# RECONOCIMIENTO AUTOMATICO DEL HABLA

FRANCISCO CASACUBERTA NOLLA
ENRIQUE VIDAL RUIZ

Dto.de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia

## 1. INTRODUCCION.

El habla ha venido siendo tradicionalmente el medio preferente de comunicación entre seres humanos. Esta constituye un proceso altamente codificado, cuyo vehículo, la voz, transporta no sólo información semántica, sino también información fisiológica y sociológica del locutor.

Desde los comienzos de la Informática, y más concretamente de la Inteligencia artificial, se ha intentado dotar a los computadores de este medio de comunicación, aunque separando el aspecto de producción (Síntesis) del de percepción (reconocimiento). Treinta años después, los logros obtenidos están muy lejanos de conseguir una verdadera comunicación hablada entre humanos y computadores. Mientras que en la síntesis, existen productos acabados bastante aceptables, en el Reconocimiento, éstos son muy limitados y de muy restringida utilidad práctica, a pesar de las informaciones optimistas que suelen aparecer en los medios de comunicación social no especializados.

Las causas de la relativamente poco alentadora situación actual son varias y están muy relacionadas con las características intrínsecas del habla. En primer lugar cabe destacar la *continuidad,* ni los fonemas, ni las sílabas, ni siquiera las palabras constituyen elementos discretos que se puedan separar fácilmente de forma automática. Además, y debido a inercias del aparato fonador humano, éstos se influyen unos a otros, dando lugar a complejos efectos de *coarticulación*. En segundo lugar está la *variabilidad,* un mismo locutor no pronuncia dos veces una misma palabra de forma idéntica, y menos aún dos locutores distintos (Vaissière, 1985) (Casacuberta, 1987a). En tercer lugar tenemos el *ruido y la distorsión,* en la señal vocal, parte de la información irrelevante para la comprensión del mensaje hablado. Finalmente, la *imprecisión* de los conocimientos multidisciplinarios disponibles de tipo acústico, fonético, etc. es lo suficientemente alta como para que no se pueda basar en ellas el diseño de modelos exactos útiles para la interpretación de la señal vocal.

Asumiendo estos inconvenientes, se han propuesto y estudiado diversas metodologías y arquitecturas en las que se han impuesto una serie de restricciones con el objeto de simplificar el problema general del Habla, para que éste sea abordable. Las simplificaciones que se asumen afectan fundamentalmente a la talla del léxico, al tipo del lenguaje y a la variabilidad aceptable en la señal vocal. Según la naturaleza de las simplificaciones, el Reconocimiento del Habla se convierte en: *Reconocimiento de Palabras Aisladas:* el léxico es reducido (hasta unas 500 palabras), y las palabras deben pronunciarse con pausas entre ellas. En el caso más simple, el reconocimiento es monolocutor (el aprendizaje del sistema se realiza con el mismo locutor que lo va a utilizar), y en otro caso es multilocutor (en el que intervienen varios locutores) lo que suele reducir significativamente la talla del vocabulario utilizable. *Reconocimiento de Palabras Conectadas:* es similar al anterior aunque, a cambio de restringir el vocabulario, no se exigen pausas entre palabras. La sintaxis suele ser muy elemental o inexistente. *Reconocimiento de Palabras Aisladas y Conectadas con Diccionarios*

*Difíciles:* en este caso, las palabras pertenecen a pequeños vocabularios, pero suelen ser muy cortas (letras y dígitos, por ejemplo) y/o muy parecidas. *Reconocimiento de Palabras Aisladas de Grandes Diccionarios:* es similar al primer caso, aunque la talla del diccionario puede alcanzar las 50.000 palabras, lo que impide la creación y uso de prototipos como modelos de las palabras a reconocer, exigiendo por tanto una aproximación analítica al reconocimiento. *Reconocimiento del Discurso Continuo:* aquí se plantean problemas de "comprensión" en tareas con semántica restringida, y la sintaxis suele ser compleja aunque artificial o "pseudonatural".

En general, un sistema encuadrado en algunas de las simplificaciones anteriores, se compone de un conjunto de módulos (integrados o independientes), cada uno de los cuales lleva asociada una Fuente de Conocimientos (acústica, fonética, léxica, sintáctica o semántica) y un procedimiento interpretativo correspondiente. Estos módulos interaccionan entre sí con el objetivo de obtener una interpretación aceptable del mensaje hablado. El número y/o tipo de dichos módulos dependerá de la estrategia de reconocimiento adoptada y del tipo de interpretaciones que se deseen (Casacuberta, 1987b).

A continuación, describiremos una breve historia del Reconocimiento Automático del Habla, junto con las metodologías más en boga y comentaremos, según nuestro punto de vista, el futuro de esta Ciencia multidisciplinar.

