"ARTICULATORY - ACOUSTIC CORRELATIONS IN COARTICULATORY PROCESSES: A CROSS-LANGUAGE INVESTIGATION" (PROYECTO ACCOR)*

Francisco Casacuberta, Isabel Alfonso, María José Castro, Andrés Marzal, Juan Carlos Pérez, Andrés Sánchez, José Miguel Benedí, Isabel Galiano y Enrique Vidal.

Departamento de Sistemas Informáticos y Computación.
Universidad Politécnica de Valencia

Resumen

El proyecto ACCOR trata de nuevas técnicas para la obtención de representaciones de la señal vocal en el marco de los modelos de producción de la voz. Concretamente, se estudian las representaciones articulatorias, incluyéndose el fenómeno de la coarticulación como una fuente más de información.

En este artículo se hace una breve presentación del proyecto, sus objetivos y las fases del mismo, para finalizar con una descripción de los trabajos desarrollados en el seno del grupo de Reconocimiento de Formas e Inteligencia Artificial de la Universidad Politécnica de Valencia.

1.Introducción

El habla es la forma más habitual con la que se intercambian informaciones los seres humanos. Sin embargo, el diseño de un interfaz general basado en la voz que permita la comunicación entre un ser humano y un computador , hoy en día, constituye un problema abierto, al que solo se le ha dado soluciones particulares.

La primera etapa en un sistema de reconocimiento automático del habla tiene como misión la obtención de una representación adecuada de la señal vocal (parametrización). A partir de ésta, se aplican, típicamente, técnicas de Reconocimiento de Formas con el objeto de alcanzar una interpretación aceptable de la pronunciación realizada [Casacuberta, 87] . En la actualidad existen muchas técnicas de parametrización que pueden clasificarse en perceptuales y productivas; las primeras están basadas en modelos auditivos y son las más utilizadas, y las segundas lo están en modelos de pronunciación de la voz.

El proyecto que nos ocupa se denomina ACCOR y trata de desarrollar nuevas técnicas de representación de la señal vocal en el marco de la segunda de las categorías. Concretamente, se estudian las representaciones articulatorias, esto es, representaciones basadas en la geometría del tracto vocal humano durante el proceso de articulación del sonido vocal, incluyéndose la coarticulación o fenómenos de inercia como fuente de información [Marchal, 1989].

Es interesante notar que el proceso de obtención de información acústica a partir de la articulatoria es relativamente bien conocido, mientras que el proceso inverso no lo es, aunque existen muchos trabajos en esta dirección. Este consiste en reconstruir el proceso articulatorio a partir de información acústica.

* Proyecto subvencionado por la Comunidad Europea en el programa ESPRIT-II (BRA), número de acción 3279

Este proyecto está siendo financiado por la Comunidad Europea en el marco de ESPRIT-II (BRA), y para llevarlo a cabo se formó un consorcio con los siguientes participantes:

- Centre National de la Recherche Scientifique. Institut de Phonétique. Aix-en-Provence. Francia. (Coordinador principal)
- University of Dublin. Dublin. Eire.
- Ludwig Maximilians-Universität. Munich. Alemania.
- Consiglio Nazionale delle Ricerche. Padua. Italia.
- University of Reading. Reading.Reino Unido.
- Siemens AG. Munich. Alemania.
- Universidad Politécnica de Valencia. Valencia. España.
- Universidad Autónoma de Barcelona. Barcelona. España.
- University of Stockholm. Estocolmo. Suecia.

El proyecto empezó el uno de Abril de 1989, tiene una duración de 30 meses y un presupuesto de aproximadamente 2,100.000 ECUS; la aportación de la Comunidad Europea es de aproximadamente 1,700.000 ECUS.

2. Objetivos del proyecto ACCOR

Uno de los problemas de mayor dificultad en el Reconocimiento Automático del Habla es la modelización de la gran variabilidad acústica que aparece en la realización de un mismo fonema. La hipótesis de trabajo en la que se basa este proyecto consiste en considerar que esta modelización puede ser más tratable si se puede analizar la variabilidad acústica en base al comportamiento coarticulatorio subyacente.

Los objetivos de este proyecto son los siguientes:

1) La obtención de varios corpora de datos vocales en varias lenguas de la Comunidad Europea, sobre los cuales se realizarán los estudios articulatorios comparativos necesarios.

2) Estudio de los subsistemas motores que intervienen en el habla bajo el punto de vista de la coarticulación. Determinación de las restricciones que existen en la producción de sonidos.

3) Estudio de las relaciones que permiten determinar los movimientos articulatorios a partir de la señal vocal.

