Conversor texto-voz multilingüe para español, catalán, gallego y euskera

Miguel Ángel Rodríguez Crespo, José Gregorio Escalada Sardina, Doroteo Torre Toledano
Tecnología del Habla
Telefónica I+D
C/ Emilio Vargas, 6 28043 Madrid
e-mail: miguel@craso.tid.es

Resumen

Los conversores texto-voz (CTV) han comenzado a ser usados en distintos tipos de aplicaciones o servicios automáticos, entre los que por su difusión destacan aquellos en los que se proporciona información telefónica hablada a los usuarios (coste de llamadas telefónicas, saldo de una cuenta bancaria, ...).

A medida que estos servicios se han ido extendiendo y su uso se ha hecho más común, ha surgido la necesidad de que en ellos se pudiera elegir el idioma en el que el usuario desea recibir la información. Esto es de especial importancia en entornos donde el bilingüismo es habitual entre los hablantes.

La solución obvia consiste en usar diferentes sistemas de conversión texto-voz para los diferentes idiomas, pero esto presenta el inconveniente de tener que duplicar los recursos disponibles, y plantea dificultades para integrar distintos sistemas dentro de un único servicio. Es mucho más eficaz disponer de un único CTV en el que se pueda seleccionar en cada momento el idioma (en el que vendrán escritos los textos, y en el que se generará la voz sintética).

Por otro lado, mientras que es relativamente amplio el número de conversores texto-voz multilingües [Sproat94, Coile93, Lindström93] disponibles comercialmente para los idiomas con mayor número de hablantes en el mundo y con mayor importancia desde un punto de vista económico (inglés, español, francés, ...), son muy escasos (a veces inexistentes) los conversores para idiomas de uso minoritario.

Para dar respuesta a estas necesidades, se ha desarrollado un CTV multilingüe que incluye los idiomas oficiales del estado español (español, catalán, gallego y euskera).

La estructura de esta comunicación es la siguiente:

En la introducción se habla de los antecedentes del CTV multilingüe, y de la manera en la que se planteó su diseño y realización.

En el apartado dedicado a la división en módulos, se describe brevemente la estructura del CTV y sus componentes.

En los siguientes apartados se repasa cada uno de los módulos, destacando las peculiaridades que ha planteado cada idioma y la manera en la que se han abordado.

Finalmente, se incluye un apartado en el que se recogen los principales asuntos pendientes, y algunas conclusiones a las que se ha llegado durante el desarrollo de este trabajo.

1. Introducción

El CTV multilingüe que se describe en esta comunicación ha sido obtenido a partir de un CTV en español que fue desarrollado previamente, y que comenzó a ser empleado en diversas aplicaciones y servicios.

En el momento en que se planteó la necesidad de desarrollar un conversor texto-voz multilingüe, se consideró la posibilidad de realizar un diseño completamente nuevo desde el principio. Esta posibilidad fue finalmente rechazada por varias razones:

- El trabajo necesario para crear una nueva arquitectura y unas nuevas estructuras de datos llevaría mucho tiempo, dada la cantidad de

código que sería necesario realizar de nuevo.

- El CTV en español contaba con una arquitectura bastante modular que permitiría hacer las adaptaciones necesarias para realizar una tarea concreta con una cierta independencia de las demás tareas. Además, esta arquitectura estaba ya bien adaptada para facilitar el uso del CTV en aplicaciones reales.
- El CTV en español demostró en la práctica un funcionamiento muy robusto y con una buena calidad acústica, tanto en inteligibilidad como en naturalidad.

En resumen, el CTV en español ofrecía tanto una base sólida de la que partir, como la suficiente flexibilidad para introducir cambios. Sin embargo, el adaptar y modificar un conversor ya existente ha hecho que, en algunos casos, las soluciones adoptadas para tratar alguna característica particular de un idioma no sean las mejores, y se hayan visto influidas por las limitaciones que da el hecho de partir de una estructura determinada en la que las tareas se realizan en un orden concreto.

En un principio, se comenzó a desarrollar un CTV multilingüe para español, catalán y gallego. Estas lenguas comparten muchas características comunes (morfológicas, sintácticas, ...) puesto que todas ellas son lenguas romances con un mismo origen. Más tarde, se decidió incorporar el euskera. El hecho de que el euskera presente grandes diferencias morfológicas y sintácticas respecto a las lenguas previamente introducidas dio lugar a que aparecieran dificultades adicionales.

