

Construcción de Grafos de Fonemas para un sistema de RAH desacoplado

Jon Ander Gómez, María José Castro, Emilio Sanchis

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
{jon,mcastro,esanchis}@dsic.upv.es

Resumen

En este trabajo se presenta una aproximación a la decodificación acústico fonética de un sistema de reconocimiento automático del habla desacoplado, en que la información se transmite en forma de grafos de fonemas, palabras o unidades semánticas.

El módulo de decodificación acústico fonética recibe como entrada una secuencia de vectores con las probabilidades de cada una de las unidades fonéticas que reconoce el sistema, y obtiene como salida un grafo de fonemas como representación de una frase.

En la última sección se presentan resultados experimentales que demuestran la viabilidad de nuestro sistema.

1. Introducción

El Reconocimiento Automático del Habla (RAH) es un campo de la ciencia y de la técnica actualmente muy avanzado, con un esquema para sus sistemas de reconocimiento muy bien definido y ampliamente utilizado. Las diferencias entre los distintos sistemas de RAH radican en las aportaciones particulares de los desarrolladores de estos sistemas. La mayoría de ellos están basados en los modelos de Markov y en la integración en un autómata de los tres niveles de conocimiento: fonético, léxico y sintáctico (y en muchos casos también semántico). La estrategia de reconocimiento en estos sistemas es encontrar el mejor camino o derivación por el autómata dada una secuencia de vectores acústicos de entrada, siendo el algoritmo más utilizado el de Viterbi. Al esquema de estos sistemas de RAH le denominaremos clásico o estándar [1, 2, 3, 4].

Nuestro sistema de RAH tiene una estructura desacoplada, en cada fase se resuelve una parte del problema según el nivel de conocimiento (acústico, fonético, léxico, sintáctico y semántico), y la información se transmite en forma de grafo para no perder información. Existen otros trabajos que utilizan grafos o esquemas parecidos [5, 6, 7].

El módulo que vamos a presentar en este trabajo es el que trabaja a nivel fonético; su cometido es encontrar unidades fonéticas a lo largo de una pronunciación y construir un Grafo de Fonemas.

El grafo de fonemas representa adecuadamente el orden temporal en el que se encuentran las unidades fonéticas, así como el solapamiento de las distintas alternativas que aparecen. Los arcos enlazan los nodos en un solo sentido, de esta manera determinan qué nodos son alcanzables desde uno en particular.

Los grafos son una buena manera de representar el fenómeno del habla porque:

- permiten representar su naturaleza secuencial,

Este trabajo ha sido financiado parcialmente por la CICYT a través del contrato TIC2002-04103-C03-03.

- son capaces de reflejar la incertidumbre al permitir distintos caminos alternativos,
- permiten introducir de forma natural valores de confianza sobre los distintos segmentos.

2. Descripción del Sistema

A nivel global, nuestro sistema de RAH dispone de los siguientes módulos: *preprocesador*, *clasificador fonético*, *generador de grafos de fonemas*, *generador de grafos de palabras* y *reconocedor*.

El módulo *preprocesador*, que no cambia respecto de las aproximaciones estándar, extrae un vector de características acústicas cada cierto intervalo de tiempo; en nuestro caso, la *energía* y los *coeficientes cepstrales* más las primeras y segundas derivadas de todos ellos [1, 3, 4].

El segundo módulo, el *clasificador fonético*, realiza una clasificación en clases naturales a nivel acústico, utilizando para ello técnicas de *clustering* [8]. Las probabilidades de las clases acústicas se combinan con probabilidades condicionales para obtener las probabilidades fonéticas asociadas a cada trama.

El módulo *generador del grafo de fonemas* recibe como entrada la Secuencia de Vectores con las Probabilidades Fonéticas (SVPF) y construye un Grafo de Fonemas (GF). La SVPF representa una pronunciación, y el GF es una nueva representación de la misma pronunciación que sirve como entrada al siguiente módulo.

3. Generación del grafo de fonemas

La estrategia para generar un grafo de fonemas a partir de una SVPF consiste en encontrar segmentos que abarcan un intervalo temporal donde una unidad se considera suficientemente probable. Cada vez que se delimita un segmento se crea un nuevo nodo que se añade al conjunto de nodos del GF.

Los nodos representan las unidades fonéticas detectadas en una pronunciación, y los arcos las transiciones entre los nodos. La información que se guarda en los nodos y en los arcos se detalla en las tablas 1 y 2 respectivamente.

