CONSTRUCCION DE SISTEMAS DE RECONOCIMIENTO DEL HABLA MEDIANTE TECNICAS DE APRENDIZAJE AUTOMATICO*

Enrique Vidal, Pablo Aibar, José Miguel Benedí, Francisco Casacuberta, Pedro García, Natividad Prieto, Héctor Rulo †, Emilio Sanchís, Encarna Segarra

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
†Centro de Informática, Universidad de Valencia

Resumen

Se presenta el planteamiento, objetivos y estado actual de un proyecto subvencionado por la Comisión Interministerial de Ciencia y Tecnología durante el periodo 1990-1992. En este proyecto se propone el uso extensivo de técnicas de Aprendizaje Automático de Modelos Estructurales para el desarrollo de Sistemas de Decodificación Acústico-Fonética, así como la construcción de Modelos de Lenguaje en niveles Sintáctico-Semánticos. Aunque algunas de estas técnicas han sido ya puestas a punto en trabajos previos (del propio equipo investigador y/o de otros autores), otras requieren aún desarrollos fundamentales y/o trabajos de adaptación antes de ser adecuadamente aplicables a los problemas propuestos. Estos desarrollos son también objeto del trabajo presentado. Por otra parte, dado que el uso de técnicas de Aprendizaje Automático exige disponer de grandes conjuntos de datos (vocales) de entrenamiento, se ha planificado una parte significativa de recursos para la obtención de estos datos. Finalmente, se prevé la realización de los prototipos necesarios para demostrar el alcance final de las técnicas y métodos desarrollados.

1. Breve perspectiva histórica

El Reconocimiento Automático del Habla (RAH) forma parte del área conocida como "Percepción" en Inteligencia Artificial. El objetivo final del RAH es el permitir la comunicación oral hombre-máquina.

Los primeros trabajos en RAH se desarrollaron en la década de los 50 en EEUU. Sin embargo, no se produjo un empuje importante hasta los años 70 con el proyecto norteamericano ARPA-SUR (Advanced Research Project Agency - Speech Understanding Research). También en Japón, Francia, Italia y Alemania surgieron en los años 70 multitud de grupos y proyectos en RAH, tanto públicos como privados. Ya en los años 80, se incorporaron otros países como Inglaterra, España, Bélgica, etc.

En España, los grupos más importantes que se dedican al RAH proceden de grupos que venían trabajando en la década de los 70 en Tratamiento de Señales y/o Informática. Los primeros grupos surgieron en la ETSIT de Madrid; en 1980 aparece un nuevo grupo en la Universidad de Valencia, que en 1986 se traslada al Departamento de Sistemas Informáticos y Computación de la Universidad Politécnica de Valencia.

* Proyecto subvencionado por la Comisión Interministerial de Ciencia y Tecnología (TIC 448/89)

Finalmente, surge otro grupo en el Departamento de Teoría de la Senyal i Comunicació de la ETSIT de Barcelona. Existen asimismo otros grupos más recientes en la ETSIT de Madrid, Universidad de Granada, Canarias, País Vasco, etc.

2. Antecedentes directos y relación con otros trabajos

Como antecedente más inmediato hay que referirse al proyecto "Reconocimiento Automático del Habla" PA85-0086, subvencionado por la (extinta) Comisión Asesora de Investigación Científica y Técnica (CAICYT),y realizado esencialmente por el mismo equipo que el del proyecto aquí presentado durante el periodo 1987-1989. Durante la evolución del mencionado proyecto fueron cubriéndose sus objetivos y aparecieron a su vez nuevos temas y necesidades interesantes. Del examen de los resultados obtenidos en cada objetivo, así como de las nuevas y prometedoras posibilidades que fueron surgiendo, y de las nuevas necesidades que estas posibilidades plantearon, nació el presente proyecto, cuyos objetivos generales podrían sintetizarse en: (A) Extensión del uso de técnicas de Aprendizaje Automático ya desarrolladas y desarrollo de nuevas técnicas más potentes; y (B) Particularización de estas técnicas en los niveles Acústico-Fonético y Sintáctico-Semántico.

Obviamente, no solo el citado proyecto fue originario de los planteamientos del proyecto actual, sino que los resultados de otros muchos trabajos, independientes de los nuestros, han contribuido también significativamente a configurar estos planteamientos.

El RAH ha alcanzado en la actualidad cierta madurez metodológica en la que la mayor parte de los sistemas se desarrollan desde una perspectiva de Reconocimiento Estructural de Formas basada en Modelos de Markov [Rabiner, 88]. Estos modelos son utilizados a todos los niveles, desde el simple Reconocimiento de Palabras Aisladas [Rabiner, 88], hasta el Reconocimiento al nivel Acústico-Fonético [Bahl, 88], [Schwartz, 88], o la Modelización del Lenguaje en los niveles Sintáctico-Semántico [Jelinek,85], [Lee, 88]. Resultados notables han sido obtenidos con esta metodología en tareas tales como el Reconocimiento de Grandes Vocabularios (Palabras Aisladas) [Laface, 88], o Reconocimiento de Habla Continua, con vocabularios de cierta envergadura [Lee, 88], [Ney, 88], [Schwartz, 88]. Estos éxitos están claramente sustentados por la capacidad de Aprendizaje "Inductivo" (es decir, a partir de muestras) que confiere a los Modelos de Markov el algoritmo de Baum-Welch de estimación de probabilidades. No obstante, la modelización Markoviana en RAH siempre conlleva implícitamente una cierta carga "Deductiva", ya que el esqueleto estructural subyacente debe ser establecido (deductivamente) en base a la experiencia del diseñador del sistema. En cualquier caso, los éxitos alcanzados por los Modelos de Markov han hecho decrecer notablemente el interés por otras aproximaciones (puramente deductivas) en las que los sistemas de RAH se desarrollaban de una forma más "manual", en base a conocimientos humanos sobre el habla (Sistemas Expertos o Basados en el Conocimiento), y han estimulado, en cambio, la experimentación con nuevos paradigmas de aprendizaje (puramente inductivo) como los Sistemas Conexionistas o Redes Neuronales [Peeling, 88].

