Tecnología del Habla para Siete Idiomas: El Proyecto Esprit POLYGLOT-I

J.M.Pardo; E.Enríquez; S.Aguilera; A. Santos; A.Quilis (1)

1. Planteamientos generales y breve historia del proyecto

El proyecto POLYGLOT-I, de tres años de duración, (agosto, 1989-agosto, 1992), está centrado en la tecnología del habla y sus objetivos son tanto el reconocimiento de habla (habla-texto) como la síntesis del lenguaje (texto-habla).(2)

Como es sabido, dentro del área del reconocimiento automático del habla, se pueden distinguir dos tecnologías diferentes: el reconocimiento de palabras aisladas y el reconocimiento de habla continua. Ambos aspectos se consideran dentro de este proyecto, por lo que, en definitiva, los intereses de este Proyecto giran en torno a tres objetivos: el reconocimiento de palabras aisladas (RPA), el reconocimiento de habla continua (RHC) y la conversión texto-habla (CTH) o síntesis del lenguaje.

Pero aparte de su interés por tratar cuestiones punteras en tecnología del habla, POLYGLOT-I aporta un nuevo e importante aspecto, tanto desde la perspectiva de la investigación como de sus posibles aplicaciones comerciales y es el que propone unos sistemas plurilingües, es decir válidos para diferentes idiomas. En concreto, para siete idiomas comunitarios: alemán, español, francés, griego, holandés, inglés e italiano.

Polyglot-I se apoya en una experiencia previa, obtenida a partir de un proyecto Esprit anterior (3) donde se analizaron detalladamente y conjuntamente algunos aspectos lingüísticos, se desarrollaron una serie de herramientas para el tratamiento común de determinadas características lingüísticas (fonéticas, léxicas y sintácticas) y se generaron bases de datos homogéneas y semejantes en los siete idiomas implicados.

Participan actualmente en POLYGLOT-I siete países: Alemania, España, Francia, Grecia, Holanda, Inglaterra e Italia, y las Universidades y Empresas que forman el consorcio son:
- En Alemania:
- La Universidad de Bochum (Lehrstuhl fur allg. Elektrotechnik und Akustik)
- Philips (Aachen)
- Siemens
-En España:

-.La Universidad Politécnica de Madrid (E.T.S.I. de Telecomunicaciones, Dpto. de Ingeniería Electrónica.
- La Universidad Nacional de Educación a Distancia (Facultad de Filología, Dpto. de Lengua Española).
- En Francia:
- El C.N.R.S./LIMSI (Departement Communication Homme-Machine)
- BULL-S.A.
- En Grecia:
- La Universidad de Patras (Department of Electrical Engineering)
- En Holanda:
- La Universidad Católica de Nimega (Department of Linguistic Speech and Language Technology).
- Nederlandse Philips Bedrijven B.V.(Einhoven) (Institute for Perception Research)
- En Inglaterra:
- La Universidad de Edimburgo (Centre for Speech Technology Research -I.P.O.)
- En Italia:
- Olivetti Systems and Networks (responsable general del proyecto)
- Syntax Software Sistemi Sp.A.
- CRAI

2. Objetivos y Estructura General del Proyecto

Como se ha dicho, el trabajo general de POLYGLOT-I gira en torno a tres temas: RPA, RHC y CTH, sin embargo, en cada una de estas áreas se pretenden resultados específicos que conviene destacar, ya que determinan la estructura general del proyecto:

1- En RPA se pretende desarrollar un sistema de reconocimiento de palabras aisladas basado en un amplio número de formas léxicas (decenas de miles), adaptable al locutor, capaz de funcionar en las lenguas implicadas y ejecutable en un ordenador personal.

2- En CTH también se va a desarrollar un sistema de síntesis de habla (sistema de texto-habla), de alta calidad, para vocabularios ilimitados, para varias lenguas europeas y ejecutable en un ordenador personal.

3- Tanto para RPA como para CTH, se harán prototipos que demuestren la viabilidad del producto y su interés en el campo de las aplicaciones comerciales.

4- Finalmente, en RHC se realizará el estudio pormenorizado de las posibilidades de desarrollar un sistema semejante a los anteriores, que sería desarrollado posteriormente.

Pero, aparte de estos cuatro objetivos primarios se está llevando a cabo un considerable esfuerzo para integrar toda la información lingüística anterior (Esprit 860) en los nuevos sistemas de reconocimiento y síntesis.

El hecho de conseguir integrar toda esta información para diversas lenguas permitirá que el sistema obtenido pueda ampliarse más fácilmente a otros sistemas, con un coste mucho menor.