Tabla 1: Información por cada nodo.

ph_n	Unidad fonética del nodo n .
$t_i(n)$	Instante de tiempo donde empieza.
$t_f(n)$	Instante de tiempo donde acaba.
coste	Probabilidad de haber sido localizado en el intervalo de tiempo $[t_i(n), t_f(n)]$.

Tabla 2: Información por cada arco.

predecesor	Nodo del cual parte el arco.
sucesor	Nodo al que apunta el arco.
frontera	Trama donde comienza el nodo sucesor.
ajuste	Ajuste cuando existe separación o solapamiento entre los segmentos.

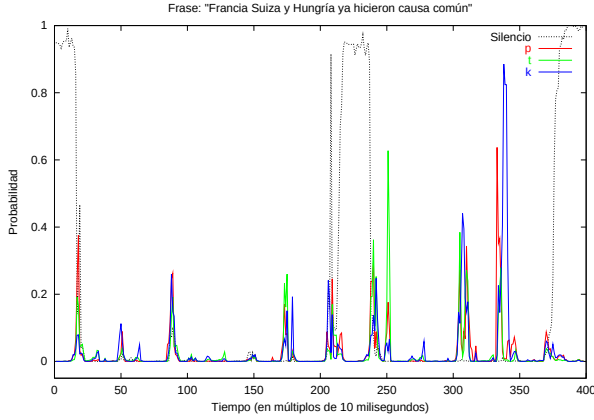


Figura 1: Probabilidad para las unidades [p], [t], [k] y los silencios.

Tras la creación de los nodos se crean los arcos que los conectan según los siguientes criterios:

- i) Mantener el orden temporal de los nodos: Nunca un arco partirá de un nodo hacia otro anterior en el tiempo.
- ii) Evitar la generación de bucles: Para ello se controlan de manera especial los distintos casos de solapamiento entre segmentos, partiendo un segmento en dos trozos cuando sea necesario.
- iii) Permitir que dos segmentos sean conectados aunque no estén solapados, siempre y cuando el lapso de tiempo entre cuando acaba el anterior y cuando comienza el posterior no exceda dos veces la duración media de un fonema (aproximadamente 100 ms).

3.1. Detección de unidades fonéticas

La búsqueda de segmentos se realiza de manera independiente por cada unidad fonética. Consiste en encontrar intervalos dentro de la pronunciación en curso donde la probabilidad de la unidad en cuestión supere cierto umbral. Un ejemplo de cómo evoluciona la probabilidad de las unidades fonéticas a lo largo del tiempo puede verse en las figuras 1 y 2.

La estrategia para detectar los segmentos de cada unidad fonética sigue los siguientes pasos:

- 1) En primer lugar se realiza una detección marcando las tramas que superan cierto umbral $U_{detector}$, obteniendo unos segmentos primarios.
- 2) A continuación dichos segmentos primarios son extendidos por ambos lados si la probabilidad de las tramas de la frontera supera otro umbral $U_{amplificador}$, menos restrictivo que el anterior.

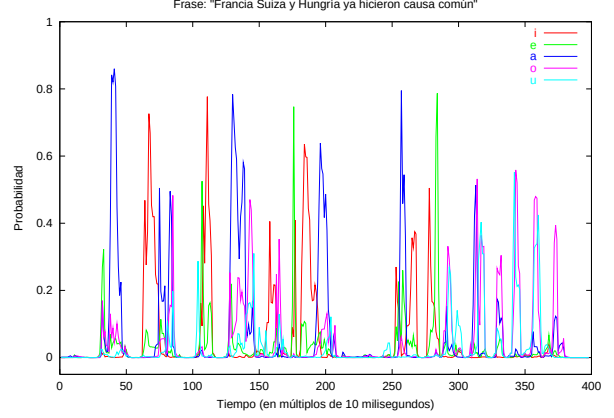


Figura 2: Probabilidad para las vocales del castellano.

- 3) Los segmentos obtenidos ya son candidatos a generar nodos, pero antes se realiza una depuración, eliminando los segmentos demasiado cortos (una trama), y uniendo los cercanos (separados por una trama).

Los umbrales $U_{detector}$ y $U_{amplificador}$ son heurísticos ajustados empíricamente, en este caso están fijados a 10% y 5% respectivamente. El ajuste de $U_{amplificador}$ está íntimamente relacionado con los márgenes de separación permitidos para considerar a dos segmentos consecutivos.

