Proyecto SALA: SpeechDat Across Latin America VENEZUELA

Elsa Mora(*),
Manuel Rodríguez(**),
Asunción Moreno(***)
Universidad de Los Andes. Mérida Venezuela
Departamento de Lingüística(*)
Departamento de Electrónica y Comunicaciones(**)
Universitat Politecnica de Catalunya. España (***)
Dep. Teoría de la Señal y Comunicaciones

Introducción

Desde la década de los 80, el desarrollo de los sistemas de reconocimiento de voz ha sufrido un cambio importante. Desde los primeros sistemas de reconocimiento de palabras aisladas, basados en plantillas y probados en condiciones de laboratorio, hasta los sistemas actuales con reconocimiento de grandes vocabularios, habla continua, pudiendo operar por vía telefónica, introducción de diálogo, traducción automática, etc. se ha avanzado tanto en el desarrollo de nuevos algoritmos como en la velocidad de cálculo de los actuales sistemas. La necesidad de entrenar sistemas de reconocimiento de voz cada vez más potentes y con mayores capacidades obliga a obtener grandes cantidades de datos.

El proyecto más ambicioso para la generación de bases de datos por vía telefónica financiado por la Unión Europea es el proyecto SpeechDat: "Speech Databases for Creation of Voice Driven Teleservices". Su objetivo es adquirir una base de datos para reconocimiento de voz en todas las lenguas mayoritarias de la Comunidad Europea. Este proyecto es continuación del proyecto del mismo nombre que finalizó en el año 96 y en el que participaron un gran número de empresas industriales y Universidades. El proyecto ha finalizado y las bases de datos obtenidas corresponden a todas las lenguas oficiales de la comunidad y alguna variación dialectal específica. Las bases grabadas constan de 1000 o 5000 llamadas por lengua dependiendo del tipo de aplicación (red móvil o fija) y el corpus está dividido en varios items siendo los más significativos los siguientes:

· Cada locutor pronuncia más de 40 frases leídas o espontáneas.
· La base de datos incluye un corpus fonético muy amplio. Cada locutor lee 9 frases de este corpus de forma que pronuncia al menos una vez todos los fonemas y en la base se ha maximizado el número de difonemas y trifonemas.
· La base de datos incluye palabras clave, números, dígitos, cadenas de números, cadenas de dígitos, fechas, horas etc.
· La base de datos contiene, para cada señal grabada, una transcripción ortográfica de lo que realmente ha pronunciado cada locutor y una serie de marcas relacionadas con los ruidos que aparecen en las señales grabadas.
· Cada base de datos es validada por un centro específico, SPEX (Holanda), y será hecha pública en breve plazo por ELRA (European Language Resources Association).

La base de datos SpeechDat europea puede equipararse a otras de características similares: MACROPHONE (en inglés) o VAHA ("Voice Across Hispano America" en Español hablado en Estados Unidos), realizadas en EEUU y que pueden obtenerse vía LDC (Linguistic Data Consortium).

EL PROYECTO SALA

El proyecto SALA, SpeechDat Across Latin America, se ha constituido para extender SpeechDat en Latino America. El objetivo del proyecto es la producción de bases de datos de voz que cubran la mayor parte de las regiones dialectales en Latino América para los idiomas español y portugués.

Organización

El proyecto SALA se ha fundado por un consorcio de partners industriales y Universidades. Hasta este momento participan: CSELT (Italia), Lernout & Hauspie(Bélgica), Lucent Technologies (USA), Philips Speech Processing (Alemania), Siemens (Alemania), SPEX (Holanda), Temic (Alemania), UPC (Espàña) y Vocalis (Reino Unido). El consorcio está abierto a otros partners industriales o públicos.

El proyecto SALA es apoyado por una red de Universidades Latino Americanas, entre ellas la Universidad de Los Andes.

La producción de las bases de datos en Español está coordinada por la Universidad Politécnica de Cataluña (España).

La validación de las bases de datos se realiza por medio del Instituto de Validación SPEX (Holanda) y serán distribuidas por ELRA.

Especificaciones

Las bases de datos orales deberán cumplir las especificaciones de las bases SpeechDat las cuales pueden obtenerse fácilmente por internet:
http://gps-tsc.upc.es/veu/sala/

El proyecto SALA tiene prevista la realización de las bases en dos partes: la primera de ellas tiene como objetivo la generación de 8 bases de datos en Español de Hispano América y Brasil, de 1000 locutores cada una. Las zonas de grabación están definidas a priori según unos criterios que incluyen información de tipo lingüístico, división política, para tener en cuenta las redes telefónicas, y demográficas. Las zonas son las siguientes:

Zona 1: Brasil

Zona 2: México

Zona 3: Caribe: Puerto Rico, Cuba, Rep. Dominicana, Venezuela

Zona 4: America Central: Honduras, Guatemala, El Salvador,Costa Rica, Nicaragua

Zona 5: Panamá, Colombia

Zona 6: Ecuador, Perú y Bolivia

Zona 7: Chile

Zona 8: Argentina, Uruguay y Paraguay.

Cada zona se divide a su vez en regiones dialectales, con el objetivo de recoger en cada región, un número significativo de informantes desde el punto de vista de reconocimiento de voz. Una vez generadas las bases de datos, se realizará un estudio de las mismas para determinar desde un punto de vista de reconocimiento del habla:
· la necesidad de una redefinición de las zonas y regiones tanto en el sentido de unificación de algunas de ellas como de separación de otras
· el número de locutores necesario por zona dialectal y región a la vista de los resultados obtenidos.

La segunda parte del proyecto completará las grabaciones hasta llegar a los 5000 locutores por cada zona dialectal.

SALA EN VENEZUELA

DESARROLLO DEL PROYECTO

El objetivo de este proyecto ha sido la puesta a punto y mantenimiento de un sistema automatizado que permita grabar, por vía telefónica, las voces de 1000 personas, a través de la lectura de un texto de 44 ítems.

Instrumento

Para la obtención del corpus de habla por vía telefónica se utilizó la base de datos creada por la UPC, la cual fue adaptada a la realidad lingüística venezolana. Los diferentes ítems de cada hoja entregada para ser leída presentaban distintas frases, leídas o espontáneas, que contenía fechas, números, frases, palabras, letras. El total de dicha base está compilada en un CD.

La adaptación del corpus es fonéticamente balanceada, hecho avalado por la SPEX (Holanda). La información detallada sobre la base de datos aparece en el Informe Técnico Moreno-Mora 1999 SALA SPANISH VENEZUELAN DATABASE.

Búsqueda de informantes

La búsqueda de informantes se realizó a través de dos personas responsables de seleccionar las personas en cada una de las áreas dialectales del país : Andes (Táchira, Mérida, Trujillo), Centro (Distrito Federal, Miranda, Carabobo, Aragua, Lara, Yaracuy,

Falcón), Llanos (Portuguesa, Guárico, Cojedes, Apure, Barinas), Zulia (Zulia), Sud-oriental(Sucre, Nueva Esparta, Monagas, Anzoátegui, Delta Amacuro, Bolívar, Amazonas)

A cada responsable se le indicó una fecha prevista para las grabaciones, pero dados los inconvenientes técnicos, así como aquellos asociados a los
sucesivos cortes de luz, los plazos debieron prolongarse, de modo que el lapso entre la primera llamada completada y la última fue de 6 meses en lugar de los 3 meses establecidos inicialmente. Esto hizo que buena parte de las personas que estaban dispuestas a llamar no lo hicieron y/o perdieron la planilla para la grabación. Hubo entonces que diseñar diferentes maneras para estimular el interés por hacer las llamadas, hasta llegar al número esperado.

Distribución de las hojas de lectura

Para unas 100 llamadas, se generó y se envió a la persona responsable en cada región, 125 hojas distintas, previendo el caso de gente que no llama, pierde la hoja etc. Esto se hizo pues de hecho si no se insistía a las personas a quienes se les solicitaba que llamaran, sólo lo hacía un 30% y, de las llamadas recibidas, muchos colgaban antes de terminar. Al insistir el porcentaje de llamadas recibidas aumentaba.

La hoja de instrucciones se le entregó a cada informante conjuntamente con la hoja que debía leer por teléfono. Luego de hacer la llamada debía pasar a firmar la hoja donde autorizaba el uso de su voz y participaba en el sorteo previsto, el cual se realizaría al culminar las llamadas completas previstas.

Todas las hojas eran distintas y cada locutor sólo podía llamar una vez. Asimismo, dos locutores no podían usar la misma hoja. Sin embargo, fue difícil evitar la repetición de algunas de las planillas.

El proceso de grabación finalizó en abril del año 2000 y ahora contamos con un corpus, de habla telefónica, de español venezolano de más de 1000 hablantes de todas las regiones del país, varios grupos de edad y de ambos sexos, distribuidos de la siguiente manera:

Andes

	<18	18-30	31-45	46-60	>60	?	Totales
Mujeres	3	62	19	13	1	3	101
Hombres	0	45	31	38	1	1	116
Totales	3	107	50	51	2	4	217

Centro

	<18	18-30	31-45	46-60	>60	?	Totales
Mujeres	0	45	32	21	1	3	102
Hombres	0	42	32	26	4	1	105
Totales	0	87	64	47	5	4	207

Llanos

	<18	18-30	31-45	46-60	>60	?	Totales
Mujeres	6	44	36	16	2	1	105
Hombres	5	35	34	16	1	0	91
Totales	11	79	70	33	3	1	196

Zulia

	<18	18-30	31-45	46-60	>60	?	Totales
Mujeres	1	37	42	20	1	1	102
Hombres	1	43	42	26	2	1	115
Totales	2	80	84	46	3	2	217

Sud-oriental

	<18	18-30	31-45	46-60	>60	?	Totales
Mujeres	0	74	25	8	0	3	110
Hombres	0	3	34	26	0	1	94
Totales	0	107	59	34	4	4	204

Totales

	<18	18-30	31-45	46-60	>60	?	Totales
Mujeres	10	262	154	78	5	11	520
Hombres	6	198	173	132	8	4	521
Totales	16	460	327	210	13	15	1041

Aspecto técnico

La plataforma de adquisición consistió en un PC conectado a dos líneas telefónicas por medio de una interfaz especial que coordinaba los siguientes pasos : reconocimiento de llamada, descuelgue, reproducción de mensajes pregrabados, digitalización y almacenamiento de cada respuesta en un archivo distinto, reconocimiento de calidad inadecuada de respuesta y finalmente cuelgue de la línea telefónica.

Originalmente el proyecto estableció que la conexión telefónica debería ser del tipo ISDN. Sin embargo, al determinar que este tipo de telefonía no está implementado en Venezuela por la compañía telefónica nacional, CANTV, se aprobó proceder con líneas telefónicas analógicas tradicionales.

Para el buen desarrollo de este proyecto estaba establecida la firma de un convenio entre la Universidad Politecnica de Cataluña (UPC) y la Universidad de Los Andes (ULA). Antes

de aprobarse dicho convenio, septiembre de 1999, se decidió empezar a trabajar con lo que estaba a nuestro alcance.

Se instaló el sistema en el Laboratorio del Grupo de Procesamiento de Voz, anexo al Laboratorio de Electrónica de la Facultad de Ingeniería.

Se gestionó y se obtuvo una instalación especial de 2 líneas telefónicas correspondientes a la central telefónica de la universidad, con acceso directo y se instaló la tarjeta de adquisición.

El 5 de junio se recibió la tarjeta de interfaz Dialogic Proline/2V. Se instaló y probó exitosamente el software de instalación y el software de adquisición de señales de la Universidad Politécnica de Cataluña. Se presentaron algunos inconvenientes con este software, por ejemplo, una versión estaba programada para trabajar con una codificación logarítmica ley A en vez de la que sirve para el Proline/2V, que es la ley mu. Finalmente con la tercera versión del programa Ada, el 25 de junio, se pudo realizar la primera grabación completa de prueba. Lamentablemente, aún así, el sistema no era apropiado para el trabajo de 24 horas diarias de la adquisición automática de llamadas, se detuvo entonces el trabajo esperando la compra del nuevo equipo.

El 21 de septiembre, se puso a la disposición del proyecto un PC 486, para comenzar a realizar las grabaciones definitivas. La primera grabación completa del corpus se realiza entonces el 27 de septiembre, trabajando con las 2 líneas de la central de la ULA.
Empezaron a llamar localmente, alumnos, amigos y voluntarios que no tuvieran inconveniente en llamar desde la misma universidad o pagando la llamada, pues aún no contábamos con una línea gratuita para el informante.

El 23 de octubre, una tormenta eléctrica dañó una de las líneas de entrada de la tarjeta Dialogic, hecho que retrasó el proceso de llamadas. Un porcentaje elevadísimo de llamadas, aproximadamente un 70 %, resultaron incompletas.

Esto se reportó a España, y como respuesta mandaron otra versión del programa de adquisición, llamada Ana. Sin embargo, no se observó un cambio notable. Una mejoría se logró al correr el programa Scandisk, programa que encuentra y repara los sectores dañados de la memoria. A partir de allí, el porcentaje de llamadas completadas subió casi al 50 %. Este era un mejor porcentaje, pero evidentemente, había mucho margen aún para llegar a un nivel satisfactorio.

En este momento se comienza a mandar por Internet los archivos grabados a
España, para lo cual se habilitó una dirección especial, y con un asistente en España para descomprimir los archivos enviados, dos a tres veces por semana se lograba comprimir entre diez y veinte archivos.

A esta fecha ya se habían hecho las gestiones administrativas y técnicas necesarias para la solicitud de la línea 800. Es importante señalar que este proceso no puede hacerse independiente de la universidad. De hecho, las oficinas de telefonía de la ULA son quienes gestionan este servicio cuando se trata de instalar la línea en recintos universitarios. Finalmente, el 30 de noviembre se realiza la primera llamada a través de la línea 800 y así puede iniciarse la toma de la muestra desde las otras partes del país.

A principios de diciembre se recibe de la Universidad Politécnica de Cataluña el equipo para el proyecto, un Pentium II de 450 MHz, con 13 GB de disco duro, sistema multimedia, sistema operativo Windows98. El 14 de diciembre se mudó la tarjeta Dialogic y el software para este equipo. Lamentablemente hubo problemas técnicos que impedían la grabación, unidos a los problemas de líneas telefónicas ocurridos a raíz de las tragedias