XIII Congreso de Estudios Vascos Ciencia, tecnología y cambio social en Euskal Herria = Zientzia, teknolqia eta gizarte aldaketa Euskal Herrian. (13. 1995. Zamudio
ISBN: 84-89516-07-3 .- Donostia: Eusko Ikaskuntza, 1996. p. 317-323

RECONOCIMIENTO AUTOMATICO DEL HABLA: PERSPECTIVA DESDE UN GRUPO DE INVESTIGACION DEL PAIS VASCO

Luis Javier Rodríguez *Becario del Gobierno Vasco UPV/EHU;* Amparo Varona *Becaria del Gobierno Vasco UPV/EHU;* Germá Bordel *Profesor Asociado UPV/EHU;* Inés Torres *Profesora Titular (interina) UPV/EHU;* Karmele López de Ipiña *Profesora Asociada U.P. Navarra;* Jose María Alcaide *Titular E.U. (interino) UPV/EHU*

El reconocimiento del habla continua, independiente del locutor, con grandes vocabularios y gramáticas naturales es uno de los problemas más importantes en el campo de la inteligencia artificial. El presente artículo trata de introducir las cuestiones teóricas y tecnológicas implicadas en esta tarea, desde la perspectiva de un grupo de investigación del País Vasco. El lector podrá conocer la evolución histórica de las investigaciones, la estructura de un sistema de reconocimiento y cada una de las partes que lo constituyen, así como los problemas específicos, las distintas aproximaciones formales y las soluciones propuestas en cada caso. Finalmente se realizan unas consideraciones sobre el futuro más inmediato de las investigaciones en este área.

Edozein hizlarirentzat baliogarria, hiztegi handiak eta gramatika naturalak erabiltzen dituen mintzo jarraiaren ezagutza, adimen artifizialaren arloan dagoen arazo garrantzitsuenetariko bat da. Artikulu honek, gai honi dagozkien arezo teoriko eta teknologikoak aztertuko ditu Euskal Herriko talde ikertzaile baten ikuspuntutik. Ikakurleak, ikerketen garapen historikoa, ezagutza-sistema baten egitura, sistema osatzen duren atal guztiak, zenbait arazo espezifiko, hurbilketa formalak eta kasu bakoitzean proposatzen diren soluzioak ezagutu ahal izango ditu. Azkenik, arlo honetako ikerketen etorkizun hurbila aztertuko da.

Continous speech recognition is one of the most challenging tasks in the field of artificial inteligence. Speaker independence, large vocabularies and natural language are still unsolved questions. This paper deals with some theoretical and practical aspects revolved in this area, from the perspective of a research group at the Basque Country University. Historical evolution of research group at the Basque Country Universty. Historical evolution of research and the structure of a charactenstic speech recognition system are described in first place. Then some specific problems are addressed, and different approaches, with the corresponding solutions, are reviewed. Finally we point out some research derections for the future.

1. INTRODUCCION

En las últimas décadas, empresas y centros de investigación de todo el mundo han invertido un gran esfuerzo en el desarrollo de sistemas inteligentes, es decir , sistemas con capacidad de percepción, aprendizaje, razonamiento y comunicación. En esta tarea confluyen disciplinas muy diversas, que van de la matemática aplicada a la psicología cognitiva. Por otra parte, estos esfuerzos han dado lugar a campos de investigación completamente nuevos, como la visión artificial, la síntesis y el reconocimiento del habla, la inteligencia artificial, etc. En particular, la comunicación con las máquinas, realizada típicamente a través de teclados, mandos, botones, etc. (elementos rudimentarios si los comparamos con el lenguaje hablado, forma natural de comuni-

cación entre los seres humanos), constituye uno de los principales problemas pendientes. Tratar de «civilizar» a las máquinas para que aprendan a comunicarse mediante el habla requiere no sólo dispositivos tecnológicos avanzados -equivalentes al oído y al aparato fonador- sino también desarrollos a nivel teórico -lo que propiamente denominaríamos lenguaje, que engloba aspectos gramaticales, léxicos, semánticos y prácticos-, íntimamente ligados a la inteligencia artificial.

En el pequeño mundo de la comunicación hombre-máquina mediante el habla se distinguen varios campos de investigación: síntesis de voz, reconocimiento del habla, reconocimiento del locutor, traducción de lengua hablada e identificación de la lengua. El nivel de desarrollo de cada uno de estos campos no es el mismo. Aunque podemos encontrar en el mercado sintetizadores de altas prestaciones, los sistemás de reconocimiento son aún muy restrictivos: palabras aisladas, vocabulario muy limitado, dependencia del locutor, etc. Los sistemas de reconocimiento automático del habla para grandes vocabularios, discurso continuo e independientes del locutor, así como la identificación del locutor y de la lengua y los sistemas de traducción automática se desarrollan con mayor o menor éxito en laboratorios de investigación de universidades o en departamentos de I+D de grandes empresas del mundo de la informática y de las telecomunicaciones (IBM, AT&T, Philips).

