* DESARROLLO DE CORPUS PARA INVESTIGACION EN TECNOLOGIAS DEL HABLA (ALBAYZIN)#

F. Casacuberta1, R. Garcia2, J. Llisterri3,
C. Nadeu4, J.M. Pardo5 and A. Rubio6 *

1Universitat Politècnica de València (UPV), Dpto. DSIC
2Universidad Politécnica de Madrid (UPM), Dpto. DSSR
3Universitat Aut noma de Barcelona (UAB), Dpto. DFE
4Universitat Politècnica de Catalunya (UPC), Dpto. DTSC
5Universidad Politécnica de Madrid (UPM), Dpto. DIE
6Universidad de Granada (UG), Dpto. DETC

RESUMEN

El proyecto ALBAYZIN trata de superar la falta de corpus adecuados para el desarrollo de sistemas de reconocimiento automático del habla en lengua castellana. En este proyecto se diseñan y construyen tres corpus distintos con: 1) frases equilibradas fonéticamente, 2) frases correspondientes a una tarea de consulta a una base de datos geográfica y 3) habla producida en condiciones adversas. En el presente artículo se describen tanto las características generales de los tres corpus mencionados como la metodología que se sigue en su elaboración.

Proyecto subvencionado por la CICYT (TIC91-1488-C06).

* La lista de coautores ha sido ordenada alfabéticamente. Cada uno de ellos pertenece a un grupo distinto de los seis que componen el consorcio del proyecto ALBAYZIN. El grupo coordinador del proyecto es el de la UPC. Los demás investigadores participantes en el proyecto son: J. Díaz y A. Peinado (UG); S. Aguilera y J. Menéndez-Pidal

(UPMDIE); J. Gómez y J. Santos (UPM-DSSR); N. Prieto, E. Sanchís, E. Segarra y E. Vidal (UPV); D. Poch (UAB); A. Bonafonte, E. Lleida, J.B. Mariño y A. Moreno (UPC).

1. Introducción

El desarrollo de bases de datos de voz es de crucial importancia dentro del área de investigación en tecnologías del habla y especialmente en reconocimiento y comprensión del lenguaje hablado.

El interés estriba en disponer de grandes corpus que representen toda la variabilidad inherente a la señal considerada. Los factores de variabilidad pueden ser fonéticos (p. ej. influencia de los contextos en los que pueden aparecer los distintos alófonos), acústicos (diferentes ambientes acústicos de trabajo), variabilidad en el mismo locutor (distintos estados psicológicos y fisiológicos, velocidad de locución, etc.) y variabilidad entre locutores (diferencias dialectales, sociolingüísticas, etc.).

Estos corpus deben servir para el entrenamiento y evaluación de sistemas de reconocimiento y procesado del habla ya desarrollados o que se desarrollarán en un futuro cercano. Muchos de estos sistemas utilizan modelos estadísticos, cuyos parámetros deben ser estimados a partir de una gran cantidad de datos. Así mismo, el disponer de una base de datos común en una lengua determinada permite la realización de análisis comparativos de las prestaciones de los distintos sistemas desarrollados dentro de la comunidad científica interesada en dicha lengua.

Gracias a los esfuerzos realizados en los últimos años en distintos países, existen actualmente un cierto número de corpus de habla: [Fourcin et al., 89], [Garafolo et al., 89], [Kurematsu et al., 89], [Price et al., 89], [Shirai et al., 89], [Zue et al., 89], etc., principalmente para la lengua inglesa. El esfuerzo requerido para producirlos excede a menudo la capacidad de un grupo de investigación aislado [Fourcin et al., 89], [Price et al., 89], [Eskenazi, 89]. Ello conlleva la necesidad de cooperación entre distintos grupos, cooperación que, por otro lado, permite evitar cierta reiteración de trabajos y, al mismo tiempo, posibilita la creación de estándares ampliamente aceptados.

Hasta ahora, en nuestro país se han realizado esfuerzos muy limitados para desarrollar corpus de habla, los cuales ni han alcanzado una cobertura a gran escala de las fuentes más importantes de variabilidad que se encuentran en el habla, ni permiten entrenar y evaluar un sistema de reconocimiento en una tarea de ciertas dimensiones. Esta falta de corpus de habla castellana para la investigación en tecnologías del habla, y especialmente

para el desarrollo de sistemas de reconocimiento, ha sido el motivo de la formación de un consorcio constituido por seis grupos españoles de investigación en tecnologias del habla, los cuales van a diseñar y producir (en colaboración con una empresa española) un conjunto de corpus. El nombre y afiliación de los seis grupos se presenta al principio del presente artículo.

El proyecto correspondiente se denomina ALBAYZIN[1]. Financiado por la CICYT (TIC91-1488-C06), el proyecto arrancó en octubre de 1991, aunque de hecho durante casi dos años tuvo lugar una etapa previa de establecimiento de objetivos. La duración del trabajo es de dos años y el resultado final será un conjunto de tres corpus cuyas características generales se describen en este artículo.