## 2. HISTORIA Y DESARROLLO DEL RECONOCIMIENTO AUTOMATICO DEL HABLA

Aunque el origen del Reconocimiento Automático del Habla data de comienzos de los sesenta, el verdadero ímpetu no se alcanza hasta los setenta, coincidiendo con el lanzamiento de un gran proyecto de investigación: *SUR* (Speech Understanding Research) promocionado por la agencia ARPA (Advanced Research Project Agency) del Departamento de Defensa de los EE.UU. El objetivo de este proyecto

fue muy ambicioso: un léxico de talla media para una tarea restringida, un número de errores semánticos relativamente pequeño y de un coste computacional no excesivo (Klatt, 1980). Estos objetivos no fueron alcanzados totalmente, aunque se hicieron notables aportaciones a las arquitecturas y metodologías de la Inteligencia Artificial y contribuyó a un mejor conocimiento de las características del habla y de las limitaciones del reconocimiento automático.

Entre las Universidades y empresas que participaron con mayor éxito en el proyecto ARPA-SUR se deben mencionar: la Universidad de Carnegie-Mellon con el sistema *HARPY* (Newell, 1978) (Lowerre, 1980), en el que las fuentes de conocimiento estaban integradas en una red de estados finitos. Este sistema fue el que más se aproximó a los objetivos que se habían propuesto (Casacuberta, 1987a). La misma Universidad desarrolló el sistema *HEARDAY-II* (Erman, 1980), en el que se propuso la arquitectura de "pizarra", ampliamente utilizada posteriormente para la implementación de sistemas de Inteligencia Artificial. El tercer gran sistema fue *HWIM* de la Bolt Beranek and Newmann Inc. (Woolf, 1980); aunque este sistema estaba lejano de los objetivos propuestos, supuso una gran aportación en cuanto a estrategias de razonamiento en sistemas inteligentes (Casacuberta, 1987b).

En paralelo con el proyecto SUR, otras empresas desarrollaron los suyos propios. Así por ejemplo, el grupo de Proceso del Habla de IBM Thomas J. Watson Research Center (Jelinek, 1976) (Bahl, 1983) propuso un sistema basado en técnicas de Teoría de la Comunicación, en el cual el léxico y la sintaxis-semántica estaban representados mediante Modelos de Markov.

En la segunda mitad de los setenta y principios de los ochenta, se desarrollaron ciertas técnicas relacionadas con la llamada *Aproximación Global al Reconocimiento del Habla,* que permitieron resolver de forma satisfactoria un conjunto de problemas simples (Reconocimiento de Palabras Aisladas y Conectadas) (Casacuberta, 1987a). Esta aproximación considera a los objetos del habla (fonemas, sílabas, palabras) como un todo (sin es-

tructura), y está basado en almacenar ciertas representaciones de tales objetos con los que comparar cuando se intenta interpretar un nuevo objeto desconocido. No obstante, para aplicaciones más ambiciosas es necesario asumir que un objeto del habla está formado por unidades más pequeñas relacionadas entre sí (*Aproximación Analítica*). En estos casos, el problema reside en identificar estas unidades para luego ser analizadas en conjunto por algún sistema basado en técnicas sintácticas o de Inteligencia Artificial (Casacuberta, 1987b).

En paralelo a estos trabajos, y aparte de los proyectos americanos mencionados, varios equipos de investigación europeos y japoneses realizaron diversas propuestas para abordar el problema del Reconocimiento del Discurso Continuo, y éstas se caracterizaron por ser soluciones más o menos ad hoc y por su falta de unificación en cuanto a la metodología a utilizar. Cabe destacar el proyecto MYRTILLE del Centre de Recherche en Informatique de Nancy (Francia) (Pierrel, 1981) (Casacuberta, 1987b). También en Francia el proyecto ARIAL (Perennou, 81) de la Universidad Paul Sabatier. En España el proyecto TABARCA en la Universidad de Valencia primero y en la Universidad Politécnica de Valencia después (Vidal, 1985) (Casacuberta, 1987a). Otros muchos proyectos aparecieron en esa época (Haton, 1981).

Durante la primera mitad de los ochenta, el Reconocimiento Automático del Habla sufre el impacto de los llamados *Sistemas Basados en el Conocimiento,* también conocidos como Sistemas Expertos. Estos sistemas de Inteligencia Artificial, que desde pocos años atrás estaban teniendo un auge general importante, está basado en la idea de que los conocimientos que poseen los expertos sobre algún tema podrían ser transmitidos a un sistema informático, el cual pasaría, a partir de ese momento, a poseer unas habilidades similares a las del experto. Ejemplo de tales sistemas fueron y son el EVAR de la Universidad de Erlangen (RFA) (Niemann, 1985) SYSTEXP de la Universidad de Nancy (Francia) (Haton, 1985) la sociedad de Expertos de la Universidad de Concordia (Canadá) (De Mori, 1985), el SERAC del CNET (Francia) (Mercier, 1985).