Aunque el grupo ejecutor del presente proyecto no ha abandonado completamente la aproximación Deductiva [Benedí, 91] el proyecto aquí presentado se inscribe plenamente en la corriente inductiva. Sin embargo, siendo conscientes de las dificultades computacionales que puede entrañar una aproximación puramente inductiva (esta conduce, en general, a problemas intratables), hemos optado por una vía de "Aprendizaje Estructural", en la que se puede llegar a compromisos deductivo-inductivo adecuados. Los fundamentos metodológicos de esta vía los hemos buscado en el Reconocimiento Sintáctico o Estructural de Formas y su paradigma de Aprendizaje, la Inferencia Gramatical [Fu, 82]. Según esta vía, el compromiso estructural-versus-numérico en la descripción de los objetos considerados se desplaza hacia una mayor riqueza estructural, y el compromiso deductivo-inductivo en la Adquisición del Conocimiento, en la dirección Inductiva. Aproximaciones en cierto modo similares a la nuestra (aunque

metodológicamente diferentes), están empezando a ser consideradas también por otros autores [English,86], [Kohonen, 86], [Thomason, 86], [Bahl, 88], los cuales enfatizan, con nosotros, el gran potencial que parece existir en el Aprendizaje Automático de Modelos Estructurales.

Aunque el proyecto que se propone presenta ciertos aspectos originales en cuanto a metodología se refiere, sus objetivos concretos (ver siguiente sección) son similares a los planteados en otros muchos proyectos.

A nivel nacional, no obstante, los proyectos existentes son relativamente disjuntos del aquí propuesto. Como más próximo podríamos citar los trabajos del grupo de la ETSIT de Barcelona [Mariño,88], [Mariño, 90], en los que se apunta a la construcción de un sistema de reconocimiento de palabra continua basado en modelización Markoviana de unidades subléxicas (semisílabas) y un reconocimiento guiado por la sintaxis (rígida) de la tarea considerada.

A nivel internacional, los objetivos del presente proyecto apuntan en la misma dirección que la mayoría de proyectos Europeos y Estadounidenses [Lee, 88], [Ney, 88], [Schwartz, 88], aunque por el momento consideramos poco realista el aspirar a metas del alcance de las planteadas en [Lee, 88] o [Schwartz, 88], por ejemplo, y debemos pues contentarnos con metas no tan ambiciosas, aunque sí alcanzables en el entorno científico-técnico disponible.

3. Objetivos

Los resultados obtenidos en nuestras investigaciones precedentes en RAH, así como los resultados obtenidos por otros autores con el uso de métodos alternativos de aprendizaje y/o estimación, sugieren claramente la conveniencia de proseguir el desarrollo y aplicación extensiva de Métodos de Aprendizaje Automático de Modelos Estructurales. No obstante, la extensión de dichos métodos a todos los niveles lingüísticos tropieza con dos dificultades, no necesariamente independientes, que pueden exigir considerables desarrollos fundamentales antes de que la extensión pueda hacerse efectiva:

En primer lugar, estas técnicas de Aprendizaje, en su estado actual, dependen de la disponibilidad de (gran cantidad de) muestras de aprendizaje de los objetos (acústicos) a reconocer. Estas muestras pueden obtenerse fácilmente cuando los objetos son palabras aisladas, pero no están directamente accesibles cuando los objetos a reconocer son unidades subléxicas (fonemas, por ej.), o "unidades semánticas" (significados parciales), contenidas en frases de palabra continua. Una extensión más o menos trivial de las técnicas de Aprendizaje disponibles, requeriría de (gran número de) frases de aprendizaje segmentadas y etiquetadas manualmente en términos de las unidades consideradas. La dificultad y coste de esta delicada labor de segmentación y etiquetado, sugieren fuertemente la necesidad de ulteriores desarrollos (fundamentales) de las técnicas disponibles, con objeto de eliminar o reducir al máximo este requisito.

En segundo lugar, Las técnicas de Aprendizaje indicadas han sido todas ellas desarrolladas bajo el paradigma de Clasificación, en el que se aprende un modelo (red de estados finitos o Gramática Regular) por cada clase. Obviamente este paradigma es insostenible cuando la tarea de RAH abordada corresponde a lenguajes con alta complejidad semántica (¡el número de clases podría ser infinito!), lo que exige importantes cambios en el enfoque del problema de Aprendizaje subyacente. Un nuevo enfoque podría consistir en cambiar el paradigma de Clasificación por el de Interpretación, y considerar los modelos a aprender como Transductores de un lenguaje (acústico-fonético) de entrada a un lenguaje semántico de salida.

