FUNDAMENTOS DEL RECONOCIMIENTO AUTOMÁTICO DE LA VOZ

"Decodificación o búsqueda de hipótesis"

Agustín Álvarez Marquina

Introducción. Decodificación o búsqueda de hipótesis (I)

- **Objetivo: debemos buscar ahora todas las posibles cadenas de palabras *W* para encontrar la que maximiza la fórmula:**

$$\hat{W} = \arg\max_{W} P(A \mid W) P(W)$$

donde *A* son los datos acústicos y $W = w_1, w_1, \ldots, w_n$, con $w_i \in V$, denota la cadena de *n* palabras de entre un vocabulario de tamaño fijo *V*.

Introducción. Decodificación o búsqueda de hipótesis (II)

- **La búsqueda no puede ser conducida por la fuerza bruta, puesto que el espacio de *W* es enormemente grande.**
  - Esto quiere decir que se necesita un conjunto reducido de hipótesis de búsqueda que no examinará el contundente número de posibles candidatos.
  - Solamente tendrá en cuenta aquellas cadenas de palabras que de alguna manera estén sugeridas por la estructura acústica *A*.

Estrategia integrada (I)

- **La tarea de encontrar la oración que mejor satisface todas las restricciones acústicas y lingüísticas puede realizarse mediante el uso de dos estrategias básicas:** integrada y modular**.**
- **En la aproximación integrada, la decisión implicada en el proceso de reconocer se realiza considerando todas las fuentes de conocimiento de manera combinada.**
  - En principio, esta estrategia consigue los mejores resultados si todas las fuentes de conocimiento pueden caracterizarse e integrarse de manera completa.

Estrategia integrada (II)

- En este caso es posible aproximarse a la jerarquía de conocimiento lingüístico (acústico, léxico, sintáctico y semántico) y compilarlas dentro de una red de estados finitos compuesta por nodos gramaticales, nodos correspondientes a los modelos ocultos de Markov y las conexiones entre ellos.

- **De esta forma, el problema del reconocimiento se resuelve mediante el encaje de la secuencia de vectores de características de entrada con la mejor secuencia de palabras que atraviese la citada red de conocimiento.**

Estrategia integrada (III)

- **Esta es la estrategia de búsqueda que normalmente adoptan los sistemas de reconocimiento de hoy en día.**

- **Problemas con este tipo de aproximación:**

  - No todas las fuentes de conocimiento pueden caracterizarse e integrarse de una forma total. Ejemplo: la prosodia.

  - Aparte para muchas tareas donde el tamaño del vocabulario es grande la red resultante puede ser computacionalmente intratable.

Estrategia modular (I)

- La aproximación modular *(Figura 1)* emplea la información proporcionada por las diferentes fuentes de conocimiento de una forma secuencial, pudiéndose especificar cada módulo de forma separada.

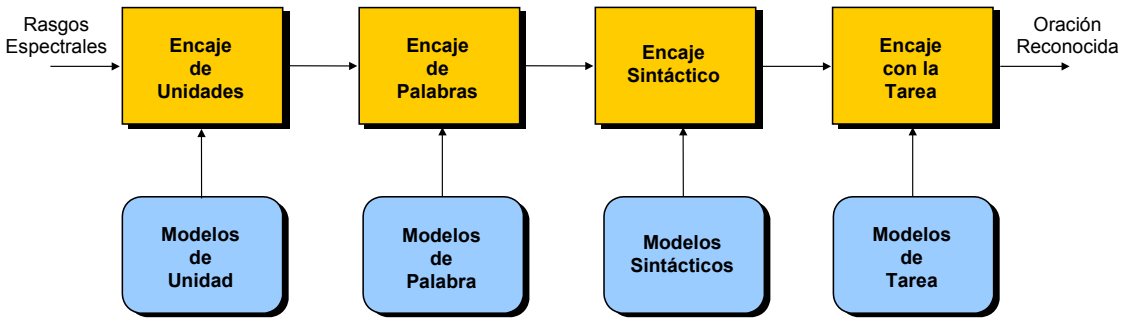

Figura 1. Diagrama de bloques de un reconocedor de habla continua modular.

Estrategia modular (II)

- La mayoría de los sistemas de diálogo funcionan con este principio de colaboración.
  - Su mayor ventaja radica en que es computacionalmente una opción más factible.
  - Por contra, la mayor limitación de este modelo es que las decisiones, que se realizan en una etapa no tienen en cuenta a las otras etapas/fuentes de conocimiento.
    - Los errores de decisión se propagan de una etapa a la siguiente y esta acumulación es susceptible de causar errores en la búsqueda a menos que, se mantengan para cada paso del proceso múltiples hipótesis.