Así pues, estos cinco objetivos se recogen y desarrollan en los cinco grupos de trabajo que están funcionando en el proyecto y cuya labor especificamos con más detalle en los siguientes apartados. Hay que destacar, sin embargo, que en todos ellos se ha diferenciado entre las cuestiones independientes de la lengua de aquellas que son dependientes del idioma y que todos los participantes en el proyecto colaboran, en mayor o menor medida, en ambas.

3. Especificaciones y Desarrollo del Proyecto

3.1. Las Tareas Comunes de POLYGLOT-I

Uno de los criterios generales que se ha tenido siempre muy en cuenta es el de establecer una serie de estándares asumidos por todos los grupos participantes y para todas las lenguas. Estos estándares se refieren, fundamentalmente, a las bases de datos habladas y a su tratamiento informático (contenido, herramientas y procedimientos de grabación, y configuración hardware necesaria).(4) Además, siempre que ha sido posible se procuró también que nuestros estándares coincidieran con los establecidos por otro proyecto Esprit: el Proyecto SAM (5). Pasamos, pues, a continuación, a enumerar estos estándares.

Respecto a las grabaciones, para RPA se realizaron grabaciones de un solo canal, utilizando un micrófono RCF MD2700; las muestras de habla se digitalizaron, con una frecuencia de muestreo de 16kHz y un filtro anti-solapamiento específico (de Olivetti). Para RHC y CTH las grabaciones se han realizado con dos canales y una frecuencia de muestreo de 20 kHz (16 bits), utilizando la placa OROS-AU22, con el fin de conseguir una compatibilidad total con SAM. Uno de los canales de RHC y de CTH utiliza el micrófono Shure SM-10. El otro canal, en RHC, utiliza el micrófono de mesa PCC-160-W; en CTH, el segundo canal queda reservado para un laringograma. Todas las grabaciones para RPA y RHC se realizaron en habitaciones acondicionadas acústicamente o en oficinas muy silenciosas. Para CTH se determinó utilizar cámaras anecoicas.

Respecto a las bases de datos utilizadas para RPA, en cada una de las lenguas del proyecto se consideraron 4 conjuntos de palabras. El primer conjunto (A) está compuesto por 500 palabras leídas por 10 hablantes (5 hombres y 5 mujeres). Para la selección de estas 500 palabras se consideraron las 200 palabras más frecuentes de la lengua en cuestión y otras 300 palabras que representaran la distribución fonémica de esa lengua. Todas ellas deben estar etiquetadas fonémicamente. Este conjunto se utiliza para la construcción del sistema de reconocimiento. El segundo bloque (B) consiste en 100 palabras tomadas del grupo A pero leídas por 5 informantes diferentes (3 hombres y 2 mujeres) y sin etiquetar fonémicamente. Se utilizan para el proceso de entrenamiento de la dependencia del locutor. El tercer grupo (C) está formado por otras 500 palabras diferentes, con una representación adecuada de las secuencias fonémicas del idioma, leídas por los 10 hablantes del grupo A más los cinco del grupo B. No tiene etiqueta fonémica asignada y se utiliza para la comprobación del sistema de reconocimiento. Para el cuarto conjunto (D) se utilizó aquí un texto de 700 palabras,las cuales fueron leídas como formas aisladas (marcando una breve pausa entre una y otra y sin marcar la entonación oracional) por los mismos 5 informantes del grupo B. También se utiliza para la comprobación del sistema de reconocimiento.

Para RHC se consideraron dos bases de datos diferentes: la primera está constituida por 6 hablantes (tres hombres y tres mujeres), tomados de la Base de Datos DARPA. Es, pues, inglés americano y tienen etiqueta fonémica a nivel de frase. La segunda la constituyen textos, libres tomados de periódicos europeos de gran tirada. Se consideran enunciados oracionales completos, tomados de 6 hablantes (3 hombres y 3 mujeres) y un mínimo de 5000 palabras por cada hablante, todas ellas etiquetadas fonémicamente.

Por último, para CTH, las bases de datos constan de: una historieta breve, con una buena cobertura de las secuencias fonémicas y de la prosodia del idioma, leídas por dos hablantes (un hombre y una mujer) y una base de datos de segmentos, ambas bases de datos con segmentación fonémica.