Es de esperar que estas bases de datos permitan abordar el diseño de sistemas de reconocimiento y comprensión del habla más ambiciosos que los planteados hasta el momento en los trabajos de los grupos integrados en el consorcio, lo que posibilitará la aportación a la comunidad científica de resultados de mayor significación, y facilitará la solución a los problemas que actualmente se padecen en el intercambio de experiencias.

2. Características generales de los corpus de habla ALBAYZIN

El objetivo del proyecto ALBAYZIN es el diseño y construcción de una base de datos que contenga una colección suficientemente amplia de unidades lingüísticas (frases, palabras, fonemas, etc.) en castellano, de forma que permita el desarrollo y la evaluación de sistemas de reconocimiento y procesado del habla.

Una vez producida, la base de datos ALBAYZIN estará compuesta por tres corpus distintos que, en el momento de redactar este artículo, se encuentran en vias de definición detallada. A continuación se explican las características más sobresalientes de cada uno de ellos.

1) El primer corpus se basa en pronunciaciones de frases entresacadas de un conjunto de unas 200 frases equilibradas fonéticamente desde el punto de vista del reconocimiento del habla. Estas frases no presentan restricciones sintáctico-semánticas y se diseñan con

[1] La denominación Albayzin proviene del barrio granadino del mismo nombre. En él tuvo lugar una reunión del entonces futuro consorcio que resultó decisiva para la consolidación del proyecto.

el objetivo de cubrir adecuadamente un amplio margen de variabilidad fonética. Como factores de variabilidad fonética [1]se consideran, entre otros, los siguientes:

a) factores dependientes del locutor, como el sexo, el dialecto o la rapidez de

pronunciación;

b) factores de tipo fonético tales comoel conjunto de alófonos, el contexto (alófonos anterior y posterior), la posición en la sílaba y el grado de acentuación.

2) El segundo corpus es más específico y es dependiente de la aplicación, estando formado por frases correspondientes a una tarea de consulta a una base de datos geográfica. Las frases están sometidas a una fuerte restricción semántica, con el fin de incluir información relativa a este aspecto en el reconocimiento. Las construcciones sintácticas se extraen analizando la transcripción de entrevistas a diferentes personas, con y sin experiencia previa en el manejo de bancos de datos, en las que éstas intentan obtener información sobre la geografía española. Así mismo, el léxico se controla para que el número total de palabras diferentes oscile alrededor del millar.

3) El tercer corpus tiene como característica propia el hecho de contener habla producida en un ambiente adverso, para reflejar el hecho de que la voz cambia significativamente cuando el hablante soporta un nivel sonoro alto en sus oidos (efecto Lombard). No se considera necesaria la grabación de habla con ruido de fondo porque, en primera aproximación, puede considerarse que el ruido ambiental es de tipo aditivo, y por tanto puede grabarse separadamente de la señal y luego añadirse a ésta con el nivel deseado.

Se utiliza aproximadamente un número de 150 locutores, repartidos por igual entre sexos. Al final se dispondrá de unas 5000 frases por corpus, de 3 o 4 segundos de duración cada una. El texto de las frases se selecciona a partir de muestras de habla espontánea.

La base de datos final estará constituida por las señales de voz convenientemente filtradas y muestreadas, organizadas de forma que permitan acceder a diferentes subgrupos atendiendo a características como el sexo, la variedad dialectal, etc. Así mismo, contendrá las consiguientes documentaciones ("cabeceras"), con información sobre las señales, las características relevantes de cada locutor, rasgos lingüísticos de la frase y entorno en el que se pronunció.

En un proyecto con voluntad estandarizadora como es ALBAYZIN, es esencial considerar atentamente las características del software y hardware ligados a la producción de las bases de datos de habla. En este sentido, ALBAYZIN tiene la voluntad de seguir lo más de cerca posible los estándars propuestos en el proyecto europeo Speech Assessment Methods (SAM) de ESPRIT. Respecto al software, dentro de SAM se han desarrollado herramientas para adquisición, estructuración y manejo de los datos contenidos en los corpus [Castagneri et al., 89], [Hendriks, 89], las cuales han sido diseñadas para resultar de fácil manejo al usuario.

En cuanto al hardware es de destacar que los corpus que se producirán llegarán al usuario en un soporte físico CDROM, es decir, el mismo que se utiliza en el proyecto SAM. De hecho, aunque el soporte hardware de los corpus disponibles actualmente en las diversas lenguas no está estandarizado y pueden observarse una gran variedad de soportes -desde la tradicional cinta magnética de carrete abierto hasta los modernos DAT y los discos ópticos- la mayoria de corpus de habla de cierta relevancia están decidiéndose cada vez más por el CDROM como soporte final para el usuario, tanto por su alta capacidad de almacenamiento y su facilidad de uso, como por su precio, que resulta relativamente bajo cuando la existencia de muchos usuarios hace posible la producción de una gran cantidad de copias.

3. Descripción de la metodología que se sigue en el proyecto

La metodología que se sigue para la realización de los corpus se puede resumir en los siguientes puntos:

1) Aspectos comunes previos

En primer lugar, es necesario definir el proceso de adquisición de datos, fijando aspectos prácticos como son: frecuencia de muestreo, número de bits por muestra, tipo de instrumentos de grabación, etc. Asimismo, se debe decidir la estructura final de los mismos en el dispositivo de almacenamiento, así como los criterios de agrupación y diferenciación de los distintos tipos de datos.

