sobre el contorno se indica una primera estimación de las marcas de* pitch.

2.2.2 Estimación de la periodicidad

Tal y como se indica en la figura 2.b, sobre el contorno de energía se puede realizar una estimación inicial de la posición de las marcas de *pitch*. Esta primera aproximación presenta distintos problemas: posición arbitraria de las marcas en las zonas aperiódicas y duplicación (pico residual) u omisión (emmascaramiento) de marcas en las zonas sonoras. Para corregir estos defectos y aprovechar esta estimación inicial, se genera una matriz de energía ($\underline{\underline{E}}$) sobre la que se aplicará el algoritmo de programación dinámica denominado *Dynamic Programming Path Finding Algorithm*.

Obtención de la matriz de energía En primer lugar, se divide el contorno de energía en T tramas de corta duración. En segundo lugar, se genera una estructura en dos dimensiones (*pitch* respecto a tiempo), definida como la matriz de energía $\underline{\underline{E}}[p_{max} \times T]$.

Para cada marca (m_i) obtenida de $\underline{c_e}[n]$ se realiza una doble estimación de su periodicidad (para absorber posibles picos espúreos), calculando la separación respecto a las marcas anterior y posterior. En la fila indicada por esta diferencia y en la columna indicada por el índice de la trama a la que pertenece la marca estudiada (t_i), se introduce el valor del pico de energía (e_i) anterior y posterior, respectivamente:

$$\begin{aligned} if\,(p_{min} \le (m_i - m_{i-1}) \le p_{max}) \\ \underline{\underline{E}}(m_i - m_{i-1}, t_i) = e_{i-1} \\ if\,(p_{min} \le (m_{i+1} - m_i) \le p_{max}) \\ \underline{\underline{E}}(m_{i+1} - m_i, t_i) = e_{i+1} \end{aligned} \qquad (2)$$

Si se observa este algoritmo, se puede comprobar que el proceso de estimación de la periodicidad está restringido a un determinado margen de valores de frecuencia fundamental: $(F_{min} = 60Hz) \le F_0 \le (F_{max} = 400Hz)$.

Dynamic Programming Path Finding Algorithm Sobre la matriz $\underline{\underline{E}}$ se aplica un algoritmo de programación dinámica [10] que tiene como objetivo determinar cuál es la mejor distribución de las marcas de *pitch*. El camino óptimo (p) será aquel que presente una energía acumulada máxima (ecuación 3), pasando por el mayor número de picos de

energía elevada.

$$E_{total} = \sum_{i=1}^{T} \underline{\underline{E}}\big(p(i), i\big) \qquad (3)$$

Este algoritmo se caracteriza por restringir los caminos de la estructura *trellis* (generada en el proceso *forward*), según una pendiente máxima (ecuación 4); definiendo, así, la variación máxima de *pitch* permitida entre períodos consecutivos (figura 3).

$$\big|p(i) - p(i-1)\big| \leq S_{max} \qquad (4)$$

Este valor de pendiente máxima depende de la frecuencia de muestreo (f_s) utilizada. Según [7], para $f_s = 8000\ muestras/s$, se escoge $S_{max} = 3$; en cambio, para $f_s = 16000\ muestras/s$ (usada en el sistema), se ha escogido una $S_{max} = 7$, después de haber realizado un barrido de $S_{max} = 1:9$. Como se puede observar, existe una proporcionalidad 2:1, aproximadamente, entre las parejas de valores.

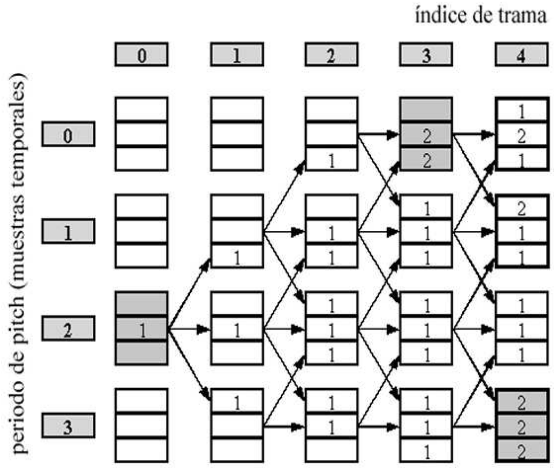

Figura 3: *Estructura trellis restringida por una $S_{max} = 1$ sobre una matriz teórica de energía $\underline{\underline{E}}$, en la cual el cero indica baja energía y el uno alta energía (los unos se representan mediante las casillas sombreadas).*

A continuación se aplica el proceso de *backtracking*, que es el encargado de escoger las casillas de la estructura *trellis* que conforman el camino óptimo (figura 4) a lo largo de la matriz de energía. Se obtiene de este proceso el valor de *pitch*, en muestras temporales, para cada una de las tramas de análisis de la señal de voz.