También existen otras limitaciones que no se deben al hecho de partir de un CTV determinado. Estas limitaciones son de tipo práctico y se puede decir que afectan a todos los idiomas por igual. El CTV multilingüe debe funcionar en tiempo real y sobre un hardware determinado (placas Antares, de la empresa DIALOGIC) que cuentan con una cierta cantidad de memoria y de velocidad de proceso.

El objetivo fundamental en la adaptación del CTV en español para convertirlo en un CTV multilingüe ha sido llegar a obtener un código ejecutable único, en el que las dependencias con el idioma aparecieran recogidas en distintas tablas o ficheros de configuración. Para hacer que el CTV funcione para un idioma determinado, basta con cargar y seleccionar las tablas correspondientes a ese idioma. Sólo en algunos casos ha sido necesario desarrollar algunas pequeñas partes de código que son dependientes del idioma.

Es bien sabido que la mayor parte de las tareas que es necesario realizar en un CTV exigen un conocimiento lingüístico del idioma bastante profundo, y saber emplear ese conocimiento lingüístico para construir procedimientos prácticos. Para la realización de las tareas de tipo lingüístico se ha contado con la intervención en este desarrollo de grupos de trabajo competentes en el uso de cada idioma y con experiencia en conversión texto-voz (Departamento de Filología Española de la Universidad Autónoma de Barcelona, Departamento de Tecnologías de las Comunicaciones de la Universidad de Vigo, y Departamento de Electrónica y Comunicaciones de la Universidad del País Vasco).

2. División en módulos

Como ya se ha indicado anteriormente, el CTV multilingüe parte de la estructura del conversor en español, que se describe en [Rodríguez93]. En esta estructura, se distinguen dos grandes bloques funcionales: bloque de proceso lingüístico y bloque de síntesis de voz.

El bloque de proceso lingüístico se compone de los siguientes módulos: Normalizador, Preproceso, Categorizador, Estructurador-pausador, Conversor grafema-alófono y Generador de parámetros prosódicos.

La síntesis de voz se realiza mediante la concatenación de unas unidades acústicas que previamente han sido diseñadas y grabadas, y que se encuentran recogidas en un inventario. Se dispone de dos modelos distintos para generar la voz sintética: un modelo LPC síncrono con el pitch [Olive90, Talkin90], y un modelo sinusoidal [Rodríguez96]. La concatenación debe realizarse de manera controlada para obtener el discurso deseado sin discontinuidades, ajustándose a las duraciones y contorno de frecuencia fundamental obtenidos por el módulo generador de parámetros prosódicos.

En los siguientes apartados se hace un repaso más detallado de los módulos del CTV.

3. Normalizador

Lo primero que hay que tener en consideración es escoger un juego de caracteres que contenga todos los símbolos necesarios para representar adecuadamente un texto escrito en

todos los idiomas. Se decidió adoptar el alfabeto ISO-Latin1, por contener todos los caracteres necesarios para los cuatro idiomas (incluyendo las distintas variedades de vocales acentuadas) y por ser un estándar de amplia difusión.

La tarea principal del normalizador es decidir dónde se acaba una frase. La frase detectada por el normalizador es la mayor unidad de trabajo (y de información) para el resto de los módulos del CTV. Esta tarea presenta numerosas ambigüedades y dificultades, pues no siempre se puede decir que determinado signo ortográfico marque final de frase.

Quizás el caso más complejo que se presenta es el del punto. Desde luego, un punto puede indicar fin de frase, pero también se emplea en otros muchos casos como son las abreviaturas ("ptas."), números ("10.423"), iniciales de nombres ("J.L. Serrano"), representación de la ele geminada en catalán ("col.legi"), ... Distinguir estos casos no es siempre fácil. Además, al considerar el euskera, se suman más casos conflictivos, pues el punto se emplea también en los ordinales ("1."), y las abreviaturas con punto pueden llevar añadida una desinencia de acuerdo al caso ("pta.ko").

La solución adoptada ha sido permitir que el normalizador sólo detecte el final de frase en los casos más claros, y que deje pasar los casos más ambiguos. Por tanto, en ocasiones, la frase detectada por el normalizador contendrá en realidad más de una frase. El preproceso, que hace un análisis más detallado de cada palabra, será el encargado de decidir después si los posibles finales de frase intermedios lo son o no.

4. Preproceso

Las tareas fundamentales de este módulo consisten en reducir a palabras cualquier elemento que aparezca en el texto (salvo los signos ortográficos), silabicar cada una de las palabras, y decidir su acentuación fonética.