Por otra parte, también es necesaria la definición y puesta a punto de los equipos informáticos, con el fin de compatibilizar los de los diferentes grupos que intervienen en este proyecto, haciendo así posible tanto el acceso a los datos finalmente obtenidos, como el intercambio de otros datos, informaciones, programas, etc.

Antes de proceder a la grabación del corpus es indispensable seleccionar y adiestrar a los locutores que pronuncian las frases. La selección se hace atendiendo a criterios de edad, sexo y variedad dialectal.

2) Tareas comunes:

La confección de los corpus requiere también la realización de tareas que por su naturaleza son comunes a todos ellos, si bien han de llevarse a cabo independientemente para cada uno. Así, en los tres casos es necesario lo siguiente:

a) adquisición de las señales y grabación en el formato elegido;
b) determinación automática, con revisión manual, del principio y fin de cada frase;
c) transcripción fonética del material grabado;
d) segmentación manual de una pequeña parte de las señales adquiridas con objeto de indicar las fronteras entre alófonos; y
e) estructuración y documentación de las grabaciones.

3) Tareas específicas:

a) Corpus fonético:

Selección de alófonos y determinación de frecuencias de aparición.
Estudio de los factores principales de variación fonética.
Diseño del conjunto de frases.

b) Corpus de aplicación:

Diseño de la aplicación.
Recolección de frases mediante entrevistas.
Análisis de las frases y propuesta del conjunto de frases.
Extracción de la estructura sintáctica.

c) Corpus en ambiente adverso:

Selección del conjunto de frases.
Diseño del ambiente ruidoso.

4. Consideraciones finales

En nuestra opinión, el interés del conjunto de corpus ALBAYZIN reside en los siguientes puntos:

a) Constituye un marco de estudio apropiado para el estado actual del reconocimiento del habla, suficientemente amplio, y progresivo en cuanto a su complejidad.

b) Servirá de apoyo fundamental a la investigación en reconocimiento del habla, siendo la base de futuros proyectos en el área; y no solamente para los grupos solicitantes del proyecto, sino para cualquier grupo que investigue en reconocimiento (y, en general, procesado) del castellano.

c) Eliminará el trabajo redundante que se realiza en la actualidad, ya que para cada sistema concreto se adquiere un corpus diferente y excesivamente limitado.

d) Y, principalmente, permitirá la definición de un conjunto de datos estándar, con lo cual se posibilitará en gran medida la contrastación de resultados y los intercambios entre los diferentes grupos, lo que es fundamental para el progreso en la materia.

Asimismo, es de esperar que la realización conjunta del proyecto por los seis grupos produzca una dinámica de colaboración que se puede extender más allá de la duración del proyecto, siendo apoyada por el hecho de compartir un mismo sistema informático de soporte de bases de datos. Evidentemente, esta ventaja es extensible a cualquier otro grupo de investigación en el área, puesto que el resultado del proyecto será público y, por otro lado, la presente agrupación de equipos de trabajo de ningún modo se plantea como cerrada.

REFERENCIAS

- Castagneri, G. Vacchetta, L., Di Carlo, A.: "An application of relational database torecognizer testing workstation". *Workshop on Speech Input/Output Assessment and Speech Databases,* Noordwijkerhout, The Netherlands, 20-30 Sept. 1989.

- Eskenazi, M.: "On coordinated assessment efforts in France". *Workshop on Speech Input/Output Assessment and Speech Databases,* Noordwijkerhout, The Netherlands, 20-30 Sept. 1989.

- Fourcin, A.S. and SAM partnership: "Progress overview of the SAM Project".*EUROSPEECH'89*, p. 308, Paris, 1989.

- Garafolo, J.S., Pallet, R.S.: "Use of CD-ROM for Speech Database Storage andExchange".*EUROSPEECH'89*, pp. 309-312, Paris, 1989.

- Hendriks, J. P.M.: "An acoustic-phonetic formalism for database access". *Workshop on Speech Input/Output Assessment and Speech Databases.* Noordwijkerhout, The Netherlands, 20-30 Sept. 1989.

- Kurematsu, A., Takeda, K., Kuwabara, H., Shikano, K.: "ATR Japansese speechdatabase of speech recognition and synthesis". *Workshop on Speech Input/Output Assessment and Speech Databases.* Noordwijkerhout, The Netherlands, 20-30 Sept. 1989.

- Price, P., Fisher, W., Bernstein, J., Pallet, D.: "The DARPA 1000-word resourcemanagement database for continouous speech recognition". *Proc. ICASSP'88*, pp. 651654, 1988.

- Shirai, K., Fujisaki, H., Itahashi, S.: "Speech database projects in Japan -present andfuture". *Workshop on Speech Input/Output Assessment and Speech Databases,* Noordwijkerhout, The Netherlands, 20-30 Sept.1989.

- Zue, V., Seneff, S., Glass, J.: "Speech database development: TIMIT and beyond".*Workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 20-30 Sept. 1989.