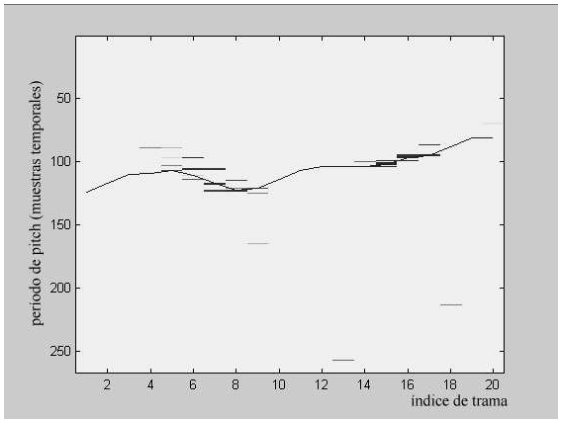

Figura 4: *Representación de la matriz de energía $\underline{\underline{E}}$ sobre la que se dispone el camino óptimo obtenido. Las zonas más oscuras indican picos de energía elevada.*

2.2.3 Estimación de la 'fase' del *pitch*

El último bloque del proceso se encarga de determinar la posición final de las marcas de *pitch*, cosa que puede ser interpretada como el ajuste de la 'fase' temporal de la periodicidad.

Después de obtener la curva de *pitch* para toda la señal (Nmuestras) mediante un proceso de interpolación, se genera la distribución inicial de las marcas de *pitch*. Se parte de una marca inicial, fijada arbitrariamente en la muestra $n = 1$ (referencia), y se van distribuyendo espacialmente todas las marcas siguiendo los valores de *pitch* que se extraen de la curva temporal:

$$\begin{aligned} &n = 1; \\ &while\ (n \leq (Nmuestras) \\ &\quad array\_marcas[n] = 1; \\ &\quad n+ = curva\_pitch[n]; \\ &end \end{aligned} \qquad (5)$$

El vector $array\_marcas$ será todo nulo menos en las posiciones que contengan las marcas de *pitch*, de valor igual a la unidad. De este modo se consigue obtener una marca sobre cada uno de los períodos de la señal de voz (figura 5), aunque ésta debe de acabar de alinearse con el máximo de energía de los períodos de la señal.

Para conseguir colocar cada marca de *pitch* sobre el máximo del período de señal de voz al que corresponde, se vuelva a recurrir al algortimo *Dynamic Programming Paht Finding*. Esta vez el dato de entrada es una ma-

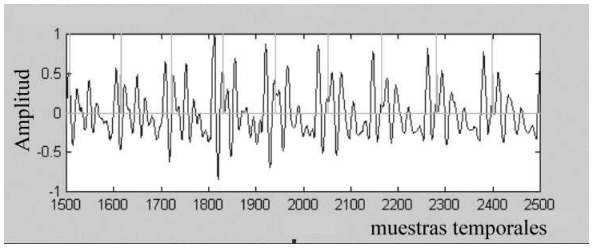

Figura 5: *Zona sonora de una señal de voz sobre la que se ha superpuesto la primera estimación de las marcas de* pitch.

triz de tramas de señal en valor absoluto ($\underline{\underline{S}}$) ponderadas por una ventana *hanning*. Estas tramas están centradas en las posiciones indicadas por el vector $array\_marcas$ y tienen una duración de $2 \times p_{max}$, para abarcar la desviación máxima permitida. Se aplica el algoritmo descrito en 2.2.2 con una $S_{max} = 9$ (según [7], para $f_s = 8000\ muestras/s$, se escoge $S_{max} = 4$).

El resultado que nos aporta el camino óptimo obtenido se interpreta como el *offset* (figura 6) que se debe aplicar a las marcas de *pitch* para ajustarlas a sus posiciones finales, es decir, la desviación del camino respecto a los máximos de la señal de voz.

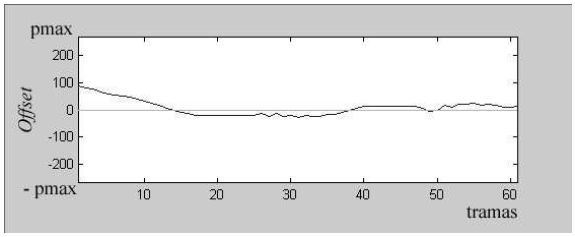

Figura 6: *Valor del offset (en muestras temporales) para cada una de las marcas de* pitch.