Al interés de universidades y empresas se ha sumado en los últimos años la Comunidad Económica Europea, potenciando las tecnologías relacionadas con las lenguas mediante la financiación de ambiciosos proyectos en los que colaboran grupos de diferentes países. Este esfuerzo de investigación ha producido sus frutos y en la actualidad se pueden comenzar a vislumbrar aplicaciones reales que hace solamente unos pocos años la sociedad en general consideraba como pertenecientes al mundo de la ciencia ficción. Consecuencia directa de este proceso es el creciente interés del sector empresarial por los resultados obtenidos en universidades y centros de investigación. Así en el conjunto del Estado, y en particular en la Comunidad Autónoma Vasca, han surgido en los últimos años empresas cuyo objetivo es comercializar sistemas que incorporan voz digitalizada y/o sintética e incluso algunas aproximaciones a lo que pudiera ser un sistema de reconocimiento autom tico del habla,

El desarrollo de muchos aspectos de los sistemas de reconocimiento autom tico del habla es total o parcialmente dependiente de la lengua y de la aplicación. El estado actual de la tecnología en este campo permite abordar el desarrollo de numerosos productos de interés: control de robots industriales o de electrodomésticos mediante voz, dictado automático, ayuda a discapacitados (por ejemplo, a ciegos), acceso y navegación por bases de datos, operaciones comerciales (venta de billetes en una terminal de aeropuerto), sistemas de seguridad (controles de entrada mediante voz), operaciones telefónicas automáticas, etc. Los sistemas multimedia, que permiten la transmisión y el tratamiento simultáneo de voz, imagen y datos en tiempo real, presentan un amplio abanico de posibilidades en nuestro entorno industrial: trabajo en casa, disminución de costes de transporte y almacenamiento de datos, tele-reuniones, acceso rápido e integrado a bloques selectivos de información, etc. El pleno desarrollo de estos sistemas requiere un módulo eficiente de reconocimiento automático del habla. En el caso de lenguas con mercado (inglés, castellano, francés, alemán, chino, etc.), el concurso de las multinacionales puede permitir un desarrollo más rápido de estos sistemas, independiente de la inversión institucional. Pero en el caso de lenguas de implantación local, con relativamente pocos hablantes -caso del euskera-, el interés comercial de las multinacionales es más bien escaso, y los sistemas adaptables que ofertan como solución no proporcionan un rendimiento óptimo. Por ello, la superación de las numerosas restricciones de los sistemas actuales requiere un esfuerzo conjunto por parte de los laboratorios de investigación, el sector empresarial y la Administración.

2. EL PROBLEMA DEL RECONOCIMIENTO AUTOMATICO DEL HABLA

Bajo la denominación de Reconocimiento Automático del Habla (RAH) se consideran, en realidad, tareas de diferente complejidad: reconocimiento de palabras aisladas, reconocimiento de palabras conectadas, identificación de palabras clave en discurso continuo, reconocimiento de discurso continuo, etc. Son las características propias del habla: continuidad temporal, variabilidad en la pronunciación, redundancia informativa y multi-interactividad de niveles de conocimiento y de capacidades perceptivas, que afectan negativamente al proceso de reconocimiento, las que nos obligan a imponer ciertas, entre ellas acotar el nivel de ruido o las condiciones ambientales esperables en la señal de entrada, establecer el tamaño del vocabulario (reducido, medio o grande), la dependencia o independencia del locutor, la capacidad de adaptación a nuevos locutores, el tipo de gramática de la aplicación, etc,

2.1. Antecedentes históricos

Los orígenes del RAH [Rabiner93] se remontan a los años 40, momento en el que se desarrollan los primeros espectrógrafos, que permítian observar el espectrograma de una señal, es decir, la evolución temporal de la energía en las distintas bandas de frecuencia, dato que podía servir para caracterizar y reconocer la voz humana. El primer dispositivo automático de reconocimiento, desarrollado en 1952 en los laboratorios Ben para distinguir los diez dígitos ingleses pronunciados de forma aislada por un único locutor, se basaba en identificar las frecuencias de resonancia de la parte vocálica de los dígitos. En otro trabajo paralelo e independiente (Laboratorios RCA, 1956) se trataba de reconocer sílabas mediante distancias espectrales obtenidas a partir de un banco de filtros analógico. Otros trabajos de la época también se basaban en dispositivos analógicos que obtenían información acerca del contenido espectral de las señales, y utilizaban como criterio de clasificación las frecuencias de resonancia de las vocales.

En los años 60 se publicaron ideas fundamentales sobre reconocimiento de patrones. De hecho, los primeros trabajos que emplearon medios informáticos aplicados al RAH datan de esta época. Estos trabajos se centraban en el reconocimiento de palabras aisladas monolocutor, y utilizaban técnicas de programación dinámica para comparar la secuencia de vectores de entrada, mediante alineamiento temporal no lineal («Dynamic Time Warping», DTW), con los patrones de las palabras del diccionario. Ademas de RCA y AT&T, entran en escena los laboratorios japoneses de NEC, a los que se suman los trabajos realizados en la Carnegie Mellen University (CMU), que continuarán hasta nuestros días.

En los años 70, parcialmente solucionada la tarea de reconocimiento de palabras aisladas, se atacó con optimismo la tarea de reconocimiento de discurso continuo. Se

perfilaban dos aproximaciones al problema: los Modelos Estructurales Estocásticos (MEE) y los Sistemas Basados en el Conocimiento (SBC). Se abordaron grandes proyectos de investigación, de los cuales el más ambicioso y conocido de la historia del RAH -que tomaba como referencia la segunda de las aproximaciones-, fue el proyecto ARPA-SUR («Advanced Research Projects Agency - Speech Understanding System»), iniciado en 1971 y financiado por el Departamento de Defensa de los Estados Unidos, proyecto que nunca alcanzó sus objetivos pero que contribuyó en gran medida a conocer los mecanismos de producción del habla y a tomar conciencia sobre la verdadera magnitud del problema. En IBM se forma también un grupo de reconocimiento del habla, que ataca varias tareas sobre grandes vocabularios. En este grupo se apuesta desde el principio por sistemas estadísticos-probabílisticos basados en aprendizaje inductivo. Mientras tanto, en AT&T continúan las investigaciones con palabras aisladas y DTW, ahora tratando de obtener sistemas independientes del locutor, para lo cual se desarrollan algoritmos de agrupamiento de muestras que generen patrones robustos.