Aunque el esfuerzo de este grupo de trabajo se ha centrado en las bases de datos de habla, también se ha dedicado cierta atención a la integración de la información lingüística obtenida en el proyecto 860 para la finalidad actual. Entre estas tareas señalaremos el trabajo realizado para la construcción de un diccionario, que ha integrado el ya generado en el proyecto anterior, pero en forma, ahora, de base de datos relacional, permitiendo un acceso directo y rápido a la información lingüística de cada forma léxica y, además, la generación de palabras derivadas (morfológicamente) y la creación de diccionarios secundarios para usos específicos. (6)

Dentro de las tareas comunes se contempla también la necesidad de desarrollar herramientas para manejo de bases de datos habladas y para etiquetado fonético automático de las mismas, partiendo,

para ello, del conocimiento de las series de fonemas que corresponden a cada palabra.

Así pues, se están desarrollando en el proyecto un sistema estándar de acceso a bases de datos de habla (fichero, locutor, etc.) basado en DBIII, un sistema de análisis estadístico de frecuencias basado en el lenguaje "S" y un sistema de autoetiquetado fonético de bases de datos habladas basado en modelos de Markov.

3.2. El Sistema de reconocimiento de Palabras Aisladas de POLYGLOT-I

Se parte de un sistema de reconocimiento de palabras aisladas, adaptable al locutor y de vocabulario extenso, ya existente y que desarrolló Olivetti para el italiano (v. Billi, 1989), ampliándose ahora a las restantes lenguas del proyecto. El sistema es adecuado para las aplicaciones que exigen el uso de vocabularios extensos o muy extensos (típicamente, aplicaciones de dictado). Entre las características más útiles podemos mencionar el rápido procedimiento, independiente de la aplicación, para entrenar el sistema a la voz del usuario.

El sistema tiene tres bloques: el primero es la etapa de Preselección; el segundo es el Análisis Fonético fino y el tercero corresponde al análisis contextual. La etapa de preselección consiste en la adquisición y digitalización de la señal. La señal de voz (o señal oral) se captura mediante un micrófono cardioide dinámico tipo RCF MD2700, muestreado a 16 kHz y digitalizado con una precisión de 12 bits. Los vectores de 20 coeficientes de autocorrelación y de LPC se calculan para cada trama de 10ms. Estos vectores se comparan con un conjunto de vectores fonéticos de unos prototipos, (previamente calculados y dependientes del locutor), usando una medida de distancia espectral WLR modificada. El número de prototipos varía según las necesidades específicas de cada lengua. (7) La salida de esta primera etapa es una lista de, entre 40 y 200 formas candidatas, entre las que debe encontrarse la palabra que sirvió de entrada.

La etapa de Análisis Fonético Fino genera, para cada una de las palabras candidatas obtenidas por preselección, una segmentación fonética que viene dada por la representación fonética de cada palabra contenida en el diccionario. La segmentación está asociada a una puntuación, que es una medida del grado de concordancia entre los parámetros relacionados con los fonemas y la estadística de su distribución. El proceso está basado en una verificación bayesiana multirasgo.

La etapa de análisis contextual se basa en las propiedades combinatorias del idioma, aplicadas a las clases grmaticales de las palabras. Las fuentes de conocimiento utilizadas son: frecuencia de clases de palabras, probabilidades de dos clases gramaticales (bigramas) y tres clases gramaticales (trigramas), tabla de transiciones de palabras función y reglas determinísticas que consideran

algunas características específicas del idioma que no quedan convenientemente expresadas mediante los bigramas.

Un algoritmo de búsqueda en haz de izquierda a derecha busca el mejor camino en la lista de palabras que produce el Análisis Fonético Fino. Las puntuaciones acústicas y las probabilidades combinatorias (de dos clases gramaticales) se combinan entre sí. Para los contextos izquierdo y derecho se utilizan de una a tres palabras dependiendo de cada caso.

3.3. Reconocimiento de Habla Continua en POLYGLOT-I

Las características del trabajo sobre reconocimiento de habla continua que se está desarrollando en POLYGLOT son como siguen: el reconocimiento será dependiente del locutor, en el modelo no se considerará la robustez frente al ruido y la longitud del vocabulario oscilará entre 1000 y 5000 palabras. Importa, además, su carácter plurilingüe y que el registro hablado mantenga un alto grado de naturalidad. Ya hemos destacado, sin embargo, que el trabajo que se realiza en RHC es, fundamentalmente una primera aproximación que permitirá establecer las líneas de trabajos posteriores.