3 Estudios y resultados obtenidos

Para poder evaluar el funcionamiento del sistema se trabaja con un corpus de voz de habla continua en catalán etiquetado (134 frases con 5.500 unidades), que ha sido marcado mediante un proceso semiautomático revisado manualmente.

En los estudios que se presentan a continuación, el proceso de verificación no se realiza a nivel de desviación entre marcas, porque los sistemas no siguen los mismos criterios de disposición de marcas dentro del período de la señal de voz. El criterio escogido se basa en la comparación a nivel del *pitch* medio: a partir de las marcas de segmentación[1], se determinan las marcas de *pitch* que corresponden a cada unidad y, según su distribución temporal, se calcula su F_0 media. Este proceso se realizará para todas las unidades sonoras del corpus (74.9% del total), dejando de lado las sordas o silencios, ya que las marcas de referencia no siguen una pauta concreta.

Los resultados de los estudios se presentan en tanto por ciento de acierto respecto al total de unidades estudiadas, según una determinada desviación (D en las tablas) entre el *pitch* medio obtenido por el sistema analizado respecto al de referencia.

3.1 Estudio 1

En este primer estudio, se analiza $\underline{c_e}\,[n]$ en tramas de 40 ms [7] y solapamiento del 50%. Con estos valores y los descritos durante los puntos anteriores (S_{max}), se obtienen los siguientes resultados para todas las unidades sonoras:

D [Hz]	≤ 3	≤ 5	≤ 7	≤ 10
Acierto	77.5%	83.4%	85.9%	88.1%

Tabla 1: Resultados totales de la desviación del *pitch* medio de las unidades sonoras del corpus.

Estudiando detenidamente las unidades que presentan más errores, se ha podido observar que el grupo de las unidades fricativas sonoras (4.7% de las sonoras) es especialmente crítico, debido a que las marcas de referencia para estas unidades no están bien condicionadas: presentan variaciones bruscas respecto al *pitch* medio de las unidades vecinas. Por este motivo, se realiza un nuevo test (tabla 2) dejando de lado estos fonemas. Los resultados mejoran considerablemente, del orden de un 3%.

D [Hz]	≤ 3	≤ 5	≤ 7	≤ 10
Acierto	80.5%	86.5%	88.9%	90.9%

Tabla 2: Resultados totales de la desviación del *pitch* medio sin las unidades fricativas sonoras.

3.2 Estudio 2

Una vez determinado el conjunto de fonemas válidos como referencia de los tests a realizar, se intentan mejorar los resultados presentados en la tabla 2. Además del problema comentado de los fonemas fricativos sonoros,

[1] Delimitan la posición de las unidades (fonemas) dentro de la señal de voz

se ha podido comprobar que el sistema presenta dificultades en los fonemas oclusivos y aproximantes sonoros. Estos fonemas presentan niveles de energía menores respecto a sus vecinos (vocales, normalmente). Por ello, su periodicidad tendrá poco peso en la elección del camino óptimo, sobretodo al trabajar con ventanas 'grandes'. En este caso, en una misma columna de $\underline{\underline{E}}$ coexisten picos de elevada energía (de vocales, por ejemplo) y de menor energía (de las oclusivas sonoras, por ejemplo). Esto, que es la base de la distribución casi-equispaciada de las marcas en las zonas sordas, provoca que este conjunto de fonemas se conviertan también en zonas de transición entre fonemas periódicos de mayor potencia. Por lo tanto, su periodicidad no quedará bien condicionada.

Por todo lo expuesto, se han realizado un conjunto de tests con ventanas de distinto tamaño (tabla 3), manteniendo para todas ellas un 50% de solapamiento relativo a su longitud, en busca de una mejora de los resultados obtenidos.

D [Hz]	≤ 3	≤ 5	≤ 7	≤ 10
40 ms	80.5%	86.5%	88.9%	90.9%
30 ms	82.3%	88.6%	91%	93.3%
20 ms	84.4%	91.1%	93.6%	95.5%

Tabla 3: Acierto en tanto % para distintos tamaños de ventanas de análisis.

De esta tabla se puede observar que, a medida que se disminuye el tamaño de la ventana de análisis, los resultados mejoran de forma considerable. Por otro lado, dentro del mismo contexto de mejora de la resolución del sistema, también se ha estudiado la importancia del grado de solapamiento entre ventanas consecutivas, manteniendo un mismo tamaño de ventana de análisis (30 ms, para la tabla 4).

