Pruebas y Validación de un Sistema de Reconocimiento del Habla Basado en Sílabas con un Vocabulario Pequeño.

Dr. Sergio Suárez Guerra[1], M. en C. José Luis Oropeza Rodríguez[2], Ing. Mariana Del Villar[3] Ing. Karen Suso[4]

j_orope2002@hotmail.com

1. Profesor Investigador del Centro de Investigación en Computación (CIC, IPN)
2. Alumno de doctorado del Centro de Investigación en Computación (CIC, IPN)
3. Ingeniera en Telemática (UPIITA, IPN)
4. Ingeniera en telemática (UPIITA, IPN)

CIC Av. Juan de Dios Batís s/n casi esq. Miguel Otón de Mendizábal, Unidad Profesional "Adolfo López Mateos" Edificio CIC. Col. Nueva Industrial Vallejo, 07738, México, D.F.

UPIITA Av. Instituto Politécnico Nacional 2580, Col. La Laguna Ticomán, 07034, México D.F.

RESUMEN.

El presente trabajo contempla los resultados obtenidos del análisis de un corpus de voz discontinuo segmentado en unidades silábicas. Los resultados contemplan la utilización de técnicas de reconocimiento tales como la codificación predictiva lineal y los modelos ocultos de Markov.

Se realizó un análisis a un corpus de 20 palabras, las cuales son: *cero, uno, dos, tres, cuatro, cinco, seis, siete, ocho, nueve, azul, blanco, café, gris, negro, rojo, rosa, verde, abrir y cerrar,* por medio de división silábica y basado en los Modelos Ocultos de Markov.

Dichas palabras generaron 31 sílabas como elementos de reconocimiento, las cuales nos permitieron crear un determinado conjunto de patrones (modelos) para ser reconocidos posteriormente.

PALABRAS CLAVE: Viterbi, HMM, LPC, sílaba.

INTRODUCCIÓN.

El reconocimiento de voz es una rama de las ciencias de la computación que se encuentra aún en proceso de crecimiento. Los distintos elementos que la conforman han llegado a generar distintas herramientas prácticas que plasman tales avances científicos. Sin embargo, la búsqueda de nuevas alternativas como sucede en otras ramas de la ciencia es de carácter relevante. Las sílabas conforman un patrón de estudio bastante útil en esta rama, el presente trabajo muestra tales hechos. En la figura 1 se plantea la solución que se le da al problema del reconocimiento de voz usando cadenas ocultas de Markov[1]. El cual se inicia introduciendo la señal de voz por medio de un micrófono. En la siguiente etapa se segmenta en sílabas la señal obtenida, cabe destacar que las etapas posteriores se le realizarán a cada una de las sílabas que componen la palabra introducida. El procesamiento digital de la señal (PDS) se realiza para reducir la cantidad de información que se debe de procesar, extrayendo únicamente las características acústicas sobresalientes de las sílabas.

El reconocimiento del habla se realiza con una etapa previa de entrenamiento[2], la cual es la encargada de generar un libro código global[3] con las muestras adquiridas, utilizando el método de Codificación Predictiva Lineal, después con este libro código se generan las Cadenas Ocultas de Markov. Una vez realizado lo anterior, se pueden reconocer las palabras contenidas en el diccionario.

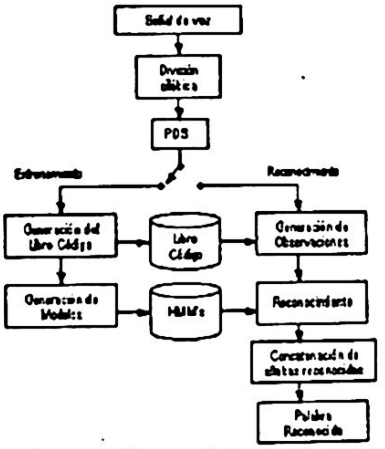

Fig 1.Diagrama a bloques de la solución propuesta.