Dentro de POLYGLOT, el RHC está basado en el uso de alófonos como unidades de decisión, en lugar de palabras, no sólo por el tamaño del léxico al que debe acceder sino también por conseguir el mayor grado de libertad posible respecto al vocabulario, en el proceso de modelado del habla. La investigación considerará el proceso completo, desde la parte acústica a la ortográfica, a nivel de oración. Para poder identificar un esquema acústico decodificado óptimo se consideran dos reconocedores fonéticos basados en concepciones teóricas diferentes. Las dos estrategias de decodicifación acústico-fonética que van a ser consideradas son:

1) Un modelo basado en los modelos ocultos de Markov (MOM)
2) Un método basado en Redes Neuronales, usando el "Time Delay Neural Networks"

Una serie de ensayos comparativos descubrirán los aspectos clave de la arquitectura par habla continua que permita la posterior extensión del proyecto. El propósito de estas pruebas no es valorar la calidad global de los diversos decodificadores fonéticos, sino comparar las estrategias. La medida de la calidad global del reconocimiento de habla usado en las comparaciones va a ser la precisión del reconocimiento de palabras y frases, teniendo en cuenta los efectos de las restricciones del lenguaje, es decir, la longitud del vocabulario y las restricciones sintácticas. La velocidad de reconocimiento fonético también se utilizará como procedimiento de evaluación, pero, principalmente, para obtener explicaciones acerca de los diferentes fenómenos que están siendo

estudiados. Con el fin de preparar resultados comparables para cada uno de los esquemas de decodificicación acústico-fonética en competición, se usará en todos ellos una misma etapa de alto nivel (es decir, procedimiento de búsqueda, modelo lingüístico).

Para limitar el número de parámetros en los experimentos de evaluación, los experimentos se centrarán en una única lengua: el inglés. A continuación la estrategia elegida será adaptada a otras dos lenguas (francés y alemán). Será empleada,para el entrenamiento de todos los decodificadores acústico-fonéticos una base de datos de habla común, e, igualmente, se utilizará una misma base de datos para las pruebas. Los diversos modelos serán evaluados con una base de datos de inglés británico y manejarán un amplio léxico (de unas 5000 palabras) y una sintaxis con pocas restricciones (bigramas entrenados sobre un gran conjunto de artículos de periódico).

También se llevarán a cabo experimentos sobre una base de datos de inglés americano (DARPA RESOURCE MANAGEMENT) que contiene oraciones con un amplio léxico (1000 palabras) pero con una sintaxis muy limitada (con una perplejidad en torno a 60), lo que permitirá la comparación de los diferentes modelos en una sintaxis muy restringida.

Dos procedimientos de búsqueda (léxica) se emplean en el trabajo. El primero se basa en una estrategia abajo-arriba, usando cadenas fonéticas. El segundo es un procedimiento de búsqueda integrada, sobre una red de símbolos de fonemas de 10ms.

3.4. El Sistema de Conversión Texto-Habla de POLYGLOT-I

Los sistemas convencionales de conversión texto-habla, disponibles en el mercado como productos comerciales se basan en una estructura modular, en la que cada módulo tiene como entrada una estructura de datos esencialmente lineal y cuya salida es también una estructura de datos similar, que sirve de entrada al próximo módulo. No hay realimentación de un módulo a su predecesor, y los sucesores no tienen acceso a la entrada de sus predecesores.

Trabajando con estos sistemas lineales, propios de los setenta, sabemos que pueden conseguirse salidas de habla de alta calidad, pero con ciertas limitaciones de calidad que son casi imposibles de superar. Una de estas dificultades aparece cuando se adapta un sistema ya existente a otro idioma. La forma en que se modulariza el complejo proceso de transformar el texto de entrada en habla audible puede ser óptimo para tratar las características de la lengua primaria, pero no tiene por qué ocurrir así para otras lenguas.

Parece que una modularización estrictamente lineal del proceso de transformación texto-habla no es

adecuada para tratar problemas tales como la prosodia (el punto más débil de los CTH comercialmente disponibles). Actualmente se considera que, para modelar adecuadamente la producción de palabras, es necesario un proceso donde todos los módulos tengan acceso a toda la información que sirve de entrada o que se genera en el sistema durante la producción de una emisión. En un sistema así, los módulos todavía operan uno tras otro, pero, debido a que operan sobre una estructura compleja de varios niveles, desde la cual cada módulo puede leer y escribir, el orden de las operaciones ya no está fijado por la propia arquitectura del sistema. También resulta más sencillo omitir un módulo que no es necesario para una lengua concreta, o añadir un nuevo módulo para otra lengua. POLYGLOT ha diseñado una de estas arquitecturas complejas para el CTH (v. Vittorelli et al. 1990). Además, siguiendo su filosofía general de desarrollar sistemas que sean plurilingües, se ha desarrollado un identificador de lenguas que permita acceder al módulo de conversión texto-habla adecuado sin necesidad de una indicación previa del idioma que se maneja.