Por otra parte, la conveniencia de incrementar el uso de técnicas de Aprendizaje Automático (de Modelos Estructurales) en el Nivel Acústico-Fonético, viene indicada por la

necesidad de mejorar las prestaciones de este nivel cuyo funcionamiento es siempre especialmente crítico, a causa de su carácter de interfaz. Los resultados notables alcanzados por sistemas de Decodificación Acústico Fonéticos (DAF) basados en modelos de Markov, cuyas redes de estados finitos subyacentes se especifican ("manualmente") a priori, hacen pensar que estos resultados podrían superarse significativamente si las redes representaran realmente la estructura concreta de cada objeto acústico (fonema) considerado. En este sentido, podrían ser de utilidad algunos de los métodos de Aprendizaje desarrollados previamente por nosotros. No obstante, dada la similitud existente entre ciertos (grupos de) fonemas, parece indicado el uso de modelos estructurales discriminativos, lo que exige adecuar dichos métodos al aprendizaje con muestras positivas y negativas (ejemplos y contraejemplos).

Por último, la aplicación de técnicas de Aprendizaje Automático en el Nivel Sintáctico-Semántico constituye un planteamiento nuevo que no parece haber sido considerada explícitamente en la literatura hasta el presente. Sin embargo la posibilidad de obtener Modelos (Estructurales) de Lenguaje adecuados, de forma inductiva (a partir de muestras), despierta un interés considerable, ya que ello podría flexibilizar el diseño y uso de reconocedores de palabra continua, en el sentido de que sería el propio usuario el que especificaría la sintaxis del lenguaje a utilizar en la tarea de reconocimiento considerada, mediante la simple pronunciación de frases representativas del lenguaje propio de dicha tarea. Esta flexibilización tendría además el atractivo de que ya no se impondría al usuario sintaxis rígida alguna, sino que ésta podría ir enriqueciéndose progresivamente con las sucesivas frases por él utilizadas. Obviamente, la conveniencia o no de esta aproximación a la especificación y construcción automática de Modelos de Lenguaje, dependerá de la efectividad de los modelos obtenidos para reducir, en el nivel semántico, los errores procedentes de niveles inferiores, así como en su capacidad de recuperación cuando una nueva frase pronunciada no está exactamente contemplada sintácticamente por el modelo correspondiente.

Todas las anteriores consideraciones configuran los objetivos concretos del proyecto, los cuales pueden agruparse en las siguientes tres categorías:

(I) Objetivos tecnológicos.

1 Desarrollo de sistemas de Decodificación Acústico Fonética (DAF), basados principalmente en metodologías de Aprendizaje Automático, con tasas de error para la lengua Castellana no superiores al 20% (Dependiente del locutor) o 30% (Multilocutor).

2 Desarrollo de sistemas de Aprendizaje capaces de obtener automáticamente Modelos Estructurales de Lenguaje, a partir de frases (pronunciadas) representativas del Lenguaje (hablado) correspondiente a la tarea considerada. Estos sistemas trabajarán directamente a partir de representaciones simbólicas ("microfonéticas") directas de la señal vocal, o bien a partir de cadenas fonéticas suministradas por el sistema de DAF, y deberán alcanzar tasas de

reconocimiento semántico superiores al 95%, en tareas con complejidad semántica media (similar, por ejemplo, a la del lenguaje de los números Castellanos del CERO al MIL).

(II) Objetivos científicos.

1 Extender los resultados conocidos sobre Inferencia Gramatical para establecer un marco teórico adecuado al Aprendizaje Automático de Transductores Racionales. Este marco deberá determinar las limitaciones computacionales inherentes al problema propuesto.

2 Extender los resultados fundamentales clásicos, y/o establecer nuevos resultados, en Inferencia de Lenguaje Regulares con muestras positivas y negativas, así como en Inferencia de

Lenguajes Incontextuales.

En ambos casos, los resultados teóricos deberán servir de base y ser de utilidad para el desarrollo de (algunas de) las técnicas y/o metodologías requeridas par alcanzar los objetivos tecnológicos (I).

(III) Objetivos técnicos y auxiliares.

1 Adquisición de varios corpora de gran talla con datos vocales multilocutor correspondientes a varias tareas vocales de palabra continua, y segmentación manual de una pequeña parte de los mismos. Estos datos incluirán no menos de 100 locutores y deberán ser convenientemente divisibles en conjuntos independientes de entrenamiento y de test, adecuados para el aprendizaje y evaluación de los sistemas desarrollados.

2 Implementación de prototipo(s) adecuado(s) para demostrar el alcance de los objetivos tecnológicos (I).

4. Metodologías

El planteamiento general del proyecto, así como los objetivos concretos especificados en la sección anterior, dependen fuertemente de la posibilidad de aplicación de métodos más o menos generales de Aprendizaje Automático. Por otra parte, muchos de estos métodos, y otros procedimientos también requeridos para el correcto desarrollo del proyecto, descansan a su vez en el uso de ciertas técnicas algorítmicas básicas. En este sentido, la mayor parte de las metodologías "clásicas" actualmente disponibles son bien conocidas por el equipo investigador (el cual ha contribuido, de hecho, al desarrollo de algunas de ellas), gracias a sus anteriores trabajos en RAH y su reciente incursión en temas de Aprendizaje Automático:

(a) **Técnicas Algorítmicas Básicas**: *Programación Dinámica* y su utilización en Alineamiento Temporal No Lineal (DTW), Algoritmo de Viterbi, Búsqueda en Haz ("Beam Search"), "Backward-Forward", etc.; *Descenso por Gradiente* o "*Hill-Climbing*" y otras estrategias voraces, y su uso en técnicas de Reestimación ("bootstrapping"); *Ramificación y Poda* ("Branch & Bound") y su aplicación al desarrollo de algoritmos rápidos de búsqueda en Espacios Métricos.