3.2. Creación de los nodos

Una vez detectados los segmentos de todas las unidades fonéticas es necesario estimar la probabilidad acumulada de cada segmento:

$$\text{probabilidad}_{acumulada}(n) = \prod_{\forall t \in [t_i(n), t_f(n)]} P(ph_n | x_t) \quad (1)$$

donde $t_i(n)$ y $t_f(n)$ son los índices inicial y final del segmento tal como se describe en la tabla 1, ph_n representa la unidad fonética, y x_t el vector de características acústicas en el instante t .

Por cuestiones de estabilidad numérica y de eficiencia, internamente trabajamos con logaritmos, por lo tanto el coste de un nodo o segmento que representa una unidad fonética se estima según la siguiente expresión:

$$\text{coste}_{acumulado}(n) = \sum_{t=t_i(n)}^{t_f(n)} -\log P(ph_n | x_t) \quad (2)$$

Con toda esta información se crea un nuevo nodo $(ph_n, t_i(n), t_f(n), \text{coste}_{acumulado}(n))$ que se añade a la lista de nodos del grafo en construcción. El siguiente paso es crear los arcos que enlazan aquellos nodos susceptibles de serlo.

3.3. Creación de los arcos

La generación de los arcos se realiza explorando nodo por nodo, y por cada nodo buscando aquellos candidatos a ser sucesores. Las condiciones son estar relativamente cerca o tener cierto grado de solapamiento. El problema aparece en los casos en que el solapamiento es excesivo, o bien un segmento pequeño correspondiente a una unidad está inmerso en otro más grande de

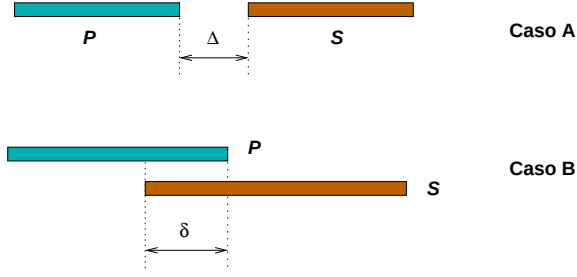


Figura 3: Situaciones típicas de dos segmentos consecutivos.

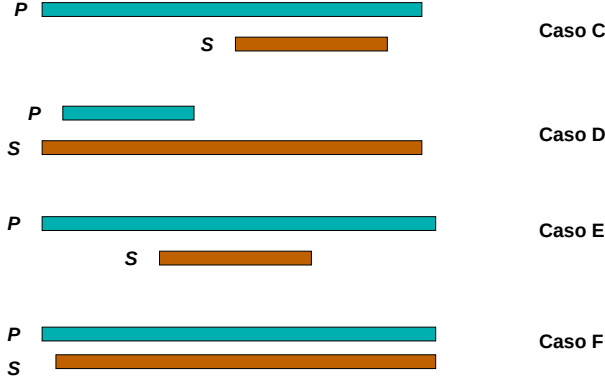


Figura 4: Situaciones críticas de dos segmentos.

otra unidad, o bien ambos son prácticamente igual de grandes. Si estos casos no se tratan de manera especial pueden provocar bucles en el grafo de fonemas, lo que podría llegar a colapsar al siguiente módulo, el GGP.

Los casos que nos podemos encontrar entre cada par de segmentos que pueden considerarse consecutivos son los que aparecen en las figuras 3 y 4, donde el sentido del tiempo es el natural de izquierda a derecha, P se refiere al segmento o nodo predecesor, y S al segmento o nodo sucesor.

Las situaciones A y B son las ideales. El parámetro Δ indica la separación máxima entre dos segmentos para que el sistema los considere consecutivos y cree un arco entre ellos. El parámetro δ denota el solapamiento. Cuando δ es demasiado grande la situación degenera en alguno de los casos que aparecen en la figura 4.

El caso F también contempla que el segmento S acabe antes o después del segmento P . Es el caso donde el solapamiento δ supera el 70% de la extensión que ocupan los dos segmentos.

En la tabla 3 se resumen las distintas soluciones que se adoptan para evitar bucles.

Tabla 3: Solución respecto de los diferentes casos.

Caso	Solución
A	$P \rightarrow S$
B	$P \rightarrow S$
C	$P \rightarrow S$
D	$P \rightarrow S$
E	$P1 \rightarrow S, S \rightarrow P2$
F	$P1 \rightarrow S2, S1 \rightarrow P2$

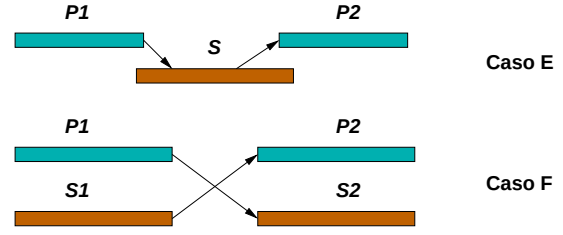


Figura 5: Soluciones a los casos E y F.