4) Propuesta de mejores procedimientos para la obtención de la señal vocal a partir de las configuraciones del tracto vocal.

3. Plan de Trabajo en el proyecto ACCOR

Para alcanzar los objetivos descritos en el apartado anterior se propuso el siguiente plan de trabajo que está dividido en cuatro bloques: (Work-Packages -WP-)

WP1: Adquisición de datos (30 meses)

1. Mejora de las técnicas disponibles: Electropalatografo, glotografo, etc.
2. Investigación de nuevos métodos: Tomografía computerizada, ultra-sonidos, etc.
3. Evaluación de las metodologías.

4. Registro de parámetros articulatorios y de la señal vocal.
5. Propuesta de una estación de trabajo multicanal.

WP2: Análisis (24 meses)

1. Revisión bibliográfica. (4 lenguas)
2. Definición de los corpora.
3. Análisis fonético.
4. Análisis automático.
5. Construcción de un banco de datos .

WP3: Interpretación (30 meses)

1. Adecuación de modelos lingüísticos y físicos
2. Relaciones temporales.
3. Discontinuidades.
4. Estudio estadístico.
5. Reglas de producción.

WP4: Simulación (30 meses)

1. Desarrollos algorítmicos.
2. Modelización.
3. Simulación con restricciones articulatorias .

El WP1 fue concebido para la obtención de los datos vocales tanto en forma de señal obtenida mediante un micrófono, como en la forma que se considerase necesario para el estudio de los movimientos articulatorios (palatografo) (Objetivo 1). El segundo y tercer WP están dedicados a la revisión del estado del arte sobre relaciones acústico-articulatorias y los consiguientes modelos que se han establecido y se puedan establecer (Objetivos 2 al 4). Finalmente, el último WP está dedicado a la posible utilización de los modelos articulatorios existentes y que puedan surgir de este proyecto en el reconocimiento automático del habla (Objetivo 3).

4. Descripción de los trabajos realizados por el grupo de reconocimiento de formas e inteligencia artificial.

El trabajo del grupo de Reconocimiento de Formas e Inteligencia Artificial (RFIA) ha repercutido en algunas de las tareas descritas en el apartado anterior. Resumidamente, estas tareas consisten en en el desarrollo de un entorno para la segmentación y etiquetado automático (WP 1.5 y WP 2.4), el diseño de un software de manipulación para una base de datos con ciertas peculiaridades (WP 2.5), y en la prueba de sistema de parametrización articulatoria en reconocimiento de palabras aisladas con Modelos de Markov y Redes Neuronales (WP 4.3).

4.1 Sonografía: un sistema interactivo de segmentación de señal vocal

Uno de los objetivos del proyecto de investigación ESPRIT-ACCOR es la segmentación y etiquetado de señal vocal en ciertas unidades de nivel subléxico, así como el desarrollo de

aplicaciones para la visualización de señal en los dominios temporal y frecuencial.Sonografía es un sistema interactivo WIMP (windows, icons, mouse, pull-down menus) de segmentación y visualización para ordenadores compatibles PC que cubre los mencionados objetivos [Marzal & Puchol,91a], [Marzal & Puchol,91b].

Como herramienta de visualización, Sonografía proporciona la posibilidad de estudiar en detalle fragmentos de la señal y su espectrograma (secuencia de transformadas discretas de Fourier), así como realzar aspectos de éste último mediante ajustes en factores de escala o contraste, o incluso representándolo en tres dimensiones.

Sin embargo, es en la tarea de segmentación de señal acústica donde Sonografía revela todo su potencial. Hoy por hoy, la segmentación y etiquetado de la señal debe ser desempeñada manualmente por expertos en fonética. Esta tarea es monótona, repetitiva y suele resultar plagada de errores al tener que efectuarse sobre corpora de centenares (o incluso millares, como es el caso de este proyecto) de pronunciaciones.

Sonografía, además de ser un cómodo editor de segmentación, incorpora ayudas al experto que hacen su tarea mucho más fácil y minimizan la posibilidad de cometer errores.

Una de dichas ayudas consiste en la "discretización" del espacio de posibles ubicaciones para marcas de inicio/final de segmento en base a la la denominada "detección del cambio espectral". Con esta opción activada, el usuario sólo puede situar marcas en aquellos puntos de la señal en los que el espectro experimenta variación, pues es en esas zonas donde empiezan/finalizan segmentos acústicos homogéneos [Vidal & Marzal,90].

Otra de las posibilidades ofrecidas por Sonografía es la de empezar a segmentar apoyándose en el proceso denominado "Segmentación Multinivel" [Glass & Zue,88], [Zue et al.,89]. Dicho proceso parte de una "sobresegmentación" (usualmente obtenida por detección del cambio espectral) y proporciona una jerarquía de posibles segmentaciones (de mayor a menor número de marcas) de entre las cuales el usuario escoge la que considere más adecuada.