D [Hz]	≤ 3	≤ 5	≤ 7	≤ 10
0 %	73.9%	80.4%	83.2%	85.8%
33 %	80.7%	86.7%	89.1%	91%
50 %	82.3%	88.6%	91%	93.3%
66 %	84.7%	91.3%	93.8%	95.6%
75 %	85.8%	91.4%	93.7%	95.8%
80 %	86.4%	92.1%	94.4%	96.3%
83 %	86.5%	92.7%	94.9%	96.7%

Tabla 4: Acierto (en %) para distintos grados de solapamiento con una ventana de análisis de 30 ms.

Tanto el estudio presentado en la tabla 3 como el de la tabla 4, se ha detenido cuando los resultados ya no mejoraban o empezaban a empeorar. El etiquetado de los fonemas sonoros de poca energía, es el causante principal de la mejora en los resultados. De la observación de ambas tablas, se llega a la conclusión que en el proceso de análisis del contorno de energía se debe utilizar una ventana de dimensiones reducidas ($\sim 20ms$) y un alto grado de solapamiento ($\sim 80\%$).

En la tabla 5 se presentan los resultados finales para determinar la configuración óptima del sistema. Se estudian tres posibilidades: (a) 12.8 ms (tamaño del pico de energía) y 83% solapamiento (*step*=2.1 ms), (b) 15 ms 83% solapamiento (*step*=2.5 ms) y (c) 20 ms y 87% de solapamiento (*step*=2.5 ms).

D [Hz]	≤ 3	≤ 5	≤ 7	≤ 10
(a)	87.5%	93.5%	96.1%	97.8%
(b)	87.3%	93.4%	96.2%	98.1%
(c)	87.3%	93.4%	96.2%	98.1%

Tabla 5: Grado de acierto (en %) para distintas configuraciones de ventana de análisis.

Prácticamente las tres configuraciones presentan los mismos resultados y costes computacionales bastante similares. De ellas se escoge trabajar con la configuración de 20ms y *step* 2.5ms, que es la que más se ajusta al compromiso que existe entre la extensión de la periodicidad hacia las zonas aperiódicas y la resolución en las zonas sonoras de menor energía.

3.3 Estudio 3

Para realizar una segunda evaluación del sistema, se ha marcado el corpus de 134 frases mediante otro sistema de etiquetado de marcas de *pitch*. Este sistema está incluido en el paquete informático Entropic y también trabaja sobre las muestras de la señal de voz [12]. Es un método muy utilizado y es un claro representante del 'estado del arte' en este tipo de sistemas.

D [Hz]	≤ 3	≤ 5	≤ 7	≤ 10
Entropic	86.8%	94.5%	96.4%	97.7%
C. ópt.	87.3%	93.4%	96.2%	98.1%

Tabla 6: Comparativa entre el acierto (en %) de la configuración óptima del sistema descrito respecto al método desarrollado por Entropic.

En la tabla 6 se presentan los resultados del sistema desarrollado, utilizando la configuración óptima (C. ópt., en la tabla) descrita en el apartado anteior, respecto a los resultados del etiquetado mediante Entropic. Se puede observar que ambos métodos presentan resultados muy similares.

3.4 Estudio 4

Otra posibilidad que se ha estudiado ha sido la de combinar ambas técnicas, es decir, se suprime el primer bloque del sistema desarrollado (extracción del contorno de energía y de sus máximos) por el sistema de Entropic y a continuación se aplica el resto del sistema como postprocesado del mismo. La idea consiste en partir de una buena estimación de las marcas de *pitch* y utilizar la programación dinámica para extenderlas a lo largo de toda la señal de voz. De este modo las marcas de *pitch* en las zonas sordas mantendrán una mejor consistencia respecto a sus vecinas que si se colocan arbitrariamente [1][2].

En este nuevo proceso (PDEnt) aparece un problema: no se dispone del nivel de energía de cada pico para ser introducido en la matriz $\underline{\underline{E}}$. Por lo tanto, después de aplicar el algoritmo (2) la matriz sólo contendrá casillas con ceros y unos, cosa que equivale a ponderar todas las marcas por igual. En este contexto, al finalizar el proceso *forward* pueden aparecer distintas casillas que presenten una 'energía total' acumulada idéntica (en la figura 3 se presentan tres posibilidades), habiéndose de realizar el proceso de *backtracking* para todas ellas. Se escogerá el camino óptimo que: a) pase por el mayor número de 'máximos de energía' (siga mejor las marcas de Entropic) y presente una variación global menor (evolución suave de la curva de *pitch*).

Después de realizar un conjunto de tests similares a los expuestos en el punto 3.2, se obtiene la configuración óptima para 20ms y *step* 5ms (75% solapamiento). En la tabla 7 se presentan los resultados comparando los tres sistemas. Se puede observar que el sistema híbrido también presenta unos resultados muy buenos y que puede ser también utilizado para etiquetar un corpus de voz.