(b) **Aprendizaje de pesos** en *Funciones Discriminantes Lineales* (Algoritmo "Perceptrón" básico, etc.)

(c) **Estimación de probabilidades** en *Gramáticas Regulares* y *Modelos de Markov* (Algoritmo de Baum-Welch).

(d) **Técnicas de Aprendizaje no supervisado**: Agrupamiento o "clustering"; *Cuantificación Vectorial*, etc.

(e) **Nuevas metodologías de Aprendizaje**: Aprendizaje de pesos en Funciones Discriminantes Lineales Multietapa (*Perceptrón Multicapa*, y "Backward Error Propagation"); aprendizaje de pesos en *Modelos Estructurales Discriminantes* generales (Funciones Discriminantes Lineales Extendidas) y su particularización a Gramáticas Discriminantes, y métodos alternativos de estimación de probabilidades en Modelos de Markov; ciertos métodos de *Inferencia Gramatical* y su aplicación al aprendizaje de la estructura en modelos estructurales.

Aunque algunas de las técnicas indicadas están ya suficiente maduras y son más menos

directamente aplicables a los problemas a resolver, la mayoría de ellas deben considerarse todavía en fase de desarrollo y/o adaptación. En particular, *Inferencia Gramatical* es un tema que requiere aún de trabajos científicos de tipo fundamental para establecer resultados teóricos en los que basar las técnicas correspondientes. Algunas de estas técnicas, como la "Error Correcting Grammatical Inference" (ECGI) [Rulot, 87], y la "Morphic Generator Grammatical Inference" (MGGI) [García, 87], y/o su aplicación al RAH, han sido desarrolladas e introducidas en la comunidad científico-técnica gracias a los trabajos de nuestro equipo investigador, y es objeto del presente proyecto continuar estos trabajos. Así mismo, miembros del equipo han introducido recientemente el concepto de *Funciones Discriminantes Lineales Extendidas*, y su aplicación a ciertos problemas de Reconocimiento de Formas y RAH. Este formalismo podría contemplar como casos particulares ciertas variantes al algoritmo de estimación de probabilidades en Modelos de Markov, recientemente propuestas por otros autores, y en las que se se tienen puestas grandes expectativas en lo que a su aplicación al RAH se refiere ("Corrective Training", "Maximal Mutual Information Estimation" (MMIE), etc.). Finalmente, aunque el *Perceptrón Multicapa* y su correspondiente algoritmo de aprendizaje han sido aplicados con éxito a ciertos problemas simples de RAH, es obvio que esta metodología deberá sufrir aún considerable desarrollo y/o adaptación al tipo de problemas (de RAH) a resolver antes de que pueda considerarse suficientemente madura para su aplicación a problemas de interés en RAH.

A continuación se discute la aplicación de las metodologías y técnicas que se acaban de mencionar a los objetivos tecnológicos indicados en la sección anterior.

Por una parte, el desarrollo de un sistema de *Decodificación Acústico Fonética* eficaz exigirá la puesta en juego de la práctica totalidad de las metodologías y técnicas mencionadas. En particular se prevé realizar un estudio comparativo de tres aproximaciones diferentes: (A) Uso de prototipos ("plantillas") de segmentos fonéticos establecidos mediante algoritmos de Descenso Rápido, Ramificación y Poda, y/o Programación Dinámica, para minimizar cierta función objetivo relacionada con la Distorsión acústico-fonética global del conjunto de frases de aprendizaje. (B) Uso de Modelos de Markov obtenidos mediante técnicas más o menos convencionales. (C) Uso de modelos estructurales de Estados Finitos obtenidos mediante nuestros algoritmos de aprendizaje por Inferencia Gramatical. (D) Combinación de (B) y (C). En todos los casos, excepto posiblemente en (A), se procederá por Reestimación o "Bootstrapping" basado en una (pequeña) fracción del corpus de aprendizaje convenientemente segmentado manualmente. Se asume que los modelos a obtener representarán "fonemas en contexto"; es decir, se deberá capturar toda la variabilidad alofónica exhibida por cada fonema en el corpus de aprendizaje utilizado. Así mismo, en todas las aproximaciones se considerará un "postproceso fonológico" para corregir errores de transcripción que violen las reglas fonológicas de la lengua (y/o tarea específica) considerada. Este postproceso se basará en Modelos Fonológicos estructurales obtenidos mediante nuestras técnicas de Inferencia Gramatical. Por otra parte se estudiará el uso de muestras negativas en el caso (C), así como la posibilidad y viabilidad de la aplicación de Funciones Discriminantes Multietapa (Perceptrón Multicapa) en algunos de los temas propuestos.