Tal procesamiento de información es obtenido tras la solución a los tres problemas de Markov, donde el primero de ellos (obtención de la probabilidad de que una observación haya sido producida por un modelo) es de la forma[1]:

J. Díaz de León, G. González, J. Figueroa (Eds.): Avances en Ciencias de la Computación, pp. 206-211, 2003.

1.- INICIALITACION

$$\alpha_1(j) = \pi_j b_j(o_1), \quad 1 \le j \le N$$

2.- INDUCCION

$$\alpha_{t+1}(j) = b_j(o_{t+1}) \sum_{i=1}^{N} \alpha_t(i) a_{ij}, \quad 1 \le j \le N, \quad 1 \le t \le T-1$$

3.- TERMINACION

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

(1)

El problema 2 (obtención de la mejor trayectoria) hace uso del algoritmo de Viterbi, de tal manera que:

1. INICIALIZACION

$$\delta_1(j) = \pi_j b_j(o_1), \quad 1 \le j \le N$$

2. RECURSION

$$\delta_{t+1}(j) = b_j(o_{t+1}) \left[\max_{1 \le i \le N} \delta_t(i) a_{ij} \right], \quad 1 \le j \le N, \quad 1 \le t \le$$

3. TERMINACION

$$P^* = \max_{1 \le i \le N} [\delta_T(i)] \qquad \text{para } 1 \le i \le N$$

$$q_T^* = \arg \max_{1 \le i \le N} [\delta_T(i)] \qquad \text{para } 1 \le i \le N$$

(2)

Para terminar, el problema 3 (que se encarga de optimizar los parámetros del modelo) es de la forma:

$$a_{ij} = \frac{\text{número esperado de transiciones del estado } q_i \text{ al estado } q_j}{\text{número esperado de transiciones del estado } q_i}$$

$$b_j(k) = \frac{\text{número esperado de veces que estando en el estado } j \text{ el símbolo observado es } v_k}{\text{número esperado de veces que estuvo en el estado } j}$$

(3)

En la figura 2 se muestra un diagrama más detallado del reconocimiento.

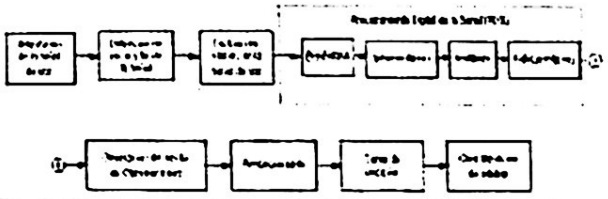

Fig. 2. Diagrama a bloques del reconocimiento de una palabra.

El procesamiento digital es el tratamiento realizado a la señal de voz para que sea más fácil realizar operaciones computacionales, además de que es más exacto que el analógico.

Con base en los vectores de autocorrelación determinados en el procesamiento digital de señales (PDS) y el libro código global obtenido en el entrenamiento, se genera el vector de observaciones.

El reconocimiento se realiza con el vector de observaciones de la sílaba a reconocer y las Cadenas Ocultas de Markov de las sílabas del diccionario generadas en el entrenamiento, con cada uno de ellos obtendremos una probabilidad, la máxima será la que indica que modelo es el que se apega mejor a la secuencia, de esta manera se toma la decisión.

Una vez que se reconozca cada una de las sílabas que componen la señal de voz introducida, se procede a concatenarlas para así construir la palabra que el sistema determina que fue pronunciada.
En la figura 3 se muestra el diagrama a bloques de la etapa de entrenamiento del sistema[4].

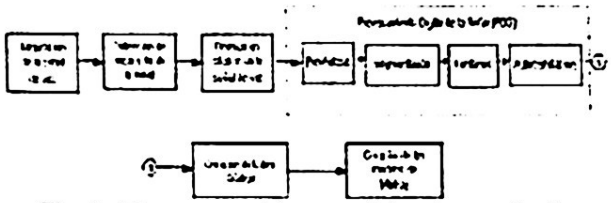

Fig. 3. Diagrama a bloques del entrenamiento

Después del bloque de división silábica de la señal de voz, a cada una de las sílabas encontradas se les realizan las operaciones de los bloques siguientes. Se lleva a cabo el PDS, encontrando los vectores de autocorrelación con base en estos se extraen los coeficientes LPC, después se genera el libro código global y por último se crea la cadena oculta de Markov para la sílaba que estamos procesando[5,6].

II. DESARRROLLO.

El presente trabajo cumple con las siguientes características:

- Es capaz de reconocer los diez dígitos, los siguientes colores: azul, blanco, café, gris, negro, rojo, rosa y verde, así como los comandos: abrir y cerrar.
- El reconocimiento esta basado en sílabas y realizado con Cadenas Ocultas de Markov.