Demostrada en los primeros 80 la ineficacia de los SBC, se invierte todo el esfuerzo en desarrollar sistemas capaces de extraer conocimiento de forma inductiva, es decir, a partir de muestras, Se utiliza a partir de entonces, siguiendo los trabajos de IBM, una modelización acústica basada en Modelos Ocultos de Markov («Hidden Markov Models», HMM), discretos y continuos, y se optimizan los algoritmos de aprendizaje para entrenar los sistemas a partir de grandes bases de datos. Se mejoran también los sistemas de DTW para el reconocimiento de palabras conectadas, más concretamente se desarrollan algoritmos de búsqueda eficientes con los que determinar la secuencia optima de patrones para una secuencia de vectores acústicos.

Mediada la década de los 80 se presenta la aproximación conexionista como alternativa a la aproximación estadístico-probabilística. Las redes neuronales artificiales («Artificial Neural Networks», ANN) comparten con los HMMs su carácter inductivo, es decir, el aprendizaje a partir de muestras, pero sus configuraciones clásicas -como los perceptrones multicapa («Multi-Layer Perceptron», MLP)- no son capaces de representar fenómenos dinámicos como la señal de voz, por lo cual tuvieron que desarrollarse arquitecturas recursivas especificas, con objeto de superar estas limitaciones. Otros autores han optado por configuraciones híbridas en las que un MLP es utilizado para estimar las probabilidades de emisión de un HMM [Renals94].

Actualmente la investigación se concentra, por un lado, en mejorar el rendimiento de la modelización acústica (generación automática de unidades acústicas contextuares, entrenamiento discriminativo de los modelos) y, por otro lado, en la integración de niveles de conocimiento superiores (estrategias de búsqueda heurísticas en grandes autómatas que representan modelos del lenguaje). también se está invistiendo un gran esfuerzo en diseñar y adquirir grandes bases de datos para el entrenamiento de sistemas de reconocimiento de discurso continuo.

2.2. Descripción de un sistema de RAH

Un sistema de RAH se compone de una etapa de Preproceso (modelado de la señal vocal), una etapa Acústico-Fonética (modelado acústico de unidades subléxicas y/o léxicas) y una etapa Sintáctico-Semántica (modelado del lenguaje). Estas dos últimas pueden combinarse de forma secuencial o integrarse en un único módulo.

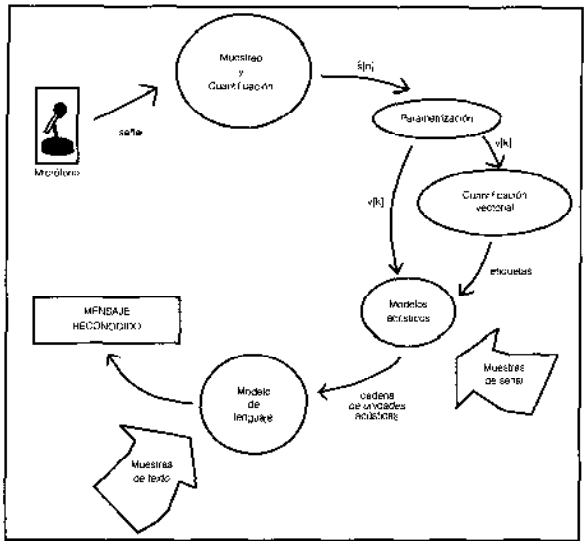

FIGURA 1 **: Flujo de reconocimiento en un sistema de RAH.**

El sistema parte de una señal de voz, a la cual se aplican técnicas de procesamiento de señal y reconocimiento de patrones para generar una cadena de unidades acústicas y a continuación, dependiendo de la aplicación, producir una interpretación semántica que, a su vez, genere una acción u otro tipo de representación de alto nivel. Habitualmente estas etapas se realizan secuencialmente mediante módulos independientes que por sí mismos definen áreas específicas de investigación. Así, el módulo de adquisición proporciona una señal digital , muestreada a intervalos fijos y con niveles discretos de cuantificación. Esta señal constituye la entrada del módulo de parametrización, donde se realiza un análisis en tiempo y en frecuencia de sucesivos segmentos de señal, generando como salida una secuencia de vectores de parámetros acústicos Esta transformación debe producir una representación compacta con una pérdida mínima de información. A continuación se procede a la decodificación acústico-fonética (DAF) de la secuencia de vectores, es decir, a la generación de la secuencia óptima de unidades acústicas -habitualmente fonemas- según el conjunto de modelos disponible. En muchos sistemas (por ejemplo, en los HMM discretos), entre el modulo de parametrizacion y el módulo DAF se inserta un modulo de cuantificación vectorial («Vector Quantization», VQ), proceso en el que cada vector de características es sustituido por la etiqueta del vector centroide más cercano según una cierta métrica definida en el espacio de los parámetros acústicos. En este proceso se pierde parte de la información, pero también se reduce drásticamente el coste computacional del modulo de DAF. Como paso previo, deberá generarse el diccionario de centroides, con el criterio de minimizar el error de cuantificación para un conjunto estadísticamente significativo de vectores (habitualmente se utiliza el algoritmo LBG [Linde80]). El módulo de modelización del lenguaje toma como entrada la secuencia optima de fonemas, o típicamente las N secuencias «mejores», a partir de las cuales, y según unas ciertas reglas gramaticales más o menos restrictivas, se determinará la secuencia «correcta» y, por tanto, el mensaje que pretendía extraerse de la señal acústica. La figura 1 ilustra las distintas etapas del proceso.