Para los casos A, B, C y D se busca la mejor frontera entre los segmentos P y S , y se crea el arco que los enlaza. Con el objeto de evitar bucles los casos inversos de C y D no se contemplan, es decir, nunca se buscará solución a una situación similar a la de C en que S esté en la parte inicial de P , ya que sería el caso D. Ídem para el caso D. Además, para el caso D nunca se generará un arco que intente enlazar S como predecesor de P , aunque éste comience primero.

Los casos E y F implican la división de los segmentos. En la figura 5 se presentan las soluciones que se adoptan para estos casos, donde $P1$ es un nodo que surge de la parte inicial de P , y $P2$ de la parte final, ídem para el segmento S .

La mejor frontera entre los segmentos del par de nodos a ser enlazados por un arco se obtiene mediante la expresión 3:

$$\Upsilon = \arg \min_{\tau \in [t_i(n)+2, t_f(n')-1]} \left(\sum_{t=t_i(n)}^{\tau-1} -\log(P(ph_n|x_t)) + \sum_{t=\tau}^{t_f(n')} -\log(P(ph_{n'}|x_t)) \right) \quad (3)$$

donde Υ representa la frontera entre ambos segmentos y toma el valor de la trama donde empieza el segmento sucesor, ph_n es la unidad fonética del nodo n , $ph_{n'}$ la de n' , $t_i(n)$ es la trama inicial del nodo n , y $t_f(n')$ es la trama final del nodo n' . Las constantes en la expresión " $\tau \in [t_i(n) + 2, t_f(n') - 1]$ " están para evitar que uno de los segmentos quede reducido a una trama.

El ajuste en cuanto a coste de mover las fronteras de dos nodos para ser enlazados se calcula aplicando la expresión 4: en caso de que no haya solapamiento el ajuste será positivo, y en caso de solapamiento negativo, para evitar que algunas tramas se computen por duplicado.

$$ajuste = ajuste_p(ph_n, \Upsilon) + ajuste_s(ph_{n'}, \Upsilon) \quad (4)$$

$$ajuste_p(ph_n, \Upsilon) = \begin{cases} -\sum_{t=\Upsilon}^{t_f(n)} -\log(P(ph_n|x_t)) & \text{si } \Upsilon < t_f(n) + 1 \\ 0 & \text{si } \Upsilon = t_f(n) + 1 \\ \sum_{t=t_f(n)+1}^{\Upsilon-1} -\log(P(ph_n|x_t)) & \text{si } \Upsilon > t_f(n) + 1 \end{cases} \quad (5)$$

$$ajuste_s(ph_{n'}, \Upsilon) = \begin{cases} -\sum_{t=t_i(n')}^{\Upsilon-1} -\log(P(ph_{n'}|x_t)) & \text{si } t_i(n') < \Upsilon \\ 0 & \text{si } t_i(n') = \Upsilon \\ \sum_{t=\Upsilon}^{t_i(n')-1} -\log(P(ph_{n'}|x_t)) & \text{si } t_i(n') > \Upsilon \end{cases} \quad (6)$$

4. Decodificación Acústico-Fonética

Aunque el propósito de los grafos de fonemas es servir como punto de partida para la construcción de los grafos de palabras, si buscamos el mejor camino encontramos una secuencia fonética que puede servir para estimar la calidad de estos grafos de fonemas.

La búsqueda del mejor camino sobre el grafo de fonemas se realiza mediante técnicas de programación dinámica. La medida para evaluar el coste de un camino se calcula según la ecuación 7.

$$\begin{aligned} coste_{camino} &= \\ &= \sum_{i \in [1..|camino|]} \left(coste(n_i) + ajuste(n_{i-1} \rightarrow n_i) \right. \\ &\quad \left. - \rho \cdot \log P(ph_i|ph_{i-1}) \right) \end{aligned} \quad (7)$$

donde $coste(n_i)$ es el coste del nodo n_i según la ecuación 2, $ajuste(n_{i-1} \rightarrow n_i)$ es el coste de ajustar la frontera entre los nodos n_{i-1} y n_i , y se estima según la ecuación 4, $P(ph_i|ph_{i-1})$ es la probabilidad que aporta el modelo fonológico, y representa la probabilidad de que aparezca la unidad fonética ph_i después de ph_{i-1} , ρ es una constante ajustada empíricamente para ponderar adecuadamente la importancia del modelo fonológico. En los resultados que se muestran el valor de ρ es 2.