Algoritmos de búsqueda de hipótesis en reconocimiento de voz

- **En los últimos años se ha realizado un progreso significativo en el desarrollo de los algoritmos de búsqueda [JEL94] [GUP95]. Los ejemplos más destacados son:**
  1. Algoritmo de búsqueda en haz de una sola pasada (*One-Pass Beam Search*).
  2. Algoritmos de búsqueda heurística basados en el A* [NIL80] y decodificación con pila (*stack decoding*) [PAU91], [PAU92], [GOP95a], [GOP95b].
  3. Estrategias de decisión con múltiples pasadas.

Algoritmo de búsqueda en haz de una sola pasada (I)

- **Para la parte del conocimiento que puede integrarse dentro de la red de estados finitos, el problema de búsqueda se suele resolver encontrando el mejor camino posible a través de la red.**
  - Un algoritmo de búsqueda exhaustiva por paso, como el algoritmo de Viterbi, resulta muy costoso en términos de tiempo de procesamiento y capacidad de almacenaje.

Algoritmo de búsqueda en haz de una sola pasada (II)

- En este algoritmo solamente un pequeño conjunto de todas las hipótesis parciales plausibles (palabras), que caigan dentro del haz es evaluado.
- Diversas técnicas como son el uso de árboles léxicos y la búsqueda anticipada de fonemas se han implementado con objeto de reducir el número de hipótesis durante la búsqueda.

Algoritmo de búsqueda en haz de una sola pasada (III)

- La anchura del haz, que determina los costes computacionales y el posible número de errores producto de la poda en el árbol de búsqueda, se establece de manera empírica, siendo dependiente de la tarea y del modelado empleado.
- Hay que hacer notar que la solución proporcionada por este método de búsqueda no es la óptima.

Algoritmo de búsqueda en haz de una sola pasada (IV)

- En la práctica, sin embargo, se observa que la probabilidad calculada con este algoritmo presenta la propiedad de dominio, que consiste en que el camino así hallado concentra la mayor parte de la probabilidad total [MER91].

Búsqueda heurística basada en el A* y decodificación con pila (I)

- Teniendo en cuenta que la señal de voz porta información lingüística de alguna manera localizada, no todos los eventos lingüísticos están activos y necesitan ser evaluados en todo momento.

- Con el fin de asimilar esta propiedad se puede usar una estrategia de búsqueda del tipo "el mejor primero" *(best first)*.

Búsqueda heurística basada en el A* y decodificación con pila (II)

- **Esta búsqueda se suele implementar mediante el uso de una pila que mantiene para cada instante de tiempo una lista ordenada con las hipótesis parciales.**

- **Entonces, la mejor hipótesis de la pila se intenta extender a una pequeña lista de palabras, elegida de acuerdo a la bondad de los encajes en los niveles acústico y gramatical.**
  - Una de las ventajas de este método es que los modelos de lengua a largo plazo pueden integrarse de forma natural en la búsqueda.

Búsqueda heurística basada en el A* y decodificación con pila (III)

- **Para controlar el crecimiento exponencial de la pila se suelen emplear estrategias de búsqueda basadas en el algoritmo A*.**
  - De esta manera se emplea, no sólo la estimación del coste de un camino hasta el punto alcanzado, sino también una evaluación del camino que falta por recorrer.

- **Otro ejemplo de algoritmo basado en el A*, pero aplicado a la búsqueda en un grafo constituido por elementos silábicos lo encontramos en [GUP88].**

Búsqueda heurística basada en el A* y decodificación con pila (IV)

- La gran ventaja de este método es el menor tamaño de la red como consecuencia del menor número de elementos básicos (sílabas en este caso).

- **El algoritmo del enrejado en árbol *(tree-trellis)* [SOO91] es una forma eficaz de controlar el tamaño de la pila al mantener todas las posibles opciones de continuación en un árbol y recombinarlas con las hipótesis parciales hacia atrás, que presentes en la pila, se han ido clasificado en un segundo árbol de búsqueda.**

Estrategias de decisión con múltiples pasadas (I)

- **Este tipo de algoritmos realiza una primera pasada con el fin de establecer las hipótesis iniciales y en vueltas posteriores se completan de forma progresiva [MUR93], [ALL93] [AUB94].**
  - En oposición a los algoritmos tradicionales de búsqueda de izquierda a derecha (una sola pasada).

- **Los métodos muti-paso, como el mencionado anteriormente, se suelen diseñar con el fin de proporcionar las mejores N hipótesis (oraciones) [SCH90], [SCH92].**

Estrategias de decisión con múltiples pasadas (II)

- **Este paradigma de búsqueda resulta ideal cuando se desea integrar múltiples fuentes de conocimiento a diferentes niveles de abstracción.**

- **Con objeto de mejorar la flexibilidad, se pueden emplear modelos acústicos más sencillos para producir un retículo con segmentos *(segment lattice)* o con fonemas *(phone lattice)*, en la primera pasada de la búsqueda.**
  - Los modelos léxicos y de lengua pueden además incorporarse para generar un retículo de palabras *(word lattice)*.