Por otra parte, el desarrollo de sistemas capaces de obtener *Modelos Estructurales de Lenguaje* (sintáctico-semánticos) de forma más o menos inductiva (a partir de ejemplos), es un tema que (salvo algunas raras excepciones) no ha sido explícitamente considerado en la literatura hasta el presente. Así pues, no es posible extender o modificar técnicas precedentes, y el desarrollo deberá ser esencialmente original. Las metodologías previstas derivan todas del área de Inferencia Gramatical. En este caso será un requisito importante el disponer de algunos de los resultados teóricos mencionados anteriormente, en especial en el caso de Aprendizaje Automático de Transductores. Las metodologías a desarrollar (o algunas ya disponibles) en Inferencia de Gramáticas Incontextuales, por otra parte, se aplicarán para estudiar comparativamente el alcance de los (robustos y eficientes) Modelos Sintácticos Regulares de Lenguajes, frente a modelos sintácticos más potentes (Incontextuales), aunque

computacionalmente menos interesantes y posiblemente aprendidos de forma menos robusta.

5. Estado actual

Transcurrido el primer año del proyecto, la mayor parte de los objetivos que se presentaban en el apartado 3 se encuentran en la fase de desarrollo prevista. A continuación se pasa a desglosar el estado actual de cada uno de ellos:

(I) OBJETIVOS TECNOLÓGICOS.

1 Los trabajos que en la actualidad se están desarrollando en el marco de la **Decodificación Acústico-Fonética** se pueden dividir en cuatro grupos, cada uno asociado a una modelización concreta. Todos estos grupos comparten una metodología bien conocida consistente en segmentar iterativamente las muestras de entrenamiento a partir de los modelos obtenidos hasta ese momento. Con los nuevos segmentos se aprenden nuevos modelos que se utilizan en el paso siguiente de la iteración. No obstante, para cada tipo de modelización se han tenido que desarrollar las técnicas que le son propias para realizar la segmentación y el aprendizaje.

El *primer grupo* de trabajos trata de la modelización de las unidades subléxicas mediante plantillas, utilizando algoritmos clásicos de alineamiento temporal no lineal. En un principio, cada unidad quedaba representada por una plantilla; posteriormente, se permitían varias gracias a la introducción de algoritmos de agrupamiento (clustering) posterior a la fase de segmentación de las muestras de entrenamiento. Los resultados alcanzados, el 61% para la tarea de las Hierbas (209 hierbas aromáticas y medicinales, monolocutor), y del 52% para un subconjunto (50 frases x 4 locutores, dependiente del locutor) de la tarea de las Frases fonéticamente balanceadas, han dado lugar a dos publicaciones en congresos internacionales [Aibar, 90] [Castro, 90]. Actualmente, se están realizando experimentos de decodificación independientes del locutor y habla continua.

En el *segundo grupo* de trabajos se trata de modelizar las unidades subléxicas mediante autómatas finitos estocásticos. Para ello se están utilizando las técnicas de inferencia gramatical (ECGI y MGGI) citadas en el apartado anterior. En una primera fase, las unidades subléxicas representadas son independientes del contexto, y las representaciones, discreta para la primera técnica, y semicontinua para la segunda. Los resultados obtenidos en tareas de palabras aisladas monolocutor (Hierbas, 75% MGGI y 65% ECGI) y de habla continua dependiente del locutor (subconjunto de las Frases fonéticamente balanceadas, con el 66% para el MGGI y el 53 % para el ECGI) han demostrado lo adecuado de estas técnicas en DAF. En la actualidad se están realizando experimentos con habla continua independiente del locutor, y en fases posteriores se utilizarán unidades subléxicas contextuales. Los trabajos realizados y los resultados alcanzados hasta este momento han sido publicados en diversas actas de congresos internacionales [Sanchís, 90] [Sanchís, 91], [Galiano, 91a] y en documentos técnicos [Carpi, 90], [Galiano, 91b].

Los trabajos del *tercer grupo* acaban de empezar, y en ellos se van a utilizar redes neuronales artificiales para la representación de las unidades subléxicas.

Finalmente, en el *último grupo* de trabajos se utilizan técnicas ya consideradas clásicas como son los Modelos de Markov Ocultos. Algunos de estos trabajos han sido orientados al estudio de la modelización de la duración [Russell,90] y su posterior incorporación en autómatas finitos [Ferrer, 91]. Otros, en cambio, están sirviendo para el desarrollo de sistemas de referencia [Más, 89], [Sánchez, 91] con el objetivo de poder comparar los resultados que se están obteniendo con los métodos propuestos en este trabajo con los que se obtengan mediante

técnicas clásicas.

2 En el apartado de **modelización de lenguaje** mediante técnicas de aprendizaje automático se está trabajando en tareas de comprensión de Discurso Continuo en las que, si bien se permite una sintaxis flexible, la semántica está restringida; este es el caso de aplicaciones del tipo consulta a una base de datos, modelización del lenguaje de acceso a una centralita telefónica, etc.

Esta línea de trabajo, que nosotros llamamos Modelización Estructural y Automática de Lenguajes [Prieto,91], se caracteriza por realizar el aprendizaje automático de los modelos de lenguaje, mantener cierta flexibilidad en cuanto a las construcciones sintácticas admitidas, modelizar de forma adecuada las restricciones del lenguaje (semántico) y posibilitar así mismo el aprendizaje adaptativo de la semántica de la tarea.

El proceso de comprensión del discurso continuo se plantea como un proceso de traducción, a partir de secuencias acústicas que simbolizan señales vocales, en alguna representación adecuada de las acciones o "mensajes semánticos" contenidos en dichas señales. El problema es, desde este punto de vista, el aprendizaje automático de un transductor a partir de un conjunto de pares de entrada-salida que ejemplifican su comportamiento [Vidal,89] [Prieto,91].