Además de las características antes mencionadas, Sonografía proporciona ayudas para el etiquetado de segmentaciones, permitiendo la edición de etiquetas asociadas a cada marca, importando tiras de etiquetas de un fichero de texto y sugiriendo a partir del etiquetado en curso cual es la siguiente etiqueta a añadir.

Actualmente, se sigue trabajando en el desarrollo de Sonografía con el fin de transportarlo a plataformas UNIX/XWINDOWS e incorporar nuevas prestaciones (gestión de un procesador digital de señal, audición de fragmentos de señal, etc.).

4.2 La base de datos de ACCOR

En este proyecto se van a manipular una gran cantidad de datos acústicos (aproximadamente 21 Gbytes) en distintos experimentos, por lo que surge la necesidad de crear un sistema que proporcione herramientas para buscar y copiar toda la información que se necesite en un momento dado, de forma fácil y rápida. La base de datos que se presenta a continuación pretende llevar a cabo esta tarea [Pérez, 91a y 91b].Los datos manejados corresponden a pronunciaciones representativas de diferentes aspectos de los lenguajes estudiados en el proyecto ACCOR y bajo distintas representaciones (señal vocal, señal de palatografo, etc.). La idea es facilitar el acceso a los datos y al diseño de experimentos con tales corpora voluminosos.

La base de datos no contiene directamente dichas pronunciaciones fonéticas, sino información acerca de ellas, tal como la ubicación en el dispositivo de almacenamiento, locutor, pronunciaciones,etc. es decir, hay una separación de los datos de voz propiamente dichos y los datos útiles para su localización en el medio de almacenamiento elegido, los cuales se encuentran en ficheros de texto (hay 1 para cada pronunciación); la información que contienen dichos ficheros es lo que realmente se almacena en la base de datos.

Algunas de las operaciones que proporciona el prototipo son: añadir/borrar un disco de la base de datos, extracción de una selección de ficheros del disco duro, añadir ficheros seleccionados por locutor, sesión, pronunciación, etc. El tipo de almacenamiento usado hasta el momento es el de discos flexibles [Pérez, 91a].

Una prueba preliminar con los usuarios potenciales de la base de datos ha demostrado la viabilidad de su estructura lógica. Sin embargo, para un producto software final se necesitará realizar: primero, una lista completa de la especificación de los campos requeridos, y segundo, la elección del medio de almacenamiento, esto es: discos ópticos, WORM o DAT. Esta última solución es la más lenta pero de menor coste. Quizá una solución híbrida de discos ópticos y DAT sea la más adecuada. Otras consideraciones a realizar son la distribución y seguridad de los datos todo lo cual conducirá a una implementación del producto software final [Pérez, 91a].

4.3 Reconocimiento de palabras aisladas utilizando una representación articulatoria.

El grupo investigador de SIEMENS ha desarrollado, parcialmente en este proyecto, un sistemas que proporciona una representación articulatoria de la señal vocal [Schmidbauer, 89] [Hoge 91], y que fue utilizado para la modelización de fonemas. El objetivo de nuestro trabajo fue el estudio comparativo entra la representación articulatoria mencionada y una convencional (perceptiva). La representación articulatoria (RA) consistía en 7 parámetros para el modo de articulación y 11 parámetros para el punto de articulación [Casacuberta, 91].

El corpus utilizado en los experimentos está compuesto por 10 pronunciaciones de los diez dígitos castellanos realizadas por diez locutores (5 femeninos y 5 masculinos) con un total de 1000 pronunciaciones. La adquisición del corpus se realizó a una frecuencia de 16 Khz.Con estos datos se han realizado experimentos con dos aproximaciones: Perceptrón Multicapa (PM) y Modelos de Markov Ocultos (MMO). Los experimentos con PM utilizan directamente los vectores de RA. Para los experimentos con MMO se han utilizado 2 representaciones: una etiqueta fonética única por cada segmento fijo de voz ("frame") y un vector con todos los posibles etiquetas con sus probabilidades asociadas (MMO Difusos).Además, para contrastar estos resultados, se han realizado experimentos con los dígitos parametrizados convencionalmente (vectores de coeficientes Cepstrales, obteniendo, posteriormente, cadenas de etiquetas mediante Cuantificación Vectorial para los experimentos con MMO).Se ha realizado un experimento del tipo "Leaving-One-Out" [Duda,73] sobre todo el corpus (parametrización realizada con las dos aproximaciones) con el objetivo de procesar 1000 muestras efectivas de test, siendo los resultados más significativos los que se detallan en la siguiente tabla.