En el presente documento: con la ayuda de un software desarrollado, se realizó el reconocimiento por división silábica y por palabra, para poder evaluar en cual se presentaba un mejor desempeño.
El menú de gráficas incluye un submenú, en el cual el usuario puede elegir si desea ver gráficamente la señal de voz, la señal de energía o la división silábica que realizó el sistema, la siguiente figura 4 ilustra esto:

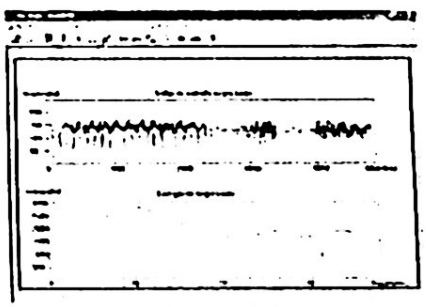

Fig. 4. Gráficas de señal de voz y energía.

Se puede elegir la opción de "ver" y se visualizarán las gráficas de las sílabas que componen la señal de

voz original, se mostrará la caja de diálogo de la figura 5.

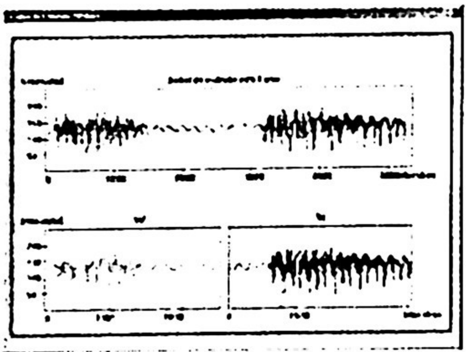

Fig. 5. Gráficas de sílabas.

III. PRUEBAS Y RESULTADOS.

En la tabla 1 se muestra el promedio de duración, en segundos, y el número de muestras que contiene cada palabra del diccionario., cabe destacar que se usaron 10 repeticiones por palabra del diccionario.

Promedio de duración de cada palabra		
Palabra	Tiempo (seg)	# Muestras
[illegible]	0.396	44[illegible]
Dos	0.432	45474
Seis	0.44	49372
tres	0.445	5047
cero	0.502	50346
uno	0.572	66322
nueve	0.596	66318
siete	0.626	69392
cinco	0.64	71356
verde	0.649	72[illegible]
café	0.652	73024
rosa	0.669	73652
rojo	0.642	73339
negro	0.662	[illegible]
[illegible]	0.676	75324
cuatro	0.677	75694
azul	0.677	766[illegible]
[illegible]	0.720	80482
[illegible]	0.764	83128
[illegible]	0.766	84218

Tabla1. Promedio de duración de las palabras del diccionario.

En la figura 6 se muestra la gráfica del promedio de duración de las palabras del diccionario. En ésta se puede ver claramente que estos parámetros (la duración y el número de muestras) se pueden utilizar para este proceso.

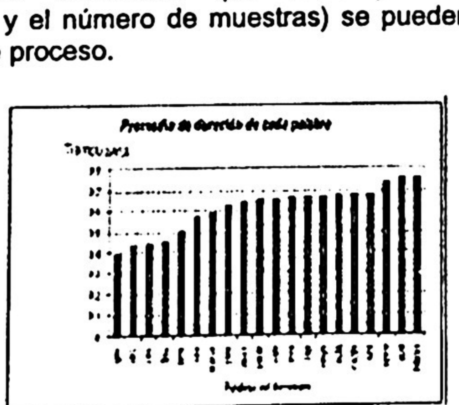

Fig. 6. Gráfica del promedio de duración de cada palabra.

Distribución de regiones

Sabemos que dado un libro código se puede partir un espacio muestra X en N regiones disjuntas y a cada una de estas se les asocia un centroide. Éste optimiza para convertirlo en el mejor representante de la región, el que en promedio difiere menos de cada uno de los elementos de la región[3].

Podemos observar la distribución de vectores en figura 7, los cuales en conjunto forman 4 regiones, su centroide fue optimizado 7 veces, a medida que los vectores cambiaban de región, se obtenía nuevo centroide, y había que recalcularlo, por ello observamos variaciones al principio de la gráfica, pero también observamos que se vuelve estable, pues el centroide ya no cambia en las últimas optimizaciones y por consiguiente, ya no hay saltos de vectores hacia otras regiones.