Recientemente se ha aplicado esta aproximación a la tarea de Reconocimiento de los números en castellano comprendidos entre el cero y el novecientos noventa y nueve, obteniéndose tasas de acierto a nivel semántico del 95% [Prieto,91] y del 98% [Segarra,91]. Actualmente se está trabajando en otras tareas más complejas como son el reconocimiento de los números del cero al millón y la interpretación de frases de consulta a una base de datos con información sobre Geografía Española.

(II) OBJETIVOS CIENTÍFICOS.

Como se ha puesto de manifiesto, el núcleo de nuestra aproximación al problema del RAH está constituido por la adopción de un punto de vista inductivo en la construcción de los sistemas de reconocimiento. Esta perspectiva ha supuesto una continuidad con relación a trabajos precedentes en nuestro grupo, en cuyo seno se habían desarrollado varios algoritmos de aprendizaje de Lenguajes Regulares [Rulot, 87], [García, 87], [García, 90a], los cuales han sido utilizados con éxito en diferentes tareas de reconocimiento de palabras aisladas [Vidal, 88], [García, 90b]. Sin embargo, una característica común a todos estos algoritmos y, en general, a los métodos de aprendizaje disponibles a través de la literatura de Inferencia Gramatical [Fu,75],[Angluin, 83], es el uso de sólo información positiva. El no tomar en consideración la información negativa conlleva, en inferencia de lenguajes, el peligro de que los modelos que se obtienen sean poco discriminativos. En problemas de clasificación (como, por ejemplo, tareas de reconocimiento de palabras aisladas) dicho peligro puede ser grave si las clases están muy próximas entre sí.

Por otra parte, sólo en casos muy particulares pueden abordarse los problemas de RAH a través del concepto de clasificación. En general, más que discriminar los objetos en clases, de lo que se trata es de interpretar los mensajes de habla asignándoles un significado mediante su decodificación en un lenguaje (semántico) preestablecido. Más que de un problema de clasificación se trata de un problema de transducción. Si la clasificación puede modelizarse adecuadamente a través del concepto de pertenencia a un lenguaje y el aprendizaje consiste en obtener una descripción (gramática, autómata) del lenguaje a partir de ejemplos, en el caso general el aprendizaje debe conducir a la obtención de un dispositivo capaz de realizar la interpretación, esto es, un transductor. Los conceptos de Transducción Racional y de

Transductor Finito permiten modelizar tales tareas complejas de RAH.

En la consecución de los objetivos señalados los resultados obtenidos hasta el presente pueden resumirse como sigue:

1. En **aprendizaje de lenguajes regulares a partir de datos positivos y negativos** se ha desarrollado un algoritmo capaz de identificar en el límite cualquier lenguaje regular. Ello significa que si se suministran, de manera incesante, datos de un lenguaje a la entrada del algoritmo, éste converge a un autómata que acepta exactamente el lenguaje desconocido. Además, con cada entrada de un nuevo dato el algoritmo produce una nueva hipótesis en tiempo polinómico con la suma de las longitudes de los datos recibidos hasta el momento [Oncina, 90]. Una aplicación de estas ideas al RAH se presenta en [Segarra, 90]

2. En cuanto al segundo aspecto mencionado, el del **aprendizaje de transducciones**, hasta el presente sólo se había estudiado un caso muy limitado de transducciones racionales: las realizadas por máquinas de Mealy [Gold, 78], [Veelenturf, 78], [Luneou, 84]. Una familia interesante de funciones racionales que incluye propiamente a las transducciones secuenciales es la constituida por las Transducciones Subsecuenciales [Berstel, 79]. Como resultado del trabajo planteado en el proyecto, se ha desarrollado un algoritmo polinómico que identifica cualquier transducción subsecuencial en el límite [Oncina, 91a]. Entre sus características más relevantes desde el punto de vista de sus potenciales aplicaciones en tareas de comprensión del habla destaca el hecho de que no precisa de una previa segmentación de los pares de entrada-salida de entrenamiento [Oncina, 91b]. La mayor debilidad de la aproximación al aprendizaje de Modelos de lenguaje (ver apartado 3) desde la óptica de las transducciones reside en la necesidad de disponer de pares de entrenamiento segmentados, lo cual constituye un problema no bien resuelto para el que los métodos aquí indicados proporcionan una solución definitiva.

Con todo, al nivel actual de desarrollo, tales algoritmos tienen, sobre todo, un interés teórico. Su utilización en problemas reales en los que el ruido y la escasez de datos son características siempre presentes requiere de ulteriores trabajos tendentes a modelizar adecuadamente tales restricciones.

(III) Objetivos técnicos y auxiliares.

Para cumplir los objetivos, tanto técnicos como tecnológicos, se han diseñado las siguientes tareas de palabra continua multilocutor: Frases fonéticamente balanceadas y consulta a una Base de Datos Geográfica. Como valoración de los métodos de Modelización de Lenguajes se propone una tarea previa de palabra continua monolocutor: Números Castellanos del CERO al MILLÓN. Y por último, se han considerado también diversas tareas de apoyo con aplicaciones de palabras aisladas multilocutor: Dígitos Castellanos, Letras del Castellano y Ciudades Españolas (= 250 ciudades).