Tabla 1:Resultados de los experimentos usando PM y MMO con los Dígitos parametrizados por Siemens -RA- y por el GRFIA (Valencia) -Cepstrales-. (Tasa de Reconocimiento total).

Parametrización.	Modelo	%Rec.
Características Articulatorias	PM #entrada=80x18,	
	#ocultas=20, #salida=10.	92.4
Etiquetado Características		
Articulatorias	MMO-1 etiqueta	
	fonética/frame #estados=30	99.1
Etiquetado Características		
Articulatorias	MMO-representación difusa	
	#estados=30	99.1
Coeficientes Cepstrales	PM #entrada=80x11,	
	#ocultas=20. #salida=10	94.6
Etiquetado Coeficientes		
Cepstrales	MMO 32 etiquetas #estados=30	97.9

5. Comentarios finales.

Este proyecto finalizará a finales de Septiembre de 1991, y su planteamiento inicial resultó ser demasiado ambicioso como para que estuviese finalizado en esa fecha. En la última revisión del proyecto realizada en Abril de 1991, los representantes de la Comunidad Europea estuvieron de acuerdo en que los corpora adquiridos en varias lenguas era una de las principales aportaciones del proyecto, ya que pueden ser de gran utilidad para la comunidad científica. No obstante, para que estos corpora sean útiles, es necesario que vayan acompañados por un sistema de gestión de bases de datos. Uno de los objetivos de la continuación del proyecto ACCOR consiste en disponer de un sistema acabado que permita a los investigadores el uso cómodo de estos datos. En la actualidad se ha distribuido una copia del prototipo realizado en nuestro laboratorio a todos los participantes de este proyecto, con el objetivo de que sea probado.

También se ha distribuido una copia del sistema Sonografía para estudiar su adecuación para la segmentación y etiquetado de los corpora adquiridos en este proyecto. Posteriormente se estudiará su incorporación a la propuesta de estación de trabajo multicanal.

Por último, se está finalizando los trabajos comparativos entre los sistemas de parametrización articulatoria propuesta por SIEMENS y sistemas convencionales (perceptivos). Los resultados obtenidos hasta ahora no parecen ser concluyentes.

6. Referencias

F.Casacuberta & E.Vidal (1987): "Reconocimiento Automático del Habla". Marcombo.

F.Casacuberta, M.J.Castro, O.Schmidbauer (1991): "The use of Articulatory Features in Isolated Word Recognition: A Preliminary Study". **Periodic Progress Report 2. Action 3279.** Ed. A.Marchal. Abril.

R.O. Duda, P.E. Hart (1973). **Pattern Recognition and Scene Analysis.** John Wiley and Sons.

J.Glass & V.Zue (1988): "Multi-Level Acoustic Segmentation of Continuous Speech". **Proc of the ICASSP 88,** pp. 429-432

H.Höge & O. Schmidbauer (1991): "Acoustic-Articulatory Decoding". **Periodic Progress Report 2. Action 3279.** Ed. A.Marchal. Abril.

A. Marchal (1989): "Articulatory-Acoustic Correlations in Coarticulatory Processes: A Cross-Language Investigation". **Technical Annex. Action 3279.** Enero.

A.Marzal & J.Puchol, (1991a): "Sonografía: An Interactive Segmentation System of Acoustic Signals based on Multilevel Segmentation for a Personal Computer".**Periodic Progress Report 2. Action 3279.** Ed. A.Marchal. Abril.

A.Marzal & J. Puchol, (1991b):"Sonografía: User's Manual". **Informe Técnico. DSIC-II/16/1991.**

J.C.Pérez, I.Alfonso & F.Casacuberta, (1991a): "The ACCOR database: Design of the DB management system". **Periodic Progress Report 2. Action 3279.** Ed. A.Marchal. Abril.

J.C.Pérez, I.Alfonso, (1991b): "ACCORDB. User´s Manual". **Informe Técnico. DSIC-II/1991.**

O. Schmidbauer (1989): "Robust Statistic Modelling of Systematic Variabilities in Continuous Speech incorporating Acoustic-Articulatory relations". **Proc of the ICASSP 89.** pp 616-619.

E. Vidal & A. Marzal (1990): "A Review and New Approaches for Automatic Segmentation of Speech Signals". **Proc. of the EUSIPCO 90.** Septiembre.

V. Zue, J. Glass, M.Philips & S.Seneff (1989): "Acoustic Segmentation and Phonetic Classification in the Summit System". **Proc of the ICASSP 89,** pp. 389-392.