Estrategias de decisión con múltiples pasadas (III)

- **La familia de métodos que se conoce con el nombre de búsqueda progresiva *(progressive search)*, constituye una forma muy interesante de aunar diferentes niveles de conocimiento de una manera sistemática.**

- **Para reducir el número de hipótesis posibles durante el proceso de búsqueda se puede usar un procedimiento de encaje rápido conocido como *fast match (Figura 2)* sobre una red reducida con modelos acústicos simplificados.**

Estrategias de decisión con múltiples pasadas (IV)

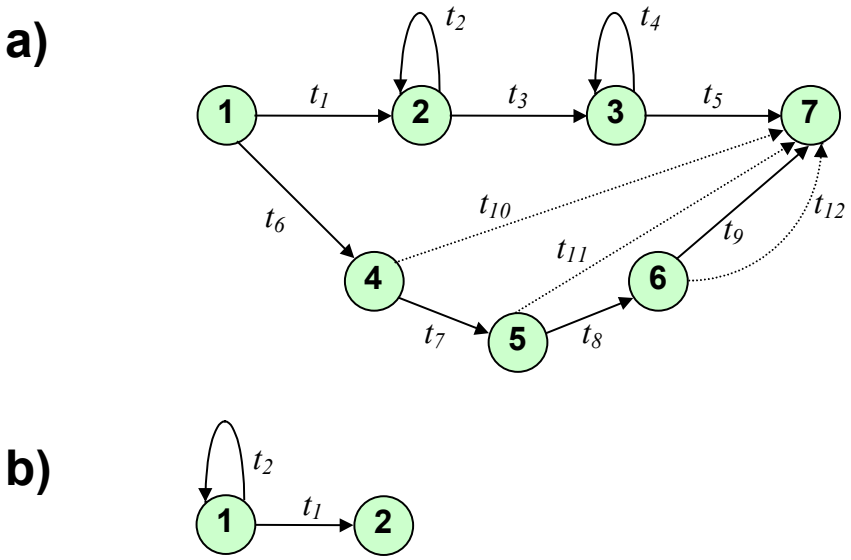

Figura 2. Estructura de un modelo oculto de Markov correspondiente a un fonema. a).- Estructura estándar. b).-Estructura reducida.

Estrategias de decisión con múltiples pasadas (V)

- Algunas modificaciones a los algoritmos de búsqueda anteriores pasan por añadir restricciones de tipo temporal a los diferentes segmentos de la red de búsqueda [BUS87], [SOO89].
- La incorporación de diferentes fuentes de información (a veces incompatibles entre sí) implica también la necesidad de establecer un modelo que permita manejar diferentes hipótesis para cada una de ellas.

Estrategias de decisión con múltiples pasadas (VI)

- Una posibilidad es la que aparece en la *Figura 3*.
  - Las soluciones parciales de los cuatro módulos se integran con el fin de determinar la solución global.

- **La combinación de este tipo de aproximación con estrategias de verificación de realizaciones resulta ser una manera bastante flexible de diseñar sistemas apropiados para tratar con grandes vocabularios [RAB96].**

Estrategias de decisión con múltiples pasadas (VII)

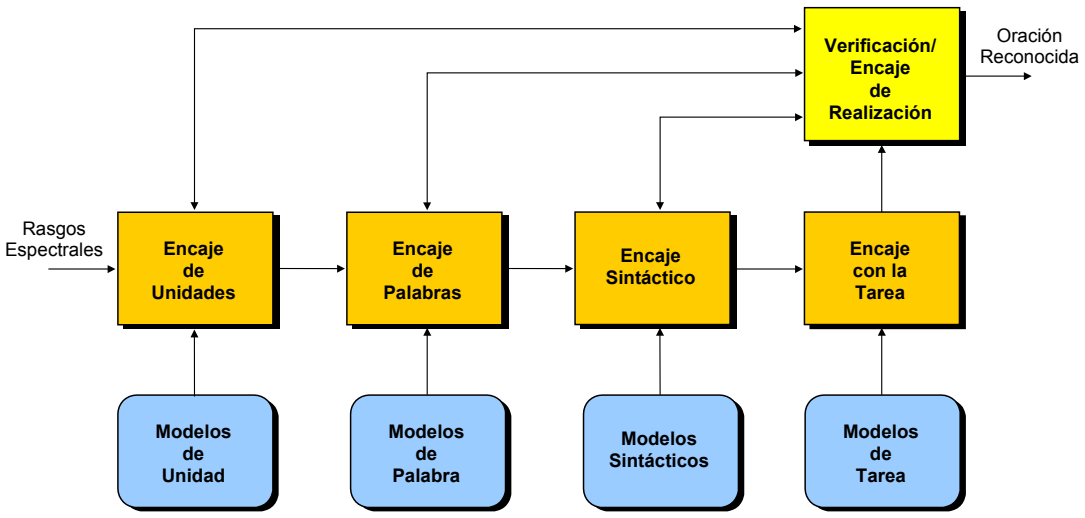

Figura 3. Diagrama de bloques de un reconocedor de habla continua modular con integración de conocimiento.