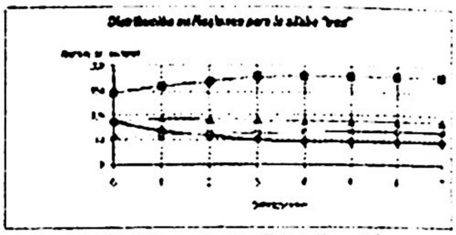

Fig. 7. Gráfica de la Distribución en regiones de la sílaba "tres" para 4 Regiones.

En la figura 8 se observa el mismo proceso, con diferencia que ahora se están modelando 16 regiones y la estabilidad se logró hasta la optimización número 12.

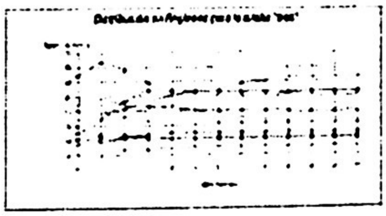

Fig. 8. Gráfica de la Distribución en regiones de la sílaba "tres" para 16 Regiones.

En las gráficas se puede observar que el desempeño del sistema fue el esperado porque fue distribuyendo los datos hasta lograr el centroide óptimo, sin adecuada distribución no se podrían obtener coeficientes LPC deseados para lograr reconocimiento. Cabe destacar que el proceso realizó hasta 128 regiones, pero no se muestra debido a la cantidad de datos que se esta manejando, por esta razón sólo mostramos la distribución regiones final, la cual se muestra en la figura 9.

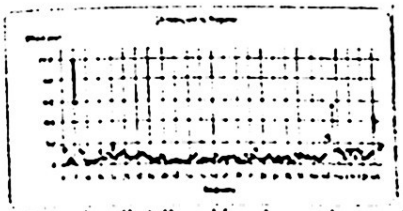

Fig. 9. Gráfica de distribución de regiones final.

En la generación de los Modelos Ocultos de Markov para sílabas se determinó el número de optimizaciones con los que trabaja mejor el sistema, se estuvo evaluando para 32 regiones, con 200 muestras con las cuales fue entrenado el sistema, 2 locutores y como se muestra en la gráfica 10 de la figura el número de errores esta por encima de 45.

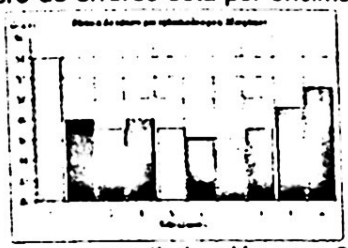

Fig. 10. Errores por optimización para 32 regiones con 200 muestras para sílabas.

En la figura 11 se muestra el desempeño del sistema para 64 regiones las mismas muestras, y como se puede observar, el número de errores ha disminuido, ahora se encuentra por encima de 14.

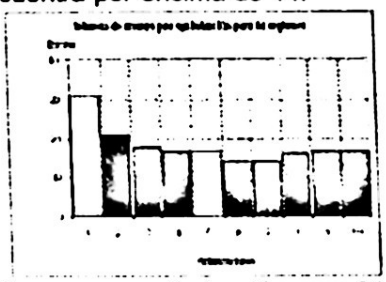

Fig. 11. Errores por optimización para 64 regiones con 200 muestras para sílabas.

Por último, la figura 12 muestra la gráfica del número de regiones elegido, el cual es de 128 pues como se puede observar, el desempeño es mucho mejor que en los casos anteriores, no continuamos incrementando el número de regiones debido al tamaño del diccionario utilizado en el sistema, otra de las razones fue que representaría un incremento innecesario en el tiempo de cómputo.

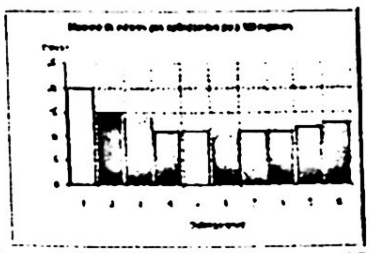

Fig. 12. Errores por optimización para 128 regiones con 200 muestras para sílabas.

Como se puede observar en las gráficas del número de errores por optimización, existe un mejor desempeño en el intervalo de optimizaciones de 5 a 7, porque después de este intervalo el número de errores comienza a incrementarse. Por ésta razón se tomó que el número de optimizaciones en la generación de los Modelos Ocultos de Markov sea de 6.