2.3. Parametrización

Los algoritmos de reconocimiento de patrones requieren reducir drásticamente el volumen de datos de la señal de voz. Para ello deberá eliminarse toda información redundante o inútil, y mantener sólo información relevante, a ser posible mediante un número pequeño de parámetros. Esta discriminación de información y reducción del volumen de datos es lo que trata de efectuarse, en primer término, mediante la parametrización. La primera etapa de procesamiento consiste en la conversión A/D de la señal de voz (filtrado «antialiasing», muestreo y cuantificación). De ella se obtiene una secuencia de números, tratable computacionalmente, que no contiene toda la información acústica de la señal de voz original, pero sí toda la información que interesa a efectos de reconocimiento.

Esta secuencia de números se divide en pequeños segmentos consecutivos y solapados, cada uno de los cuales se analiza por separado y produce un vector de parámetros o vector característico. Pueden aplicarse distintos tipos de análisis [Rabiner78], ya sea en el tiempo (energía, cruces por cero), en la frecuencia (banco de filtros, transformada de Fourier), o paramátricos (predicción lineal). La representación más utilizada consiste en la transformada inversa de Fourier del log-espectro, denominada «cepstrum», que modeliza la articulación del tracto vocal (envolvente espectral) mediante su respuesta impulsional. Es posible generar un único vector reuniendo dos o más representaciones distintas. De hecho, la representación paramétrica suele incluir primeras y segundas derivadas que ayudan a describir el carácter dinámico de la señal de voz. Se trata de almacenar el máximo de información en un espacio mínimo. Para conseguir este objetivo, en ocasiones se aplican técnicas de análisis lineal discriminante («Linear Discriminan Analysis», LDA), que generan una representación paramétrica reducida, formada por aquellas componentes del vector característico que más información aportan en el proceso de reconocimiento. Otra técnica empleada habitualmente es el denominado filtrado cepstral o «liftering», que modula la envolvente espectral y disminuye su sensibilidad a la posición exacta de las frecuencias de resonancia, permitiendo por tanto una menor dependencia del locutor y de las condiciones de transmisión de la señal (ruidos, etc.).

2.4. Modelización acústica

En el diseño de la etapa DAF se plantean dos problemás fundamentales: la elección de un conjunto de unidades subléxicas y la elección del entorno matemático adecuado para modelar esas unidades. La mayor parte de los sistemás de DAF desarrollados inicialmente trabajaban con un conjunto de unidades independientes del contexto, muy próximo al conjunto de fonemas de la lengua. Sin embargo, pronto se vio la necesidad de ampliar este conjunto básico de unidades para tratar de recoger la variabilidad contextual relevante en la discriminación acústica [LeeCH90]. La introducción de contextos ha producido incrementos notables de las tasas de reconocimiento en sistemas para grandes vocabularios y/o discurso continuo [LeeKF90a] [Bah191a]. Existe evidencia experimental de que cuanto más detallado resulta el modelado acústico, mejor es el rendimiento del sistema de reconocimiento. No obstante, si se selecciona un número grande de unidades se produce el inconveniente de que los modelos de las unidades dejan de estar bien entrenados, ya que el número de parámetros crece excesivamente. De esta forma aparece la necesidad de establecer criterios que permitan obtener inventarios de unidades con una adecuada modelización de la coarticulación y compatibles con un entrenamiento robusto y un sistema de reconocimiento de coste computacional tolerable [LeeKF90b] [Ney90] [BahI91 b] [Hwang93].

En cuanto a la metodología, los HMMs constituyen la elección más extendida, tanto en su formulación discreta [LeeKF89], que trabaja con etiquetas o símbolos, como en la continua [Rabiner89] y semicontinua [Huang92], que trabajan directamente con vectores acústicos. Para la decodificación se utiliza el algoritmo de Viterbi [Forney73] -similar a los algoritmos de programación dinámica utilizados en DTW-, que produce la secuencia de estados óptima dados una secuencia acústica y un conjunto de modelos.

Desde la introducción de los HMMs se han propuesto multitud de mejoras tanto a nivel de la propia estructura, como de las técnicas utilizadas para reestimar los parámetros de los modelos. El criterio de reestimación de parámetros más utilizado es el de máxima similitud («Maximum Likelihood Estimation», MLE) [BahI83], que maximiza la probabilidad de las muestras con respecto a su propio modelo. Con objeto de mejorar la discriminación entre los modelos, se han propuesto varios criterios alternativos: Máxima Información Mutua (MMI) [BahI86], Mínima Información Discriminante o Entropía Cruzada (MDI) [Ephraim90], Maximum A Posteriori (MAP) [Gauvain94], Aprendizaje Correctivo (CT) [BahI88], etc. Tiene particular interés el criterio de Minimización del Error de Clasificación (MCE/GPD) [Juang92] [Chou92], que parece producir los modelos más discriminantes, y además es coherente con el objetivo de disminuir las tasas de error del sistema de reconocimiento.

2.5. Modelización del lenguaje

El bloque de Modelización del Lenguaje (ML) trata de aplicar las reglas gramaticales que rigen la comunicación hablada en una determinada tarea, para facilitar el reconocimiento de la cadena de unidades acústicas generada por el modulo DAF, o la comprensión de un mensaje a partir de dicha cadena. Para ello tiene en cuenta no sólo el contenido léxico y las reglas sintácticas, sino también aspectos prácticos y semánticos.