Aunque estos sistemas han alcanzado ciertos éxitos, actualmente parecen estar perdiendo el enorme auge que tuvieron anteriormente. La causa principal hay que buscarla en las limitaciones derivadas de la insuficiencia de conocimientos que los (expertos) humanos tienen sobre el proceso del habla, así como a la dificultad de transmitir de forma efectiva a una máquina conocimientos de tipo intuitivo y/o inconsciente.

A mediados de los ochenta, cobra una gran fuerza los *Modelos de Markov* como formas de representación del conocimiento acústico-fonético y léxico. El origen de estos modelos está en el sistema de IBM propuesto en la década anterior y en el sistema DRAGON, que fue un precursor del HARPY (Baker, 1975). El gran desarrollo actual de estos modelos reside en su robusta base matemática y en su capacidad de aprendizaje automático o semiautomático a partir de muestras vocales correspondientes a la tarea de reconocimiento abordada. Las principales aplicaciones de estos modelos recaen originalmente de la representación de las palabras del léxico para el Reconocimiento de Palabras Aisladas y Conectadas, así como en la parte más débil de todo sistema de Reconocimiento basado en la Aproximación Analítica: la *Decodificación Acústico-Fonética*. Esta parte pretende obtener una representación de tipo fonético a partir de la señal vocal, y los modelos de Markov son utilizados para la representación de las unidades subléxicas escogidas. Hoy en día, son muy numerosos los equipos de investigación que utilizan tal modelización, como la Bolt Beranek and Inc. en su sistema BYBLOS para el reconocimiento del Discurso Continuo (Schartz, 1988) la Bell Laboratories (Rabiner, 1988) o SPHINX de la Universidad de Carnegie-Mellon (Lee, 1988).

Finalmente, en la última parte de esta década han surgido los *Modelos Neuronales o Conexionistas* en diversas aplicaciones y en concreto en Reconocimientos Automáticos del Habla (Bridle, 1987). Estos modelos, que en cierta medida pretenden ser una aproximación al comportamiento neuronal del cerebro humano, se basan en una idea que no es nueva. De hecho, ya surgió en los sesenta bajo el nombre de *Perceptrón* (Lipman, 87), aunque cayó en

desuso por ciertas limitaciones computacionales asociadas a los modelos propuestos. Una de las características que hacen atractivos a estos modelos es su capacidad de adaptación y aprendizaje que, a diferencia de otros métodos, está íntimamente unida a su facultad de reconocimiento.

Tradicionalmente, las Fuentes de Conocimiento de los sistemas de reconocimiento eran construidas ("aprendidas") de forma "manual" por el propio diseñador del sistema. Esto ha culminado en el uso de las metodologías de los Sistemas Expertos en los cuales hay una "transferencia" explícita, en forma de reglas, de conocimiento del experto al sistema. Con los Modelos de Markov por un lado y los Modelos Conexionistas por otro, se ha dado la vuelta la problema del aprendizaje, considerado bajo nuevas perspectivas. Según los nuevos enfoques, no sólo hay que transferir información al sistema, sino que éste debe ser capaz de "extraer conocimiento" del problema considerado, a partir de muestras del mismo. En un intento de llevar este enfoque a posiciones más extremas, han surgido recientemente diversas propuestas para la obtención, de forma totalmente automática, de Modelos de Markov, Redes de Estados Finitos, o Gramáticas Regulares (Thomason, 1986) (Rulot, 1986.)

## 3. EL FUTURO DEL RECONOCIMIENTO AUTOMATICO DEL HABLA.

Es una opinión extendida entre la comunidad de científicos que la verdadera comunicación hablada entre seres humanos y computadores no se logrará hasta bien entrado el siglo XXI, si es que alguna vez ello es posible. Por lo tanto deberemos contentarnos con desarrollar sistemas para aplicaciones concretas y restringidas con las herramientas que disponemos en la actualidad.

Uno de los campos en los que hay que hacer especial hincapié es en la Decodificación Acústico-Fonética comentada anteriormente, en particular, en la búsqueda de modelos robustos para la representación de unidades subléxicas adecuadas que contengan la mayor parte de la información

transitoria del habla. Por otro lado es necesario potenciar los Modelos de Markov y Neuronales y, en particular, todas las nuevas tendencias basadas en Aprendizaje Inductivo; es decir, aquellas en las que es el propio sistema el que extrae de forma automática, a partir de ejemplos, la mayor parte de la información (conocimiento) que le es requerida para su correcto funcionamiento.