Se hizo el análisis anterior pero con más locutores para determinar de que manera influye la variabilidad de las distintas señales de voz pues ahora se realizó con 4 locutores y con 400 muestras. En la figura 13 se muestra la gráfica del número de errores por optimización para 32 regiones, como se puede observar el menor error se encuentra en la optimización 6.

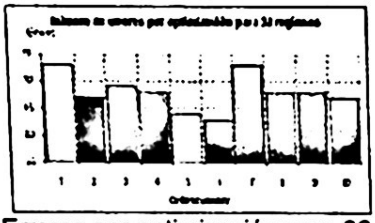

Fig. 13. Errores por optimización para 32 regiones con 400 muestras para sílabas.

En la figura 14 se muestra el desempeño del sistema para 64 regiones con las mismas muestras, y como se puede observar, el número de errores ha disminuido, ahora está por encima de 39.

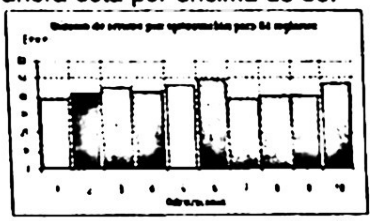

Fig.14. Errores por optimización para 64 regiones con 400 muestras para sílabas.

En la figura 15 se muestra la gráfica con 128 regiones y como se puede observar, el desempeño es mucho mejor que en los casos anteriores, pues con las mismas muestras los errores esta por encima de 20.

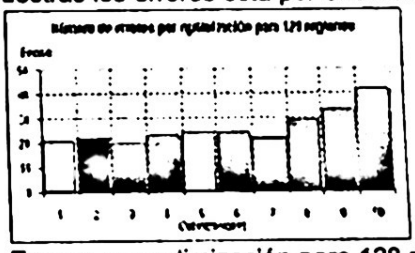

Fig. 15. Errores por optimización para 128 regiones con 400 muestras para sílabas.

A continuación se mostrarán el número de errores por optimización en la generación de los Modelos Ocultos de Markov para el caso de palabras, en la figura 16 se muestra la gráfica para 200 muestras con las

cuales fue entrenado el sistema, 2 locutores y 32 regiones, como se puede observar el menor error se encuentra en la optimización 2.

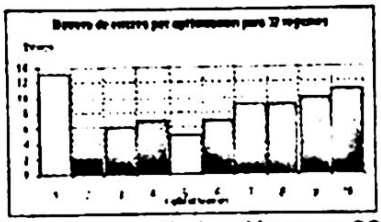

Fig. 16. Errores por optimización para 32 regiones con 200 muestras para palabras.

En la figura 17 se muestra la gráfica para 64 regiones, el menor error está en la optimización 2.

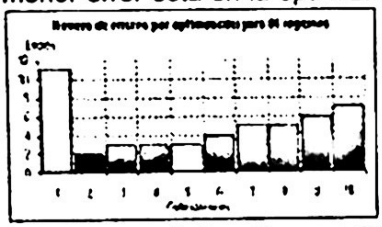

Fig. 17. Errores por optimización para 64 regiones con 200 muestras para palabras.

En la figura 18 se muestra la gráfica para 128 regiones y como se puede observar, el desempeño es mucho mejor que para 32 o para 64 regiones, pues con las mismas muestras el número de errores que se logran es cero para la optimización 2 y 3, y en las otras optimizaciones el error es muy pequeño.

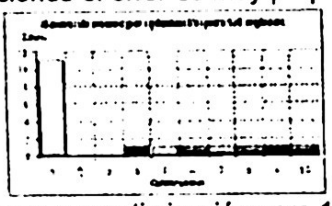

Fig. 18. Errores por optimización para 128 regiones con 200 muestras para palabras.

El análisis anterior se realizó nuevamente pero con más locutores y como consecuencia con un número mayor de muestras, ahora se realizó con 4 locutores, 400 muestras con las cuales fue entrenado el sistema, para 32 regiones, en la figura 19 se muestra la gráfica del número de errores por optimización, el menor error se encuentra en las últimas optimizaciones.

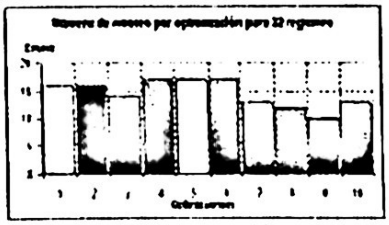

Fig. 19. Errores por optimización para 32 regiones con 400 muestras para palabras.