Se ha estudiado el sistema híbrido tanto con ajuste de 'fase' (tercer bloque del sistema (figura 1)) como sin él. Para ventanas de análisis equirvalentes, los resultados son muy parecidos. Se escoge trabajar con ajuste de 'fase' porque, aunque las marcas de Entropic vean desplazada su posición original dentro del período, estas siguen un criterio consistente a lo largo de toda la señal; en cambio, si no se aplica el *offset* temporal, las marcas quedan colocadas en los períodos sin un criterio claro, aunque siguen correctamente su periodicidad.

D [Hz]	≤ 3	≤ 5	≤ 7	≤ 10
C. ópt.	87.3%	93.4%	96.2%	98.1%
Entropic	86.8%	94.5%	96.4%	97.7%
PDEnt	87.8%	94.3%	96.6%	98.1%

Tabla 7: Comparativa entre el acierto (en %) entre los tres sistemas evaluados: Entropic, la configuración óptima del sistema desarrollado (C. ópt.) y el sistema híbrido (PDEnt).

4 Conclusiones

Se ha desarrollado un sistema de generación automática de marcas de *pitch* basado en la programación dinámica. Además se ha realizado un estudio exhaustivo de los distintos parámetros que intervienen en su funcionamiento hasta obtener una configuración óptima de trabajo. También ha sido verificado respecto a una corpus etiquetado y al funcionamiento del sistema Entropic, verificándose sus buenos resultados.

Finalmente, se desarrollado un nuevo sistema híbrido (PDEnt) que presenta unos resultados muy satisfactorios, igualando o mejorando los resultados de los dos métodos individuales. Además, este sistema presenta la ventaja adicional de distribuir las marcas de *pitch* en las zonas sordas (Entropic no coloca marcas) como adaptación entre las periodicidades vecinas. De este modo se mejora el etiquetado de estos fonemas y se refina el funcionamiento de los sistemas de síntesis *pitch*-síncronos.

Agradecimientos Este trabajo se ha realizado con el apoyo del Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya mediante la beca 2000FI-00679 del DOGC 07/02/01. También quiero agradecer al Dr. Antonio Bonafonte de la Universitat Politècnica de Catalunya su colaboración en este trabajo, tanto por el etiquetado del corpus de voz mediante la utilidad de Entropic como por las ideas sugeridas.

Referencias

[1] A. W. Black and P. Taylor. The festival speech synthesis system: system documentation. Technical report, Technical Report HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK, 1997.

[2] A. W. Black and P. Taylor. Chatr: A generic speech synthesis system. *Proceedings of COLING-94*, II:983–986, Japan,1997.

[3] A. W. Black and P. Taylor. Automatically clustering similar units for unit selection in speech synthesis. *uropean Conference on Speech Communication and Technology*, pages 601–604, Rodes, Greece, 1997.

[4] O. Boeffard and F. Violaro. Improving the robustness of text-to-speech synthesizers for large prosodic variations. *Conf. Proc. of second ESCA-IEEE Workshop on Speech Synthesis*, pages 111–114, New Paltz, USA, 1994.

[5] G. Cohen and D. Malah. Speech analysis and synthesis using a glottal excited ar model with dtw-based glottal determination. *IEEE 18th Conv. of EE in Israel*, pages 3.2.3–1–3.2.3–5, Tel-Aviv, 1995.

[6] J. Droppo and A. Acero. Maximum a posteriori pitch tracking. *Proc. ICSLP'98*, pages 943–946, 1998.

[7] V. Goncharoff and P. Gries. An algorithm for accurately marking pitch pulses in speech signals. *Proc. IASTED Intern. Conference. Signal and Image Processing*, pages 281–284, Las Vegas,1998.

[8] Lee Tan Lo, W. K. and P. C. Ching. Development of cantonese spoken language corpora for speech processing. *Proceedings of 1998 International Symposium on Chinese Spoken Language Processing*, pages 102–107, Singapore,1998.

[9] T. V. Ngoc and C. d'Alessandro. Robust glottal closure detection using the wavelet transform. *Proc. of the European Conference on Speech Technology, EuroSpeech*, pages 2805–2808, Budapest, 1999.

[10] L. R. Rabiner and Juang B-H. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[11] Y. Stylianou. A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech. *IEEE Nordic Signal Processing Symposium*, 1996.

[12] D. Talkin. *A Robust Algorithm for Pitch Tracking (RAPT).* Kleijn, W. B. and Paliwal, K. K. (Eds.), New York, 1995.