3.5. Las Aplicaciones de los sistemas desarrollados en POLYGLOT-I

Puesto que una de las claves de la actual convocatoria del Programa Esprit es la de fomentar el desarrollo de aplicaciones concretas, existe, en POLYGLOT-I, un grupo de trabajo sobre aplicaciones cuyo objetivoes demostrar el interés comercial de los sistemas de RPA y CTH desarrollados en el proyecto. Como hemos venido señalando, en el entorno del proyecto parece esencial el desarrollo de los conceptos "plurilingüe" y "pluridominio". El primero es necesario para demostrar las propiedades independientes de la lengua que poseen los sitemas; el concepto "pluridominio" es importante para demostrar la eficacia, en distintos entornos, del modelo lingüístico general que se desarrolló en el anterior proyecto (Esprit 860).

Se están desarrollando dos aplicaciones concretas que utilizarán tanto RPA como CTH (8) y una sólo para CTH (9). Estas aplicaciones son una extensión de aplicaciones actualmente en uso con la inclusión de las posibilidades que introduce la tecnología del habla. En todas ellas, se efectuará un estudio de mercado y ergonómico y se construirán prototipos que, al final del proyecto, puedan demostrar su interés y su eficacia.

4. Conclusiones

Como se puede apreciar, el trabajo del proyecto POLYGLOT abarca un amplio espectro de aspectos críticos en la tecnología del habla. El desarrollo del sistema de RPA para los siete idiomas del proyecto constituye un importante punto de partida para el posterior desarrollo del RHC, donde se están desarrollando importantes herramientas, independientes del lenguaje, que permitirán una más fácil extensión a las otras lenguas y un manejo más eficaz de todas sus bases de datos. Por otra

parte, el diseño de una nueva arquitectura multicapa para CTH ofrece la relevante ventaja de una mejora de la calidad en la salida hablada. Finalmente, el trabajo realizado en reconocimiento de Habla Continua constituye una base para futuros trabajos. En resumen, al incorporar tanto sistemas de reconocimiento como de síntesis se consigue un completo sistema de análisis de voz para siete lenguas que, sin duda, resultará de interés para la implementación de sistemas comerciales que permitan una cómoda interacción oral entre el hombre y la máquina.

Referencias:

Billi, R. et al.: APC-based very large vocabulary isolated word speech recognition system". *Proc.Eurospeech*, 1989.

Fourcin, A.; Harland, G.; Barry, W. y Hazan, V.: *Speech Input and output assessment. Multilingual Methods and Standards*. Chichester, Ellis Horwood Ltd., 1989.

Vittorelli, V.; Adda, G.; Billi,R.; Boves, L.; Jack, M.; Vivalda, E.: "Esprit POLYGLOT Project: Multi Language Speech Technology". *Acoustics Bulletin*, Octubre, 1990, págs. 17-20.

NOTAS:

1. En este proyecto participan, además, por parte de la U.P.M., Javier Ferreiros, Manuel Leandro, José Colás, José A. Vallejo, Javier Macías, Jorge Sabater, Jesús Diego, Miguel A. Berrojo, Javier Menéndez Pidal y Sadot Alezandres. Por parte de la U.N.E.D., Victoria Marrero, Susana Cano, Angeles Romero y Miguel Martínez.
2. Su título completo es "POLYGLOT-I. A multilanguage speech-to-text and text to speech system", y es el número 2104 del Programa Esprit, de la Comunidad Europea.
3. El proyecto "Linguistic Analysis of the European Languages", número 860 del Programa Esprit-1
4. Este tratamiento común es semejante al que se realizó en el Esprit 860 con las bases de datos escritas, de las que se obtuvo la información lingüística a la que ya hemos aludido y que se utiliza en el proyecto actual.
5. "Speech Assessment Methodologies", Proyecto Esprit 1541 (v. Fourcin et al., 1989)
6. La estructura general del diccionario fue elaborada a partir de la que Olivetti ya había desarrollado para el italiano.
7. Actualmente, el italiano considera 70 prototipos, en inglés el número se sitúa en torno a los 20 y en español estamos utilizando 30.
8. Estas dos aplicaciones se centran en dos entornos específicos, el de oficina y el médico.
9. También basándose en dos entornos diferentes: el de oficina y la enseñanza.