Los MLs más utilizados son los n-gramas [Jelinek91a] (en particular, los bigramas y los trigramas), que describen la probabilidad de observar una determinada palabra dadas las n-1 anteriores. Su popularidad es debida a que existen métodos de aprendizaje automático de n-gramas a partir de texto. Los n-gramas son capaces de capturar un gran porcentaje de fenómenos sintácticos y semánticos y pueden incorporarse fácilmente en los algoritmos de reconocimiento más utilizados. Además, los bigramas no añaden ningún coste computacional en el reconocimiento con respecto al sistema sin modelo del lenguaje.

Como técnicas alternativas a los n-gramas, podemos citar ciertas metodologías de aprendizaje automático de modelos estructurales (ECGI y MGGI, [Garcia90] [Casacuberta90]) y las denominadas Gramáticas Incontextuales Probabilísticas («Stochastic Context-Free Grammars», SCFG) [Jelinek91 b], que han despertado un notable interés en los últimos años, debido a su capacidad de especificación de restricciones sintácticas complejas en el marco de un formalismo estocástico. Las SCFG permiten establecer dependencias estadísticas a largo término entre las unida-

des lingüísticas, aunque presentan dos importantes inconvenientes: un coste computacional (espacial y temporal) superior al de los clásicos modelos de n-gramas, y el problema del aprendizaje automático (estructural y probabilístico) de los modelos [Casacuberta94].

Existen, en general, dos grandes aproximaciones para definir la conexión entre el modelo de lenguaje y el módulo DAF en un sistema de RAH. En la primera, DAF y ML son módulos independientes; es decir, el módulo DAF es empleado para producir una cadena, o una lista de las N-mejores cadenas, o una malla de unidades lingüísticas, y el ML es aplicado para seleccionar la secuencia correcta. La segunda aproximación impone restricciones gramaticales durante el proceso de reconocimiento mediante un ML integrado.

Los MLs con vocabularios de más de mil palabras generan autómatas enormes, que suelen necesitar grandes recursos computacionales, lo cual muchas veces los hace inviables. Una de las primeras técnicas que se utilizaron para reducir el coste computacional es la conocida como Búsqueda en Haz («Beam Search») [Haeb93]. Otra de las técnicas más utilizadas esta basada en el algoritmo A* [Kenny93], que permite una exploración inteligente (heurística) del grafo de estados. Recientemente se ha propuesto una nueva técnica de exploración conocida como Búsqueda Hacia Adelante / Hacia Atrás («Forward-Backward Search») [Austin91]. Las técnicas mencionadas se utilizan en el marco del denominado análisis por Viterbi en arquitecturas básicamente integradas. Este análisis está basado en la búsqueda del camino de mayor probabilidad (o de los N caminos o secuencias de caminos de mayor probabilidad) [Shwartz90] [Shwartz91] en un grafo integrado de estados. Algunas de estas técnicas se emplean también en el marco del análisis sintáctico-estocástico, basado en la búsqueda de la secuencia de modelos de mayor probabilidad, donde se conocen como Decodificación por Pila («Stack Decoding») [Jelinek76] [Jelinek69]. Finalmente, la técnica de Comparación Rápida («Fast Match») [Bah193] permite seleccionar unas cuantas de todas las posibles alternativas que existen en un punto del grafo de estados, para su exploración posterior mediante alguno de los métodos citados, ya que la exploración exhaustiva de todas las posibles alternativas sería, como se ha dicho, demasiado costosa computacionalmente.

Prácticamente la totalidad de los Modelos de Lenguaje (ML) toman como unidad la palabra. El único motivo que ha llevado a ello es la sencillez de procesamiento al tratar con texto escrito, si bienes lógico plantearse otro tipo de unidades léxicas menos restrictivas que lleven a la obtención de modelos más sencillos cuando se trata de lenguaje oral. En el caso de una tarea de RAH con grandes vocabularios, la ventaja de tratar con texto escrito desaparece y la utilización de unidades alternativas puede plantearse prácticamente con un mismo coste computacional. Parece lógico, pues, que poco a poco vayan imponiéndose este tipo de unidades, más adecuadas a la tarea de modelización del lenguaje hablado.

3. EL GRUPO DE RECONOCIMIENTO AUTOMATICO DEL HABLA DE LA UPV/EHU

El Grupo de Reconocimiento Automático del Habla / Mintzo-Ezagutza Automatikoaren TaIdea (GRAH-MEAT) de la UPV/EHU, instalado físicamente en el Departamento de Electricidad y Electrónica de la Facultad de Ciencias de Leioa y formado por las personas que firman este artículo, inauguró su andadura en 1991, con la presentación de la primera tesis doctoral sobre el tema en la Comunidad Autónoma Vasca. Inicialmente el grupo estuvo interesado por la aproximación deductiva, aplicando esta metodología a la tarea de decodificación acústico-fonética. Posteriormente, a la vez que otros muchos laboratorios de investigación, el grupo decidió adoptar la metodología inductiva, aproximación que ha proporcionado resultados de reconocimiento muy superiores a los de los sistemas basados en reglas.