- Expansión de formas. Se trata de expandir en palabras los números, abreviaturas, fechas... que aparezcan en el texto. Esta tarea presenta dos dificultades generales, que hemos encontrado en todos los idiomas:

Conflictos entre los formatos de distintas expresiones. Por ejemplo "3421789" puede ser el saldo de una cuenta bancaria, o un número de teléfono; "10/12" puede ser una fecha, una expresión aritmética o un código de producto; "C" puede ser un número romano, una clase de vitamina, o un grado de la administración pública. Hemos realizado métodos heurísticos para aventurar el tipo de interpretación en función del contexto. Sin embargo, estos métodos no son infalibles, y hemos decidido permitir que el usuario pueda decidir el tipo de interpretación, pues él sabe qué tipo de textos va a leer el CTV. Para ello ha sido necesario disponer mecanismos por los que el usuario del sistema de conversión pueda decidir qué clases de expresiones van a aparecer en el texto, y cómo deben leerse.

Aunque se haya identificado correctamente el tipo de expresión, a veces falta una norma o referencia que indique cómo debe pronunciarse (por ejemplo, en la expansión de una fecha, ¿cuándo se utiliza "de" y cuándo "del"?, o al tratar una sigla, ¿cuándo se deletrea, cuándo se expande, o cuándo se lee como una palabra cualquiera del idioma?). Estas decisiones quedan al buen juicio del desarrollador, pero este buen juicio no tiene porque ser bueno de acuerdo al juicio del usuario del sistema. Por eso, también en este sentido hemos intentado dar la máxima flexibilidad al usuario.

Además de estas dificultades, centrándonos en el aspecto multilingüe, aparecen otros problemas.

La primera labor que hubo que realizar fue localizar todas las dependencias del español presentes en este módulo, e intentar trasladarlas a tablas o a procedimientos más generales, que permitieran compartir el código entre los distintos idiomas. Así, por ejemplo, en la expansión de los números se eliminaron del código todas las referencias a formas de palabras, se pasaron a tablas, y se generalizaron los mecanismos de transcripción decenas-unidades para poder tratar tanto "42 -> cuarenta y dos" como "42 -> quaranta-dos".

En español, catalán y gallego es necesario cuidar la concordancia en la expansión de los números y las abreviaturas. En el caso de los números, éstos deben concordar con el sustantivo al que acompañan, si existe, y las abreviaturas deben expandirse en singular o plural, en función de las cantidades que las acompañen. Para realizar esta tarea de manera eficaz, es necesario retrasar la decisión sobre el género de

las expansiones de los números hasta después de haber realizado la categorización, para así poder localizar el sustantivo al que acompañan (si existe), e intentar aventurar su género.

En euskera, al no haber morfemas de género, y quedar el morfema de número englobado en la desinencia de la declinación, que suele aparecer escrita a continuación de la abreviatura, no se presentaba este problema. Sin embargo, ha resultado mucho más complicado identificar correctamente los elementos que necesitan expansión, por la misma presencia de la desinencia. Además de lo ya señalado en el módulo normalizador sobre la decisión de los puntos como posibles signos ortográficos, ha sido necesario considerar la desinencia para identificar correctamente siglas ("HBren"), fechas ("1983/03/07ko"), etc.

Además, se han presentado dificultades en el tratamiento de algunos de estos elementos, por decisiones tomadas previamente para el tratamiento de los otros idiomas. Por ejemplo, fue necesario flexibilizar el orden de los campos en las fechas, por ser distinto en euskera respecto a los otros tres idiomas. Asimismo, mientras que en español, catalán y gallego los números se leen basándose en las decenas, en euskera se leen basándose en las veintenas.

- Silabicación. La silabicación se utiliza fundamentalmente para decidir la acentuación fonética de las palabras, para decidir en algunos casos sobre la conversión grafema-alófono, y como ayuda para deletrear secuencias de letras que se consideran impronunciables. Se ha diseñado un sistema de reglas, basado en la información de la secuencia de letras [Mañas87], que permite realizar esta tarea. Presenta algunas limitaciones: no es capaz de recoger fácilmente criterios morfológicos (para silabicar "transatlántico", por ejemplo); se basa en la secuencia de letras y no de sonidos, por lo que si admitimos "[eksamen]" como transcripción de "examen", no es posible silabicar entre la oclusiva y la fricativa; y no tiene en cuenta fenómenos de contacto entre palabras (si fueran pertinentes). A pesar de estas limitaciones, el sistema es adecuado para las necesidades descritas al principio, y ha demostrado ser lo bastante flexible para ser aplicado en los cuatro idiomas, sin demasiado esfuerzo.
- Acentuación. Esta tarea determina si la palabra que se está tratando es átona o tónica, y en este caso, sobre qué vocal recae el acento. Pueden aparecer formas cuya acentuación depende de la categoría gramatical de la palabra (por ejemplo, "sobre" puede ser tónica o átona). También se identifican estos casos, en los que hay que esperar a haber realizado la categorización para poder deshacer la ambigüedad.