5. Experimentación

A continuación se muestran los resultados obtenidos haciendo Decodificación Acústico-Fonética (DAF) para dos bases de datos: *FRASES* y *ALBAYZIN*.

FRASES está compuesta de 170 frases diferentes pronunciadas por 10 locutores, 5 hombres y 5 mujeres, con un total de 1700 frases (alrededor de 1 hora de voz). Para el entrenamiento de los modelos acústico-fonéticos se han utilizado 120 frases de 7 locutores [8]. Se han definido tres tipos de experimentos (véase tabla 4).

Tabla 4: Particiones de test en la base de datos *FRASES*.

<i>dlit</i>	Dependiente del locutor e independiente de la tarea. 50 frases por 7 locutores
<i>ildt</i>	Independiente del locutor y dependiente de la tarea. 120 frases por 3 locutores
<i>ilit</i>	Independiente del locutor y de la tarea. 50 frases por 3 locutores

La base de datos *ALBAYZIN* [9] está compuesta por 700 frases pronunciadas entre 40 locutores diferentes (aunque no todas las frases han sido pronunciadas por todos los locutores). Esta base de datos está compuesta de dos subcorpus, uno fonético y otro específico, con 6800 frases cada uno. En nuestros experimentos hemos utilizado el subcorpus fonético con alrededor de 6 horas de voz. Sobre este subcorpus hemos hecho dos particiones, una para entrenamiento y otra para test. La partición de test es independiente del locutor pero no de la tarea.

Los resultados de DAF obtenidos se muestran en las tablas 5 y 6, y se han obtenido comparando la transcripción fonética correcta de la frase con la correspondiente al camino con mayor probabilidad del grafo de fonemas obtenido [10]. Los resultados que aquí se muestran se obtienen a partir del cálculo de la distancia de Levenshtein entre la transcripción fonética de referencia y la secuencia fonética obtenida al recorrer el GF. El algoritmo de programación dinámica que calcula esta distancia proporciona el número de aciertos (A), el de sustituciones (S), el de borrados (B), y el de inserciones (I) [11].

Como medida de la calidad del grafo de fonemas presentamos el Porcentaje de Aciertos (PA), que se estima según la siguiente expresión

$$PA = \frac{A}{A + I + B + S} \quad (8)$$

Tabla 5: Resultados de DAF (PA) aplicando el modelo fonológico para las diferentes particiones de test de la base de datos *FRASES*. Se muestran los resultados de dos tipos de experimentos, los obtenidos restando la media de los coeficientes cepstrales (*csmDA*), y sin restarla (*ccDA*).

test	ccDA	csmDA
<i>dlit</i>	73.01	72.70
<i>ildt</i>	67.34	69.77
<i>ilit</i>	64.46	67.22

Tabla 6: Resultados de DAF (PA) aplicando el modelo fonológico para la base de datos *ALBAYZIN*.

test	ccDA	csmDA
<i>ildt</i>	66.37	67.14

Los resultados obtenidos de DAF aplicando el modelo fonológico son comparables con los obtenidos por otros sistemas.

Nótese que a diferencia de otros métodos que necesitan una segmentación manual de parte del subcorpus de entrenamiento, en nuestro caso se realiza un aprendizaje no supervisado para estimar las probabilidades fonéticas [8]. Además, los modelos acústicos utilizados en nuestro sistema son más sencillos si se comparan con los modelos ocultos de Markov o con las redes neuronales.

Cabe resaltar que por la manera en que se ha diseñado nuestro sistema, como entrada al módulo generador de grafos de fonemas se puede utilizar una secuencia de vectores de probabilidades fonéticas (SVPF) calculada por un clasificador fonético distinto.

Tabla 7: Resultados de DAF (PA) para la base de datos FRASES. Los modelos acústicos son modelos ocultos de Markov con estimación de probabilidades utilizando perceptrones multicapa.

test	MOM-PM
dltit	77.00
ildt	72.01
ilit	70.97

No obstante, con el fin de comparar nuestros resultados, se presentan resultados de DAF obtenidos con un sistema de RAH que implementa la aproximación clásica (tabla 7). Los modelos fonéticos son cadenas de Markov con la particularidad de que las probabilidades de emisión se estiman con redes neuronales [12].