Desde entonces el grupo ha colaborado con grupos similares de otras universidades. En particular, ha mantenido durante estos años una estrecha relación con el Grupo de Reconocimiento de Formas e Inteligencia Artificial (GRFIA) del Departamento de Sistemas Informáticos y Computación (DSIC) de la Universidad Politécnica de Valencia -el grupo del Estado con más experiencia en el área. Esta colaboración se ha concretado en la realización de un gran número de publicaciones internacionales en común, diseño conjunto de líneas de investigación e intercambio de software y datos. Así, próximamente tendrá lugar en nuestro grupo la lectura de una tesis doctoral codirigida por un investigador del GRAH-MEAT y otro del GRFIA. En la actualidad el GRAH-MEAT participa en un ambicioso proyecto de I+D, financiado por la CICYT, junto a otras universidades españolas (Universidad Politécnica de Valencia, Universidad Politécnica de Cataluña y Universidad de Zaragoza). A otros niveles, el grupo también ha colaborado con la Universidad pública de Navarra y la Universidad Politécnica de Madrid. Este tipo de actuaciones, así como la participación activa en congresos de ámbito estatal (AERFAI, SEPLN), con la presentación de comunicaciones y conferencias y el trabajo en comités de programa, constituyen nuestra aportación a la necesaria colaboración e intercambio de conocimientos entre grupos con amplia experiencia en el área dentro de la comunidad científica, colaboración que consideramos imprescindible para el desarrollo tecnológico propio.

A nivel internacional hemos realizado numerosas publicaciones científicas y participado en los congresos más importantes del área celebrados en Estados Unidos (ICASPP), Japón (ICSLP) y diferentes países europeos (EUROSPEECH, NATO-ASI). En los primeros años de vida del grupo se realizaron también estancias y visitas a centros de investigación europeos. Más recientemente un miembro de nuestro grupo ha participado en un proyecto intensivo de trabajo financiado por el gobierno de los Estados Unidos.

El GRAH-MEAT participa en el Máster de Electrónica y Automática impartido en el Departamento de Electricidad y Electrónica de la Facultad de Ciencias de la UPV/EHU, en colaboración con la Asociación de Industrias Electrónicas del País Vasco (AIEPV). El programa docente de este Máster incluye disciplinas relacionadas con el reconocimiento automático del habla, impartidas por miembros de nuestro grupo y por reconocidos especialistas en el tema de las Universidades Politécnicas de Cataluña, Madrid y Valencia. Dentro de este marco se han desarrollado proyectos académicos y de I+D en colaboración con algunas empresas de la AIEPV: Indelec, Filosoft, Natural Vox, etc. que han dado lugar al desarrollo de sistemas de demostración de RAH para aplicaciones restringidas.

Nuestras líneas de investigación a corto y medio plazo incluyen algunos de los problemas abiertos mencionados al principio: aprendizaje discriminativo de HMMs, generación