En la figura 20 se muestra la gráfica para 64 regiones, el menor error está en la optimización 9.

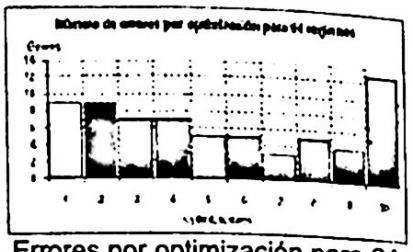

Fig. 20. Errores por optimización para 64 regiones con 400 muestras para palabras.

En la figura 21 se muestra la gráfica con 128 regiones y como se puede observar, el desempeño es mucho mejor que para 32 o para 64 regiones, pues con mismo número de muestras el número de errores está por debajo de 4, en la optimización 7 se logra que sólo exista un error.

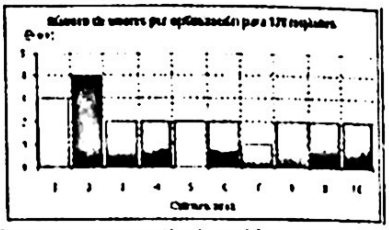

Fig. 21. Errores por optimización para 128 regiones con 400 muestras para palabras.

Los resultados obtenidos con los parámetros elegidos (128 regiones, 6 optimizaciones) y para las 200 muestras utilizadas en el entrenamiento, se pueden observar en la siguiente tabla 2:

Palabra a reconocer	Número de veces que reconoció la sílaba	Número de errores que reconoció la sílaba	Número de veces que reconoció la palabra
cero	10	-	-
uno	10	10	10
dos	10	-	10
tres	10	-	10
cuatro	10	10	10
cinco	10	10	10
seis	10	-	10
siete	10	10	10
ocho	10	10	10
nueve	10	10	10
adi	10	10	10
blanco	10	10	-
cinco	10	10	10
negro	10	-	10
rojo	10	10	10
azul	10	10	10
verde	10	10	10
café	10	10	10
canal	10	10	10

Tabla. 2. Reconocimiento de las palabras del diccionario.

CONCLUSIONES.

Podemos determinar con base en los resultados anteriores que existe un mejor desempeño utilizando reconocimiento por palabras, esto se debe a división silábica que se le realizó a las muestras, ya que en el caso del blanco y el cinco existe un silencio dentro de la primera sílaba que provoca un corte deseado. Además el tiempo de entrenamiento con sílabas es mayor que para el caso de palabras, pero esto no sería así si el vocabulario fuera más grande. Nuestro sistema se realizó mediante sílabas porque se desea en un futuro incrementar el diccionario.

Aunque el sistema es inestable se encontró un punto en donde el índice de reconocimiento fue satisfactorio en un tiempo de cómputo razonable, se realizó una prueba para determinar el número de optimizaciones en donde se alcanza la estabilidad del sistema, la cual consiste en optimizarlo hasta que la probabilidad de una observación dado un modelo (Problema 1 de Markov, algoritmo hacia adelante) no cambie, en dicha prueba se observó que el número de optimizaciones oscila entre 700 y 800 para alcanzar su estabilidad, pero el inconveniente es el tiempo de cómputo ya que se encuentra entre 6 y 7 horas.

5. BIBLIOGRAFIA.

[1] Fundamentals of Speech Recognition.
Lawrence Rabiner and Biing-Hwang Juang.
Prentice Hall, 1993

[2] A Hybrid Systems with Symbolic AI and Statical Methods for Speech Recognition.
Jesus Savage Carmona
University of Washington, 1995.

[3] Reconocimiento de palabras aisladas usando cuantización vectorial.
Ricardo Barrón Fernández.
Centro de Investigación en Computación. Octubre 1998.

[4]Reconocimiento de voz usando estructuras silábicas.
Dr. Sergio Suárez Guerra, José Luis Oropeza
CIC 2002. noviembre 2002

[5] Arquitecturas y métodos en sistemas de reconocimiento automático del habla de gran vocabulario.
Javier Macías Guarasa
Universidad Politécnica de Madrid 2001.

[6] Reconocimiento de voz y fonética acústica. Pedro Gómez Vilda.
Mc Graw Hill, 2000.