## 3. BIBLIOGRAFIA

Bahl, L., Jelinek, F. & Mercer, R.1983. A maximum Likelihood Approach to Continuous Speech Recognition, *IEEE Trans. on Pattern Anal. and Machine Intelligence,* vol. 5(2), 179-190.

Baker, J.K.1975. The DRAGON System - An Overview, *IEEE trans. Acoust. Speech and Signal Proc.,* Vol 23, 24-29.

Casacuberta, F. & Vidal, E. 1987a. *Reconocimiento automático del habla.* Ed. Marcombo.

Casacuberta, F. & Vidal, E.1987b. Reconocimiento Automático del Habla: Metodología y Arquitecturas, *Inteligencia Artificial: Conceptos, Métodos y Aplicaciones.* Ed. Marcombo.

DeMori, R. & Suen, Ch.Y. (eds.).1985a. *New Systems and Architectures for Automatic Speech Recognition and Synthesis,* NATO-ASI Series. Springer-Verlag.

DeMori, R. & Laface, P.1985b. On the Use of Phonetic Knowledge for Automatic Speech Recognition, en Demori & Suen, 1985a, 569-592.

Erman, L., Hayes-Roth, F., Lesser, V. & Reddy, D. 1980. The HEARSAY-II Speech-Understandin System: Integrating Knowledge to Resolve Uncertainly, *Computing Surveys,* Vol. 12, 2, 213-253.

Fallside F. & Woods, W.A. (eds.).1985. *Computer Speech Processing.* Prentice Hall.

Flanagan, J.L.1972. *Speech Analysis Synthesis and Perception*. Springer-Verlag.

Haton, J.P. (ed.).1981. *Automatic Speech Analysis and Recognition*. Reidel Pu. Co.

Haton, J.P.1985. Knowledge-Based, and Expert Systems in Automatic Speech Recognition, en Demori & Suen, 1985a, 249-270.

Jelinek, F.1976. Continuous Speech Recognition by Statistical Methods, *Proc. IEEE,* Vol.m 64, 4, 532-556.

Klatt, D.H. 1980. Overview of the ARPA Speech Understanding Project, en Lea, 1980, 249-271.

Lea, W.A. (ed.).1980. *Trends in Speech Recognition*. Prentice-Hall.

Lee, K.F.1988. Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System, *Tec. Rep.* CMUCS-88-148.

Lippmann, R.P.1987. An Introduction to Computing with Neuronal Nets, *IEEE ASSP Magazine,* 4-22. Abril-88.

Lowerre, B. & Reddy, R.1980. The Harpy Speech Understanding System, en Lea, 1980, 340-360.

Mercier, G. (et.al.).1985. A New Rule-Based Expert System for Speech Recognition, en DeMori & Suen, 1985a, 303-342.

Moore, R.K.1985. Systems for Isolated and Connected Word Recognition, en DeMori & Suen, 1985a, pp.73-143.

Newell, A.1978. HARPY, Production System and Human Cognition, *Doc. CMU*-CS-78-140. Dpt.of Computer Science, Carnegie-Mellon University.

Niemann, H. (et.al.).1985. The Speech Understanding and Dialog System EVAR, en DeMori & Suen, 1985a, 271-302.

Niemann, H. (ed.).1988. *Recent Advances in Speech Understanding and Dialog Systems*. Springer-Verlag.

Perennou, G.1981. The ARIAL-II Speech Recognition System, en Haton, 1981.

Pierrel, J.M.1981. Etude el mise en oeuvre de contraintes linguistiques en comprehension automatiques du discours continu, Tesis de Estado. Universidad de Nancy.

Rabiner, L.R.1988. Mathematical Foundations of Hiedden Markov Models, en Niemann, 1988.

Rulot, H. & Vidal, E.1986. Modelling (sub)String-Length-Based Constraints Through a Grammatical Inference Method, *NATO ASI* en *Pattern Recognition Theory and Applications*. Spa-Balmoral. Bélgica.

Schwatz, R.M. (et.al.).1988. Acoustic-Phonetic Decoding of Speech, en Niemann, 1988.

Thomason, M.,Granum, E. & Blake, R.E.1986. Experiments in Dynamic Programming Inference of Markov Networks with string Representing Speech Data, *Recog*. vol.19 no. 5 pp. 343-351.

Vaissière, J.1985. Speech Recognition: a Tutorial, en Fallside, 1985, pp.191-242.

Vidal, E.1985a. Diversas aportaciones al Reconocimiento Automático del Habla, Tesis Doctoral, Universidad de Valencia.

Vidal,E., Casacuberta,F., Sanchis,E. & Benedi,J. 1985b. A General Fuzzy-Parsing Scheme for Speech Recognition, en DeMori, 1985a, 427-446.

Wolf, J. & Wodds, W.A.1980. The HWIM Speech Understanding System, en Lea, 1980, 316-339.