Para realizar esta tarea, los idiomas español, catalán y gallego responden a un conjunto de reglas bien definidas. Supuesta una silabicación correcta (para este propósito), la acentuación no presenta ningún problema.

Aunque conceptualmente es un problema distinto, más propio del módulo de transcripción grafema-alófono, hemos decidido asociar la determinación del timbre de las vocales al acento, por razones prácticas. En catalán, el timbre abierto o cerrado de las vocales ("e" y "o") sólo es pertinente en las vocales tónicas, pues las átonas se reducen. En gallego, aunque también puede haber distinción en las vocales pretónicas, se ha decidido limitar la distinción de timbre a las tónicas. En estas condiciones, resultaba muy apropiado ligar la determinación de timbre a la tarea de acentuación.

En catalán, mediante un sistema de rimas y excepciones (que ha sido ajustado a lo largo del tiempo), se puede determinar el timbre de la vocal en esta fase del proceso lingüístico. En gallego, dado el peculiar comportamiento de las vocales tónicas en los verbos, se ha decidido posponer la decisión sobre el timbre hasta después de la categorización, cuando se corrige la acentuación en casos ambiguos. Entonces, tras una decisión más o menos fiable sobre la categoría de la palabra, podemos tratar de manera diferente los verbos y el resto de las palabras.

La problemática del euskera es bastante distinta. El principal problema es la falta de una directriz clara (ni siquiera un consenso) sobre la acentuación. Tras llegar a una propuesta práctica sobre la acentuación, encontramos que los mecanismos de acentuación del CTV seguían siendo suficientes para abordar la incorporación del nuevo idioma. En euskera, las palabras se acentúan siguiendo una regla general (se acentúa la segunda sílaba de la palabra, o la primera si es monosilábica) y listas de excepciones. Luego, después de haber realizado la categorización, se corrige la acentuación en los casos conflictivos (verbos y algunas palabras función).

5. Categorizador

La tarea principal de este módulo consiste en asignar a cada palabra una categoría gramatical. La información de la categoría gramatical se utiliza fundamentalmente para decidir sobre la inserción de pausas y su caracterización. Además, las categorías se emplean para la corrección del acento, en la determinación del género en la expansión de los números, y en la transcripción grafema-alófono.

La tarea de categorización se ha realizado de manera completa para español, catalán y gallego. Para euskera, tan sólo se ha realizado lo imprescindible acentuar correctamente.

Esta tarea se compone de dos fases. En una primera se decide la categoría de la palabra a partir de su forma, basándose en listas de excepciones, de terminaciones y de raíces. A continuación, se resuelven las ambigüedades mediante reglas que tienen en cuenta el contexto en el que aparece la palabra (reglas de contexto).

Para la primera fase, se han desarrollado herramientas que, a partir de un diccionario, conjugan los verbos y generan las formas flexionadas. A continuación, se eligen automáticamente listas de terminaciones y excepciones a esas terminaciones, que de manera óptima permiten la clasificación de las palabras (dentro de las limitaciones de memoria del entorno de funcionamiento del CTV). Los pronombres enclíticos plantean un grave problema, pues es prácticamente imposible conjugar completo un verbo incluyendo todas las combinaciones de enclíticos, sobre todo en gallego (que además de poseer un número elevado de posibles enclíticos, permite combinarlos prácticamente con todas las formas verbales). Por ello, estas palabras reciben un tratamiento especial, en el que se intenta aislar primero el enclítico, antes de entrar en el tratamiento de desinencias y raíces.

En la segunda fase, mediante el mecanismo de reglas de contexto, se puede eliminar la ambigüedad que haya podido quedar de la anterior fase. Se ha hecho un importante esfuerzo en construir un sistema que facilita la utilización de estas reglas (incluso para personas que no han estado implicadas en su diseño, como se ha hecho en la obtención de las reglas de categorización y agrupación-pausado de catalán y gallego).