Por último, el objetivo final de nuestros grafos de fonemas no es servir para DAF, sino conservar todas las posibles alternativas, dada la señal vocal de una pronunciación, en vistas a que otro módulo posterior genere un grafo de palabras. Sobre dicho grafo de palabras actuará un módulo reconocedor para encontrar la frase pronunciada guiado por un modelo de lenguaje. Los grafos permiten conservar todas aquellas variantes suficientemente probables y que sea el siguiente módulo el que pade, pero esta característica merma significativamente los resultados de DAF.

Una medida complementaria de la calidad de los grafos de fonemas es la comparación de la secuencia fonética de referencia con la mejor derivación que se obtiene a partir del grafo de fonemas (la secuencia fonética que corresponde al mejor camino). El PA obtenido de esta manera permite valorar la capacidad del sistema para detectar los fonemas pronunciados. Los resultados así obtenidos se muestran en la tabla 8.

Tabla 8: Resultados de DAF (PA) comparando la secuencia fonética correcta con la mejor derivación de la secuencia fonética, para las diferentes particiones de test de las bases de datos FRASES y ALBAYZIN.

	FRASES		ALBAYZIN	
test	ccDA	csmDA	ccDA	csmDA
dltit	89.05	89.34		
ildt	84.71	86.03	86.64	86.97
ilit	83.22	85.30		

6. Conclusiones

A pesar de las diferencias con otros resultados a nivel de DAF, los obtenidos a partir de los grafos de fonemas son aceptables.

En vista de los resultados obtenidos con DAF, y de los resultados que miden la capacidad de detección de unidades fonéticas, se considera factible la construcción de los restantes módulos (*generador de grafos de palabras* y *reconocedor*) a efectos de completar el sistema de RAH desacoplado. Además, nuestro sistema está abierto para utilizar otros métodos de clasificación a nivel acústico fonético que sean capaces de obtener una secuencia de vectores con probabilidades fonéticas a partir de la señal vocal.

7. Referencias

- [1] John R. Deller, John G. Proakis, John H.L. Hansen. "Discrete-Time Processing of Speech Signals". Macmillan Publishing Company, 1993
- [2] Jon A. Gómez, D. Llorens, N. Prieto, E. Sanchís. "Reconocimiento automático del habla: modelos y técnicas". Novática enero-febrero 1998 nº 131. págs: 18-23.
- [3] Lawrence R. Rabiner, Biing-Hwang Juang. "Fundamentals of Speech Recognition". Prentice-Hall, 1978
- [4] Steve Young. "A Review of Large-Vocabulary Continuous-speech Recognition". IEEE Signal Processing Magazine September 1996, 45-57
- [5] Schwartz R. and Chow Y., "The N-best Algorithm: an Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses". ICASSP'90, 1990, Albuquerque, USA.
- [6] Tran B-H., Seide F., Steinbiss V. "A Word Graph Based N-best Search in Continuous Speech Recognition", Proceedings of ICSLP 1996.
- [7] Pusateri E. and Van Thong J.M., "N-best List Generation using Word and Phoneme Recognition Fusion". EUROSPEECH 2001 - SCANDINAVIA, Norway, 2001.
- [8] Gómez, J.A. and Castro, M.J., "Automatic Segmentation of Speech at the Phonetic Level", *Structural, Syntactic and Statistical Pattern Recognition, Joint IAPR International Workshops SSPR and SPR 2002*, Lecture Notes in Computer Science, Ed. T. Caelli et al., Vol. 2396, 2002, p 672—680.
- [9] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Mariño y C. Nadeu, "Albayzin Speech Database: Design of the Phonetic Corpus". In *Eurospeech93*, vol. 1, p 653—656, Berlin, Germany, September, 1993.
- [10] María José Castro, Salvador España, Andrés Marzal e Ismael Salvador, "Grapheme-to-phoneme conversion for the Spanish language", Proceedings of the IX National Symposium on Pattern Recognition and Image Analysis, p 397—402, Benicàssim, España, 2001.
- [11] F. Prat, P. Aibar, A. Marzal y E. Vidal. "El problema de la evaluación de un sistema de reconocimiento automático del habla mediante un único valor numérico". Informe técnico DSIC II/15/94, DSIC, UPV, 1994.
- [12] Castro Bleda, M.J., "Modelado acústico de unidades subléxicas mediante una aproximación basada en métodos estructurales-conexionistas". Tesis Doctoral, 1998. Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia.