Tecnologías del habla y lenguas minoritarias*

Carmen García Mateo
Dpto. Teoría de la Señal y Comunicaciones
E.T.S.I. Telecomunicación - Universidad de Vigo
36200 - Vigo (Spain)
carmen@gts.tsc.uvigo.es

Resumen: El principal objetivo de esta comunicación es mostrar nuestra experiencia en el desarrollo de tecnología del habla en un entorno de multilingualidad: castellano y gallego, ambos oficiales en nuestra comunidad autónoma pero con un desarrollo desigual en cuanto a tecnología del habla se trata. La falta de recursos para el desarrollo de tecnología en gallego pone en peligro el uso de este idioma en los modernos sistemas de información. Esta situación es similar en otras lenguas minoritarias.
Palabras clave: tecnología del habla, lenguas minoritarias, gallego

Abstract: In this paper we show our latest developments of speech and language technology for two languages: Spanish and Galician. Special attention is devoted to the situation of this minority language: Galician, where the lack of resources puts in danger its inclusion in speech products.
Keywords: speech technology, minority language

1. Introducción

Hemos asistido en los últimos años a la puesta en marcha de algunos servicios automáticos de comunicación hombre-máquina dirigidos a usuarios finales. Algunos ejemplos exitosos son el Servicio 1003 automático de Telefonica y el Portal de emoción Voz de Telefónica Movistar(Villarrubia et al., 2002). A pesar de ello, la puesta en marcha de servicios automáticos en los que una máquina "entienda" el habla y "genere" habla no está respondiendo a las espectativas que se crearon hace unos cinco-diez años cuando quizás por presión de las agencias de financiación de la investigación, se forzó a una prematura transferencia tecnológica.

Se han creado numerosoas empresas dedicadas a estos menesteres, de las cuales a día de hoy muchas han abandonado o desaparecido totalmente. Las razones particulares del fracaso en cada caso pueden ser variopintas, pero existe un denominador común que es la falta de "madurez" de la tecnología del habla. Tanto los sistemas de conversión texto-habla como los de reconocimiento (especialmente estos últimos) necesitan ser "ajustados" con mucho cuidado a la aplicación concreta en la que se estén empleando, así como un grado de "cooperación" y "complicidad" del usuario mucho más elevado que los necesarios con otras tecnologías de interacción hombre-máquina. Es por ello que sigue siendo necesario realizar tanto investigación básica o fundamental, como aplicada para conseguir que el objetivo final de comunicación natural con los servicios automáticos sea una realidad.

* Este trabajo ha sido parcialmente financiado por el MCyT con los proyectos TRANSCRIGAL e ITACA

Si lo anterior es cierto para idomas de amplio uso como puede ser el castellano o el inglés, lo es todavía más para idiomas minoritarios o minorizados como el gallego, euskera o catalán sólo por nombrar los idiomas oficiales del estado español.

Una de las características fundamentales de las tecnologías del habla es la enorme dependencia que presentan con el idioma deque se trate. En casi cualquier sistema se suele hacer una división entre los bloques de procesado de señal, y los bloques de procesado de lenguaje. Para el desarrollo de estos últimos es necesario disponer de conocimientos lingüísticos del idioma en cuestión. También es indispensable disponer de una serie de recursos orales y de texto en el(los) idioma(s) de que se trate. Estas bases de datos, también llamadas corpora, presentan unas características particulares que hace que en muchos casos no sea posible reusar material ya existente, y por ello necesario embarcarse en procedimientos de diseño, captura y etiquetado muy laboriosos y costosos. La no disponibilidad de estos recursos es una de las causas del bajo nivel de desarrollo de la tecnología del habla en idiomas minoritarios como el gallego(García-Mateo, 2001). Sólo con el concurso de la ayuda institucional es posible hoy en día abordar este tipo de proyectos.

Nuestro grupo se ha especializado con el transcurso del tiempo en el desarrollo de tecnología del habla y el lenguaje para el idioma gallego. Es por ello que desde el campo de la codificación de voz a baja velocidad, independiente del idioma, por supuesto, hemos ido abriendo todos los campos necesarios para construir sistemas de diálogo hombre-máquina en gallego. Cronológicamente iniciamos en torno al año 1995 el desarrollo

de un sistema de conversión texto-habla bilingüe gallego-castellano que llamamos *Cotovía* ("alondra" en castellano), para poco después mediante la recogida y etiquetado de una base de datos de voz en gallego tipo SpeechDat (González-Rei et al., 2001), y la construcción de un motor de reconocimiento de habla continua y grandes vocabularios (Cardenal-López, 2001) ser capaces de incorporar reconocedores de voz a nuestros desarrollos.

Toda esta tecnología se ha desarrollado gracias a la financiación de organismos oficiales como el Ministerio de Ciencia y Tecnología (MCyT) y la Xunta de Galicia a través de convocatorias de proyectos de investigación o convenios con el "Instituto Ramón Piñeiro para la investigación en humanidades"(http://www.cirp.es). La tranferencia de esta tecnología hacia las empresas es difícil pues, según nuestra percepción, el acceso vocal a servicios telefónicos o telemáticos se ve más como un servicio de valor añadido que como una necesidad. Esperemos que esto cambie y que con la mejora de la tecnología y el aumento de la demanda por parte de los usarios de estos servicios, se consiga una transferencia efectiva de la tecnología del habla y el lenguaje para idiomas minoritarios.

A continuación describiremos nuestros esfuerzos en conseguir que la tecnología del habla en gallego tenga un grado de desarrollo similar cuando menos al del castellano, mostrando primero las características principales del sistema de conversión texto-habla, para a continuación mostrar nuestros trabajos en transcripción de programas de noticias.

2. Conversión texto-habla

El conversor texto-voz bilingüe *Cotovía* ha sido desarrollado en colaboración con un grupo de lingüistas de la Universidad de Santiago con el importante apoyo del "Centro Ramón Piñeiro para la investigación en humanidades". Se puede comprobar su funcionamiento mediante la conexión a la página web *http://www.gts.tsc.uvigo.es/cotovia*.

Consta cómo casi cualquier sistema actual de dos bloques diferenciados tal como se observa en el diagrama de bloques de la Figura 1: *el módulo acústico* encargado de generar la forma de onda a partir de una curva de entonación, pausado y secuencia de unidades, y de un *módulo prosódico* encargado justamente de generar el conjunto de parámetros anteriormente mencionados a partir del texto de entrada correctamente escrito.

En su primera versión Cotovía utilizaba un método de solapamiento de unidades acústicas pregrabadas de tipo difonema. Este conjunto de unidades cubre la fonética tanto del castellano como del gallego.

A partir de la primera versión de Cotovía, el conversor texto-voz ha sufrido modificaciones

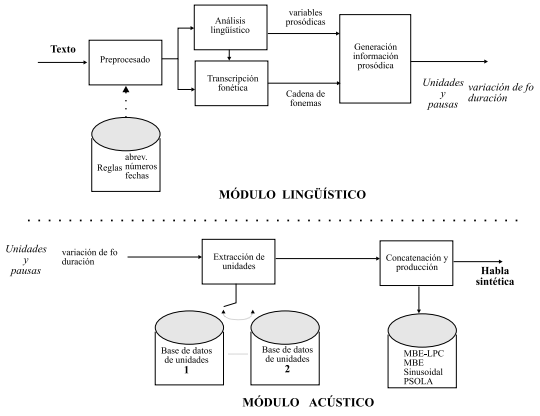

Figura 1: Diagrama de bloques de Cotovia

importantes que mejoran la naturalidad del habla sintética generada y la robustez frente a textos mal construidos. Una de las mejoras en este módulo consiste en la capacidad del conversor para generar voz femenina o masculina. Un avance claro se ha conseguido al desarrollar la versión de síntesis mediante corpus. La idea que subyace debajo de esta aproximación es la de disponer de numerorosas realizaciones de cada unidad acústica de síntesis. La selección del mejor conjunto de unidades dado un texto se realiza mediante técnicas de programación dinámica minimizando una función de coste (Banga et al., 2003). La contrapartida es la mayor carga computacional y memoria de estos sistemas.

En cuanto a procesado del lenguaje (Méndez-Pazó et al., 2003) se trabaja en la elaboración de un corpus de texto en gallego analizado morfosintácticamente por medio de un procedimiento iterativo en el que se intercalan fases de análisis automático con revisiones manuales. Actualmente, el corpus consta de casi 400.000 palabras de texto periodístico no restringido, estando previsto llegar hasta el millón a finales de año.

Del texto analizado se han obtenido dos modelos probabilísticos: uno, referido a la probabilidad de ocurrencia de cada secuencia de categorías, y otro referido a la probabilidad de que cada palabra tenga cierta categoría en el contexto de la frase. Posteriormente, se han desarrollado diferentes algoritmos para la detección de la secuencia de categorías más probable, obteniéndose unos resultados que, aunque no son directamente comparables a los reflejados en otros trabajos sobre lenguas diferentes, son realmente prometedores.

3. Descripción del proyecto Transcrigal

La transcripción automática de programas de noticias presenta numerosas aplicaciones y al mismo tiempo numerosas dificultades(Pallett y Lamel, 2002). Constituye un reto para la comunidad científica en el campo del reconocimiento automático del habla. Mediante el proyecto *Trans-*

crigal realizado en colaboración con lingüistas de la Universidad de Santiago de Compostela y con financiación del MCyT, de la Xunta de Galicia y de la Compañia de Radio Televisión de Galicia prentendemos alcanzar el doble objetivo de mejorar nuestro sistema de reconocimiento de habla continua, y recoger una cantidad importante de recursos lingüísticos en gallego no disponibles en la actualidad. Estos recursos son tanto orales como textuales.

Describimos a continuación los recursos de que disponemos para después describir el sistema de reconocimiento.

3.1. Base de datos Transcrigal-DB

Está base de datos está compuesta por grabaciones de los programas de noticias de la Televisión de Galicia, "Telexornal" y de "Telediarios" de la "Primera de TVE". Cada programa tiene una duración que oscila de los 30 a los 40 minutos. La grabación contiene tanto el fichero de video como el de audio.

Asociado a cada programa está un fichero de etiquetado en formato XML que contiene la siguiente información:

- División en tipos de noticias: titulares, nacional, internacional, etc.
- División en turnos de locutor con identificación del locutor para cada turno.
- Transcripción ortográfica con inclusión e marcas para algunos efectos "audiblesçomo pueden ser repeticiones, ruidos externos, palabras mal pronunciadas, extranjerismos, etc.
- Tipo de ruido de fondo

Esta información está sincronizada con el fichero de audio mediante el empleo de marcas temporales ("breakpoints"). La transcripción está realizada utilizando el programa Transcriber (Barras et al., 2000) específicamente construido para este tipo de bases de datos de audio. En la Figura 2 se muestra a título de ejemplo la pantalla de etiquetado de este programa. En estos momentos se dispone de unas 6 horas de material en gallego totalmente etiquetado y de 2 horas en castellano.

Parte del material de *Transcrigal-DB* se empleará para labores de entrenamiento adaptación de modelos acústicos y de lengauje, y el resto para evaluación/demostración.

3.2. SpeechDat Gal

EL entrenamiento de los modelos acústicos de partida se realiza empleando las base de datos SpeechDat Gal, pues *Transcrigal-DB* no contiene suficiente material.

SpeechDat Gal es una base de datos de 1000 hablantes que responden a un cuestionario de 46 items por teléfono (González-Rei et al., 2001). La recogida de una buena base de datos en el caso del

Figura 2: Ejemplo del etiquetado de Transcrigal-DB

gallego, así como de otras lenguas minoritarias, conlleva una dificultad extra con respecto a otras lenguas estatales o mayoritarias. En primer lugar por no tratarse de una lengua totalmente normalizada y por otra porque la especial situación sociolingüística del gallego (de dos lenguas en contacto) hace habituales las interferencias lingüísticas. El primer objetivo de nuestra base de datos es el de incluir 'buenos hablantes' en gallego. El segundo, el de obtener una representación equilibrada de la variedad de la población, con respecto a varios criterios como el sexo, la edad, el nivel sociocultural y la adscripción dialectal de los informantes.

Mediante técnicas estadísticas de adaptación se reduce el desajuste entre los modelos extraídos a partir de SpeechDat, mejorando el reconocimiento sobre el material de Transcrigal-DB.

3.3. Modelado del lenguaje

Para el entrenamiento de los modelos de lenguaje estadísticos (modelos N-grama) utilizamos el anuario del diario *El País* para el idioma castellano. Para el idioma gallego al no disponer de anuarios del tipo anterior, empleamos material extraído día a día de la Web del periódico digital *O Correo Galego* (refundado como *Galicia Hoxe* a partir del 17 de Mayo de 2003) y de la Web del servicio de metereología de la Xunta de Galicia.

3.4. Transcripción de programas de noticias

La extracción de las noticias se realiza a partir del fichero con la secuencia de sonido correspondiente a un noticiario completo, el cual está almacenado en formato PCM. Para la extracción de las noticias se siguen tres pasos:

- Extracción de segmentos de música y voz.
- Reconocimiento de los segmentos de voz.
- Extracción de las palabras clave y clasificación de la noticia.

La segmentación se lleva a cabo mediante un sistema independiente que se encarga de discriminar los segmentos de voz y los de audio del noticiario, generando marcas temporales con el instante de comienzo de cada segmento y el tipo de segmento (Pérez-Freire y García-Mateo, 2003).

El reconocedor de habla, utilizando como entrada un segmento de audio, genera una secuencia de palabras en un fichero de texto. El proceso de reconocimiento se realiza en tres fases:

- Detección y segmentación del locutor. En esta fase, con el fin de utilizar acústicas adaptadas, se intenta identificar el locutor, en caso de que este exista en la base de datos
- Detección de tópico. Se utilizan para ello diferentes modelos de lenguaje, en función de cual es el más adecuado.
- Reconocimiento. En esta fase se extrae la transcripción y los instantes de comienzo de cada palabra.

Finalmente, en el proceso de extracción de palabras clave se realizan las siguientes funciones:

- Detección de los comienzos y finales de cada noticia.
- Extracción de las palabras claves.
- Detección del tópico.
- Actualización de la base de datos.

4. Conclusiones

El mensaje básico de esta ponencia es un grito en defensa de las lenguas minoritarias que para su preservación y conservación deben subirse al carro de las tecnologías de la información y concretamente a las tecnologías del habla.

5. Agradecimientos

El desarrollo de la tecnología aqui mostrada no sería posible sin la participación de numerosos colaboradores que a lo largo de estos años han participado en nuestros proyectos realizando sus proyectos fin de carrera o contratados como ingenieros de proyecto. La lista de sus nombres sería demasiado larga para aqui explicitarla, pero vaya nuestro agradeciemnto a todos y cada uno de ellos.

Bibliografía

Banga, E. R., E. Fernández Rei, F. Campillo Díaz, y F.J. Méndez Pazó. 2003. Sistema de conversión texto-voz en lengua gallega basado en la selección combinada de unidades acústicas y prosódicas. *Procesamiento del Lenguaje Natural*, 29:153–158.

Barras, C., E. Geoffrois, Z. Wu, y M. Liberman. 2000. Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production. *Speech Communication*, 33(1–2), January.

Cardenal-López, A. 2001. *Realización de un reconocedor de voz en tiempo real para habla continua y grandes vocabularios.* Ph.D. tesis, Universidad de Vigo, Departamento de Teoría de la Señal y Comunicaciones, Diciembre.

García-Mateo, C. 2001. Recursos e actividades necesarias para desenvolver tecnoloxía da fala en galego. En *VIII conferencia internacional de lenguas minoritarias. Santiago de Compostela*, Noviembre.

González-Rei, B., A. Cardenal-López, L. Docío-Fernández, y C. García-Mateo. 2001. Problemática de la recogida y anotación de una base de datos oral para el gallego. *Procesamiento del Lenguaje Natural*, 27:–.

Méndez-Pazó, F.J., E. R. Banga, F. Campillo-Díaz, y E. Fernández-Rei. 2003. Análisis morfosintáctico estadístico en lengua gallega. *Procesamiento del Lenguaje Natural*, páginas–.

Pallett, D.S. y L. Lamel. 2002. Special Issue on Automatic Transcription of Broadcast News Data. *Speech Communication*, 37(1–2).

Pérez-Freire, L. y C. García-Mateo. 2003. Evaluación de un sistema de detección de cambios acústicos sobre programas televisivos de noticias. En *Actas de URSI2003*, Septiembre.

Villarrubia, L., R. SanSegundo, L. Hernández, y G. Escalada. 2002. Tecnología del Habla para desarrollo de aplicaciones Multilingües, Multiservicio y Multiplataforma". En *Actas de las II Jornadas en Tecnologías del Habla. Granada*, Diciembre.