Hasta la fecha se ha adquirido el siguiente material vocal: Frases fonéticamente balanceadas: 10 locutores x 1 repetición x 170 frases (≈ 70 minutos), existen 7 frases de 4 locutores segmentadas manualmente; Números del CERO al MILLÓN: 1 locutor x 1 repetición x 853 números (≈ 30 minutos), existe un subconjunto de 238 números segmentados manualmente; Dígitos Castellanos: 10 locutores x 10 repeticiones x 10 dígitos (≈ 10 minutos); y Letras del Castellano: 10 locutores x 10 repeticiones x 30 letras (≈ 15 minutos).

6. Referencias.

[Aibar, 90] P. Aibar, M.J. Castro, F. Casacuberta, E. Vidal. "Multiple Template Modelling of Sublexical Units", en **Speech Recognition and Understanding: Recent Advances, Trends and Applications** P. Laface, Ed. Springer-Verlag. NATO ASI Series, 1990.

[Angluin, 83] D. Angluin, C. H. Smith. "Inductive inference: theory and methods". **Computing Surveys**, 15 (3), pp. 237-269, 1983.

[Bahl, 88] L.R. Bahl, et al. "Acoustic Markov Models used in Tangora Speech Recognition System". **ICASSP 88**, pp.497-500, 1988.

[Benedí, 91] J.M. Benedí, et al. "Proyecto ROARS: Robust Analytical Speech Recognition System", 1991. En este boletín.

[Berstel, 79] J. Berstel. **Transductions and context-free languages**. Teubner. Stuttgart, 1979.

[Carpi, 90] S. Carpi. **Aprendizaje de Modelos Fonéticos para el Reconocimiento Automático del Habla**. Proyecto Fin de Carrera. Facultad de Informática, UPV, Valencia, 1990.

[Castro, 90] M.J. Castro, P. Aibar, F. Casacuberta, E. Vidal. "Automatic Selection of Sublexic Templates by Using Dynamic Time Warping Techniques", en **Signal Processing V: Theories and applications**, L. Torres, E. Masgrau, M.A.. Lagunas, Eds., pp. 1351-1354, 1990.

[English, 86] T. English, L. Bogges. "A grammatical approach to reducing the statistical sparsity of language models in natural domains". **ICASSP 86**, pp 742-749, 1986.

[Ferrer, 91] J.R. Ferrer. **Modelización de la duración en modelos estructurales para tareas de Reconocimiento Automático del Habla**. Proyecto Fin de Carrera. Facultad de Informática, UPV, Valencia, 1991. Pendiente de lectura.

[Fu, 75] K. S. Fu, T. L. Booth. "Grammatical Inference: Introduction and Survey, parts 1 and 2". **IEEE Trans. Sys. Man and Cyber.**, SMC-5, pp. 95-111, pp. 409-423, 1975.

[Fu, 82] K.S. Fu. **Syntactic Pattern Recognition and Applications**. Prentice-Hall, New York, 1982.

[Galiano, 91a] I. Galiano, F. Casacuberta, E. Sanchís. "On the Structure of Subword Units for a Speaker Independent Continuous Speech Task". **EUROSPEECH 91**, Genova, 1991. Pendiente de publicación.

[Galiano, 91b] I. Galiano, K. Zünkler. **A Study Comparing Structural Sub-Word Models for ASR.** Internal report, Siemens AG, Corporate Research and Development, Munich, 1991.

[García, 87] P. García, E. Vidal, F. Casacuberta. "Local Languages, the Successor Method, and a step towards a general methodology for the Inference of regular Grammars". **IEEE Trans. Pattern Analysis and Machine Intelligence**, Vol. PAMI-9(6), pp.841-845,

1987

[García, 90a] P. García, E. Vidal. "Inference of k-Testable Languages in the Strict Sense and application to Syntactic Pattern recognition". **IEEE Trans. Pattern Analysis and Machine Intelligence**, Vol. PAMI-12(9), pp. 920-925, 1990.

[García, 90b] P. García, E. Segarra, E. Vidal, I. Galiano. "On the use of the Morphic Generator Grammatical Inference (MGGI) methodology in automatic speech recognition". **Int. Journal on Pattern Recognition and Artificial Intelligence**, Vol. 4, Nº4. 1990.

[Gold, 78] E. M. Gold. "Complexity of automaton identification from given data". **Information and Control**, 37: pp. 302-320, 1978.

[Jelinek, 85] F. Jelinek. "Self-Organized Language Modelling for Speech Recognition". No publicado, 1985.

[Kohonen, 86] T. Kohonen. "Dynamically expanding context, with application to the correction of symbol strings in the recognition of continuous speech". **ICPR Proc.** Parás 1986.

[Laface, 88] P. Laface. "Recognition of Words in Very Large Vocabulary", en **Recent Advances in Speech Understanding and Dialog Systems**, H. Niemann, M. Lang, G. Sagerer, Eds. Springer Verlag, pp 235-254, 1988.

[Lee, 88] K.F. Lee. **Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System**. Ph. D. Thesis, Carnegie- Mellon Univ. Abril 1988.

[Luneau, 84] P. Luneau, M. Richetin, C. Cayla. "Sequential Learning from Input-Output Behavior". **Robotica** 1, pp 151-159, 1984.

[Mariño, 88] J.B. Mariño, C. Nadeu, E. Lleida. "Finite state Grammar Inference for Connected Word Recognition", in **Signal Processing IV: Theories and Applications**. J.L. Lacoume, A. Chehikian, N. Martán, and J. Malbos Ed. Elsevier Sc. Pub. B.V. EURASIP, 1988.