Bibliografía (I)

[ALL93] F. Alleva, X. Huang and M. Y. Hwang, "An Improved Search Algorithm Using Incremental Knowledge for Continuous Speech Recognition", *Proc. of ICASSP'93*, Minneapolis, Estados Unidos, 27-30 abril 1993, Vol. II, pp. 307-310.

[AUB94] X. Aubert et al., "Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data", *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. II, pp.129-132.

[BUS87] M. A. Bush and G. E. Kopec, "Network-Based Connected Digit Recognition", *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. ASSP-35, Nº. 10, octubre 1987, pp. 1401-1413.

[GOP95a] P. S. Gopalakrishnan, "Continuous Speech Recognition", *Modern Methods of Speech Processing*, R. P. Ramachandran and R, J. Mammone editores, Kluwer Academic Publishers, 1995, pp. 185-212.

[GOP95b] P. S. Gopalakrisnan, L. R. Bahl and R. L. Mercer, "A Tree Search Strategy for Large-Vocabulary Continuous Speech Recognition", *Proc. of ICASSP'95*, Detroit, Estados Unidos, 9-12 mayo 1995, Detroit, Estados Unidos, 9-12 mayo 1995, pp. 572-575.

[GUP88] V. N. Gupta, M. Lennig and P. Mermelstein, "Fast search strategy in a large vocabulary word recognizer", *Journal of Acoustic Society of America*, Vol. 44, Nº 6, diciembre 1988, pp. 2007-2017.

Bibliografía (II)

[GUP95] V. Gupta and M. Lenning, "Large Vocabulary Isolated Word Recognition", *Modern Methods of Speech Processing*, R. P. Ramachandran and R, J. Mammone editores, Kluwer Academic Publishers, 1995, pp. 213-230.

[JEL94] F. Jelinek, "Training and Search Methods for Speech Recognition", *Voice Communication Between Humans ans Machines*, D. B. Roe and J. G. Wilpon editores, National Academy Press, Washington D. C., 1994, pp. 199-214.

[MER91] N. Merhav and Y. Ephraim, "Maximum Likehood Hidden Markov Modeling Using a Dominant Sequence of States", *IEEE Transactions on Signal Processing*, Vol. 39, Nº. 9, septiembre 1991, pp. 2111-2114.

[MUR93] H. Murveit et al., "Large-Vocabulary Dictation Using SRI´s Decipher™ Speech Recognition System: Progressive Search Techniques", *Proc. of ICASSP'93*, Minneapolis, Estados Unidos, 27-30 abril 1993, Vol. II, pp. 319-322.

[NIL80] N. Nilsson, *Principles of Artificial Intelligence*, Palo Alto, California, Tioga, 1980.

[PAU91] D. B. Paul, "Algorithms for An Optimal A* Search and Linearizing the Search in the Stack Decoder", *Proc. of ICASSP'91*, Toronto, Canadá, 14-17 mayo 1991, pp. 693-696.

Bibliografía (III)

[PAU92] D. B. Paul, "An Efficient A* Stach Decoder Algorithm for Continuous Speech Recognition with a Stochastic Language Model", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, Vol. I, pp. 25-28.

[RAB96] L. R. Rabiner, B. H. Juang and C. H. Lee, "An Overview of Automatic Speech Recognition", *Automatic Speech and Speaker Recognition: Advanced Topics*, C. H. Lee, F. K. Soong and K. K. Paliwal editores, Kluwer Academic Publisher, 1996, pp. 1-30.

[SCH90] R. Schwartz and Y. L. Chow, "The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses", *Proc. of ICASSP'90*, Alburquerque, Estados Unidos, 3-6 abril 1990, pp. 81-84.

[SCH92] R. Schwartz et al., "New Uses for the N-Best Sentence Hypotheses within the Byblos Speech Recognition System", *Proc. of ICASSP'92*, Vol. I, pp. 1-4.

[SOO89] F. K. Soong, "A Phonetically Labeled Acoustic Segment (PLAS) Approach to Speech Analysis-Synthesis", *Proc. of ICASSP'89*, Glasgow, Reino Unido, 23-26 mayo 1989, pp. 584-587.

[SOO91] F. K. Soong and E. F. Huang, "A Tree-Trellis Based Fast Search for Finding the N Best Sentence Hypotheses in Continuous Speech Recognition", *Proc. of ICASSP'91*, Toronto, Canadá, 14-17 mayo 1991, pp. 705-708.