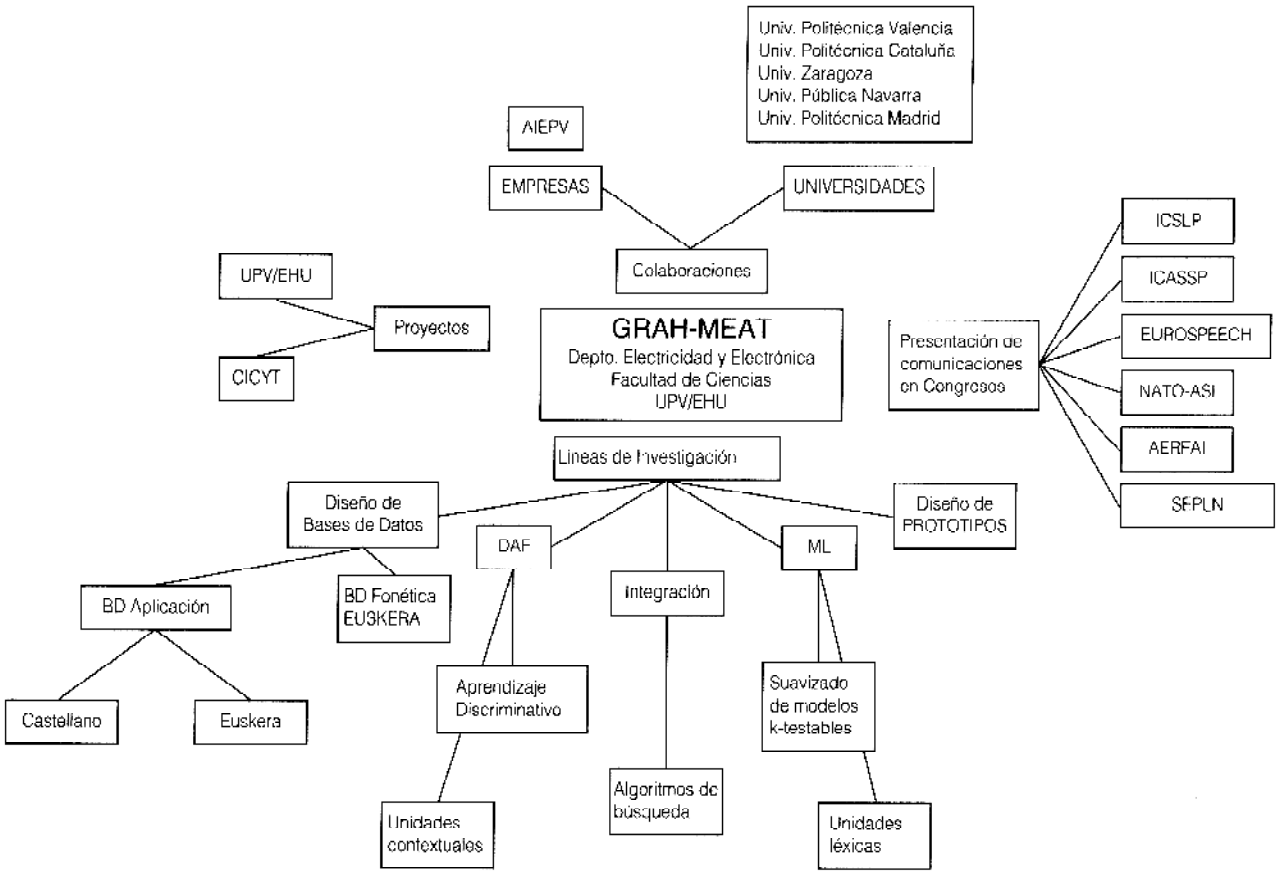

FIGURA2: **Diagrama de actividades del GRAH-MEAT.**

automática de unidades contextuares para DAF, modelización del lenguaje mediante una aproximación inductiva, con una propuesta preliminar de unidades léxicas (alternativas a las palabras), algoritmos de búsqueda eficientes para sistemas con grandes vocabularios, y particularmente el diseño y elaboración de bases de datos vocales, ya que, como se ha dicho, disponer de grandes bases de datos se ha convertido en factor clave para el desarrollo de sistemas eficientes. En este sentido, se halla en fase de finalización una base de datos fonética en euskera, diseñada y adquirida íntegramente por el GRAH-MEAT. Los primeros experimentos de DAF sobre esta base de datos tendrán lugar próximamente. Por otra parte, se han realizado ya las gestiones para la adquisición de una base de datos vocal -en castellano y en euskera- para una aplicación de operadora telefónica, que será utilizada tanto en DAF como en ML. Finalmente, dentro del proyecto CICYT citado anteriormente, se trabajará en la integración de los módulos de Preproceso, DAF y ML, para la consecución de un prototipo integrado funcionando «on-line» sobre una estación de trabajo,

4. CONCLUSIONES

El reconocimiento automático del habla ha experimentado un gran avance en los últimos años. De no poder reconocer más que un conjunto pequeño de palabras aisladas (típicamente dígitos y unos pocos comandos o palabras clave) pronunciadas por un único locutor, se ha pasado a sistemas de vocabularios medios y grandes, capaces de reconocer palabras conectadas, o de identificar palabras clave en discurso continuo, con bastante independencia del locutor. Hoy día existen máquinas de dictado automático, aunque no demasiado sofisticadas; los sistemas de seguridad mediante voz, de reconocimiento o verificación del locutor, están en plena expansión comercial, y algunos de los ordenadores personales multimedia aparecidos recientemente, capaces de combinar voz, imagen y datos en una pantalla, son también capaces de procesar órdenes habladas.

Pero quedan muchas tareas sin resolver. Actualmente, están en poceso de desarrollo sistemas de discurso continuo, con lenguajes restringidos, para el acceso automático a bases de datos, lo cual puede permitir aplicaciones como la venta automática de billetes en un aeropuerto, o la navegación interactiva por Internet mediante voz. Las investigaciones apuntan a sistemas de discurso continuo, multilocutor, con grandes vocabularios, tal vez con gramáticas restrictivas, que poco a poco irán acercándose más al lenguaje natural. Tareas como el diálogo, que mejoraría el rendimiento de los sistemas interactivos- y la traducción automática -de particular interés para la Comunidad Económica Europea- también están en fase de investigación.

En los proximos años, el GRAH-MEAT va a tratar todas estas cuestiones en el marco de un proyecto de colaboración con grupos de otras universidades, poniendo particular interés en el desarrollo de sistemas de reconocimiento del euskera. En concreto, la etapa de decodificación acústico-fonética va a mejorarse introduciendo unidades acústicas contextuares y algoritmos de entrenamiento discriminativos. Se van a desarrollar también modelos de lenguaje más robustos extraídos directamente de la lengua hablada en aplicaciones concretas. Ambas etapas deberán combi-

narse en un único sistema integrado, utilizando algoritmos de búsqueda eficientes.

5. REFERENCIAS BIBLIOGRAFICAS.

[Austin91] S. AUSTIN, R. SCHWARTZ, P. PLACENCY. «The Forward-Backward Search Algorithm», Proc. lEEE Int. Conf. Acoustic, *Speech and Signal Processing,* pp. 697-700, 1991

[BahI83] L R. BAHL, F. JELINEK, R.L. MERCER. «A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Trans on Pattern Analysis and Machine Intelligente,* Vol. 5, N. 2, pp. 179-190. March 1983.

[BahI86] L.R BAHL, P.F. BROWN, P.V. DE SOUZA, R.L. MERCER «Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition». *Proc. IEEE ICASSP-* 86, pp. 49-52

[BahI88] L.R. BAHL, P.F. BROWN, P.V. DE SOUZA, R.L. MERCER «A New Algorithm for the Estimation of Hidden Markov Model Parameters», Proc. IEEE ICASSP-88, pp. 493-496.

[BahI91a] L.R. BAHL, P.V. DE SOUZA, P.S. GOPALAKRISHNAN, D. NAHAMOO, M.A. PICHENY. «Decision Tress for Phonological Rules in Continuos Speech», Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing, pp. 185-188, 1991.

[BahI91b] L.R. BAHL ET AL. «Context dependent modeling of phones in continuous speech using decision trees», Proc. DARPA Speech, Natural Language Workshop, Feb. 1991, pp. 264-269, 1991.

[BahI93] L R. BAHL, S.V. DE GENNARO, P.S. GOPALAKRISHNAN, R.L. MERCER. «A Fast Approximate Acoustic Match for Large Vocabulary Speech Recognition», IEEE Transactions on Speech and Audio Processing, Vol. 1, nº 1, pp. 59-67, 1993

[Casacuberta90] F. CASACUBERTA, E. VIDAL, H. RULOT, B.MAS. «Learning the structure of HMM's through grammatical inference techniques», Int. Conferencie on Acoustic, Speech and Signal Processing, Albuquerque, pp. 717-720, 1990

[Casacuberta94] F. CASACUBERTA. «Statistical Estimation of Stochastic Context-Free Grammars Using the Inside-Outside Algorithm and a Transformation on Grammars», LNAI-862 Gramatical Inference and Applications, pp. 119-129, 1994.