[Mariño, 90] J.B. Mariño, et al. "Recognition of Numbers by Using Demisyllables and Hidden Markov Models", en **Signal Processing V: Theories and applications**, L. Torres, E. Masgrau, M.A.. Lagunas, Eds., pp. 1363-1366, 1990.

[Mas, 89] B. Mas. **Reconocimiento de palabras aisladas y Decodificación Acústico-Fonética utilizando Modelos de Markov**. Proyecto Fin de Carrera. Facultad de Informatica, UPV, Valencia, 1989.

[Ney, 88] H. Ney, D.Mergel, A. Noll, A. Paesler. "Overview of Speech Recognition in the SPICOS System" en **Recent Advances in Speech Understanding and Dialog Systems**, H. Niemann, M. Lang, G. Sagerer, Eds. Springer Verlag, pp 305-309, 1988.

[Oncina, 90] J. Oncina, P. García. **Un algoritmo de inferencia de lenguajes regulares usando datos positivos y negativos**. Informe de Investigación, DSIC-II/1/1990, Universidad Politécnica Valencia. Una versión resumida de este trabajo en **Proc. IV Simposium Nacional de Reconocimiento de Formas y Análisis de Imágenes,**

Granada, Septiembre 1990.

[Oncina, 91a] J. Oncina, P. García, E. Vidal. "Learning Subsequential Transducers for pattern recognition interpretation tasks". Enviado para publicación en IEEE-PAMI, 1991.

[Oncina, 91b] J. Oncina, P. García. **Aprendizaje de funciones subsecuenciales.** Informe de investigación, Universidad Politécnica de Valencia, DSIC II/5/91, 1991.

[Peeling, 88] S.M. Peeling, R.K. Moore, A.P. Varga. "Isolated Digit Recognition using the Multilayer Perceptron" en **Recent Advances in Speech Understanding and Dialog Systems**, H. Niemann, M. Lang, G. Sagerer, Eds. Springer Verlag, pp. 261-266, 1988.

[Prieto, 91] N. Prieto, E. Vidal. "Automatic Learning of Structural Language Models". **ICASSP 91**, 1991. Pendiente de publicación.

[Rabiner, 88] L.R. Rabiner. "Mathematical Foundations of Hidden Markov Models", en **Recent Advances in Speech Understanding and Dialog Systems**, H. Niemann, M. Lang, G. Sagerer, Eds. Springer Verlag, pp 183-206, 1988.

[Rulot, 87] H. Rulot, E. Vidal. "Modelling (sub)string-length-based constraints through a grammatical inference method", in **Pattern Recognition Theory and Applications**, Devijver and Kittler Eds., Springer-Verlag, pp 451-459, 1987.

[Russell, 90] M.J. Russell, et al. "The Arm Continuous Speech Recognition System". **ICASSP 90**, pp. 69-72, 1990.

[Sánchez, 91] A. Sánchez. **Decodificación Acústico-Fonética en el habla mediante HMM.** Proyecto Fin de Carrera. Facultad de Informática, UPV, Valencia, 1991.

[Sanchís, 90] E. Sanchís, F. Casacuberta, S. Carpi. "Learning Structural Models of Sublexical Units", en **Speech Recognition and Understanding: Recent Advances, Trends and Applications** P. Laface, Ed. Springer-Verlag. NATO ASI Series, 1990.

[Sanchís, 91] E. Sanchís, F. Casacuberta, I. Galiano, E. Segarra. "Learning Structural Models of Subword Units through Grammatical Inference Techniques". **ICASSP 91**, 1991.

[Schwartz, 88] R.M. Schwartz et al. "Acoustic-Phonetic Decoding of Speech", en **Recent Advances in Speech Understanding and Dialog Systems**, H. Niemann, M. Lang, G. Sagerer, Eds. Springer Verlag, pp 25-50, 1988.

[Segarra, 90] E. Segarra, P. García, J.M. Oncina, A. Suárez. "On the Use of Negative Samples in the MGGI Methodology and its Applications for difficult Vocabulary Recognition Tasks", en **Speech Recognition and Understanding: Recent Advances, Trends and Applications** P. Laface, Ed. Springer-Verlag. NATO ASI Series, 1990.

[Segarra, 91] E. Segarra, P. García. "Automatic Learning of Acoustic and Syntactic-Semantic levels in Continuous Speech Understanding". **EUROSPEECH 91**, Genova, Septiembre 1991. Pendiente de publicación.

[Thomason, 86] M.G. Thomason, E. Granum, R.E. Blake. "Experiments in Dynamic Programming Inference of Markov Networks with strings representing speech data". **Pattern**

Recognition, vol.19, No.5, pp 343-351, 1986.

[Veelenturf, 78] L. P. J. Veelenturf. "Inference of sequential Machines from sample Computation". **IEEE Trans. in Computers**. 27, pp 167-170, 1978.

[Vidal, 89] E. Vidal, P. García, E. Segarra. "Inductive learning of Finite-State Transducers for the Interpretation of Unidimensional Objects", en **Structural Pattern Analysis**, R. Mohr et al Ed. World Scientific Publ., pp 17-36, 1989.

[Vidal,88] E. Vidal, N. Prieto, E. Sanchís, H. Rulot. "Application of the Error Correcting Grammatical Inference Method (ECGI) to Multi-speaker isolated word recognition",en **Recent Advances in Speech Understanding and Dialog Systems**, H. Niemann, M. Lang, G. Sagerer, Eds. Springer Verlag, 1988.