[Chou92] W. CHOU, B. H. JUANG, C. H. LEE Segmental GPD training of HMM based speech recognizer», Proc. ICASSP-92, pp. 473-476, 1992.

[Ephraim90] Y.EPHRAIM, L.R. RABINER. «On the Relations Between Modeling Approaches for Speech Recognition». IEEE Trans. on Information Theory, Vol. 36, N. 2, pp. 372-380. March 1990.

[Forney73] G.D. FORNEY «The Viterbi algorithm». Proc. IEEE, N. 61, pp. 268-278. March 1973.

[García90] P. GARCIA, E. SEGARRA, E. VIDAL, I. GALIANO. «On the use of the morphic generator gramatical inference (MGGI) methodology in automatic speech recognition, International Journal on Pattern Recognition and Artificial Intelligence, Vol.4, nº.4, pp.667-685, 1990.

[Gauvain94] J.L. GAUVAIN, CH. LEE. «Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chams» lEEE Trans. on Speech and Audio Processing, Vol. 2, N. 2, pp. 291-298 April 1994.

[Haeb93] R. HAEB-UMBACH, H. NEY. «Improvements in Beam-Search for 10.000 word Continuos Speech Recognition», lEEE Transc ASSP, Vol. 2, pp. 353-356, 1993.

[Huang92] X.D. HUANG «Phoneme Classification Using Semicontinuous Hidden Markov Models», IEEE Transactions on Signal Processing, Vol. 40, nº 5, pp. 1062-1067, 1992.

[Hwang93] M-Y HWANG, X. HUANG. «Shared-Distribution Hidden Markov Models for Speech Recognition», lEEE Transc. SAP-1, nº 4, pp. 414-420, 1993.

[Jelinek69] F. JELINEK. «Fast Sequential Decoding Algorithm using a Stack», IBM. J. Res. Develop., pp. 675-685. Nov, 1969.

[Jelinek76] F. JELINEK, Continuos Speech Recognition by Statistical Methods», Proc. of the lEEE, Vol. 64, nº 4, pp. 532-566, 1976.

[Jelinek91a] F. JELINEK. «Up from trigrams: the struggle for improved language models», Proc. of the Eurospeech 91; Genova, Italy, pp.1037-1039, sep. 24-26, 1991.

[Jelinek91b] F. JELINEK, J. D. LAFFERTY. Computation of the Probability of Initial Substring Generation by Stochastic Context-Free Grammars», Computational Linguistics, Vol. 17, nº 3, 1991.

[Juang92] B.H. JUANG, S. KATAGIRI. «Discriminative Learning for Minimum Error Classification». IEEE Trans. on Signal Processing, Vol. 40, N. 12, pp. 3043-3054. December 1992.

[Kenny93] P. KENNY, R. HOLLAN, V. GUPTA, M. LENNING, P. MERMELSTEIN, D. O'SHAUGHNESSY. «A*-Admissible Heuristics for Rapid Lexical Access», IEEE Trans. on Speech and Audio Processing, Vol. 1, nº 1, pp. 49-58, 1993.

[LeeCH90] C.H. LEE, L.R. RABINER, R. PIERACCINI, J. G. WILPON, «Acoustic modeling for large vocabulary speech recognition», Computer Speech and Language 4, pp. 127-165, 1990.

[LeeKF89] K.F. LEE. «Automatic Speech Recognition. The development of the SPHINX System». Kluwer Academic Publishers. 1989.

[LeeKF90a] K.F. LEE, Context-dependent Phonetic Hidden Markov models for Speaker-Independent Continuos Speech Recognition», IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. 38, pp. 599-609, 1990.

[LeeKF90b] K-F LEE ET AL. «Allophone clustering for continuos speech recognition», Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing, pp.749-752, 1990.

[Linde80] Y. LINDE, A. BUZO, R.M. GRAY. «An Algorithm for Vector Quantizer Design», IEEE Transactions on Communications, Vol. COM -28, nº 1, 1980.

[Ney90] H. NEY: «Acoustic modeling of phoneme units for continuos speech recognition», Proc. EUSIPCO'90, pp.65-72, 1990.

[Rabiner78] L.R. RABINER, R.W. SCHAFER. «Digital Processing of Speech Signals». Prentice-Hall. 1978.

[Rabiner89] L.R. RABINER. «A Tutorial on Hidden Markov models and selected applications in Speech Recognition», Proceedings of the IEEE, Vol 77, pp. 257-286, 1989.

[Rabiner93] L.R. RABINER, B-H. JUANG. Fundamental of Speech Recognition, pp. 321-387, 1993.

[Renals94] S. RENALS, N. MORGAN, H. BOURLARD, M. COHEN, H. FRANCO «Connectionist Probability Estimators in HMM Speech Recognition, IEEE Trans. on SAP, Vol. 2, nº 1, Part ll, 1994.

[Schwartz90] R. SCHWARTZ, Y-LU CHOW. «The N-best Algorithm: An efficient and exact procedure for finding the N most likelihood Sentence Hypotheses», Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing, pp. 81-84, 1990.

[Schwartz91] R. SCHWARTZ, S. AUSTIN. A comparison of several approximate algorithms for finding multiple (N-Best) Sentence Hypotheses», Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing, pp. 701-704, 1991.