Al final de este módulo, con el mismo mecanismo de reglas de contexto, se realiza la corrección del acento que quedó pendiente en el módulo de preproceso.

```
# 12
INI_REGLA
INI_COND
 0 ¬ CATEG_FLAGS ¬ PPOS
 AND
 1 ¬ CATEG_TOTAL ¬ VERB ¬ NOMB
 AND
 (
  1 ¬ NO_CONCUERDA_NUMERO ¬ 0
  OR
  1 ¬ NO_CONCUERDA_GENERO ¬ 0
 )
INI_ACCI_SI
 0 ¬ CATEG_TOTAL ¬ PPOS
 1 ¬ CATEG_TOTAL ¬ VERB
 2 ¬ IR_A ¬ NO_MAS_REGLAS
INI_ACCI_NO
 0 ¬ IR_A ¬ SIGUIENTE
FIN_REGLA
```

Ejemplo de regla de contexto del módulo categorizador. "Si una palabra puede ser pronombre posesivo, y la palabra que le sigue es nombre o verbo, y no hay concordancia de género o de número entre ellas, entonces se decide que la primera palabra es pronombre posesivo, la segunda verbo y se pasa a la siguiente palabra. Si no, se pasa a la siguiente regla"

6. Estructurador-pausador

La tarea principal de este módulo consiste en la realización de pausas no marcadas ortográficamente. Además se caracterizan las pausas ortográficas y no ortográficas. Esta caracterización determina no sólo la duración de las pausas, sino también la evolución del contorno entonativo.

Al igual que sucede con el pausador, este módulo no ha sido plenamente realizado para el idioma euskera.

Para la selección de puntos donde realizar pausas no marcadas ortográficamente se utiliza el mismo mecanismo de reglas de contexto del que se ha hablado anteriormente. Con estas reglas se forman grupos de palabras en los que no está permitido realizar pausas (semejantes a sintagmas de una estructura sintáctica plana), y posteriormente se asigna un peso o probabilidad de realizar pausa a cada límite entre dos sintagmas. A continuación se elige el mejor punto para realizar la pausa, basándose en los pesos asignados y en criterios rítmicos. De momento, estos criterios son independientes del

idioma (es decir, se han tomado los que se venían utilizando en español).

Finalmente, se caracterizan las pausas, tanto las ortográficas como las introducidas por este módulo. Para esta caracterización se utiliza un conjunto de tipos de pausas posibles, que engloba las que se han considerado relevantes para la caracterización de la prosodia en español, catalán, y gallego (pues todavía no se trata el euskera).

7. Conversor grafema-alófono

Lo primero que hubo que hacer para el CTV multilingüe fue diseñar un conjunto de alófonos para los cuatro idiomas. Se amplió el conjunto de alófonos del CTV español, al que se incorporaron todos los alófonos adicionales necesarios. Siempre que fue posible, se empleó el alfabeto SAMPA para la representación de estos alófonos. El conjunto final incluye 46 alófonos, pero no todos los idiomas emplean todos los alófonos.

Se empleó un procedimiento de reglas para realizar la transcripción fonética. Cada idioma cuenta con unos ficheros de reglas propios, que son parte de las tablas que hay que cargar y seleccionar para que el CTV funcione en un idioma determinado.

La transcripción se hace partiendo de los caracteres silabicados y acentuados fonéticamente de las palabras de una frase. La acentuación fonética también indica el timbre adecuado de las vocales "e" y "o" en los idiomas catalán y gallego.

Sobre todo en el caso del catalán aparecen diversos procesos fonológicos de asimilación que dificultan la realización de la transcripción fonética de una manera secuencial, desde el principio hasta el final de la frase. Para solucionar este problema se ha dividido el proceso de transcripción fonética en dos fases. La primera fase trabaja recorriendo los caracteres silabicados y acentuados fonéticamente desde el principio hasta el final de la frase. En aquellos casos en que no se puede decidir el alófono concreto equivalente a un carácter (o caracteres), porque depende de alguna característica del alófono siguiente (que todavía no ha sido obtenido), se genera temporalmente un alófono que recoge la ambigüedad encontrada. La segunda fase de la transcripción recorre los resultados de la primera fase en orden inverso (desde el final hasta el principio) y va resolviendo las ambigüedades que quedaron pendientes.

También fue necesario permitir que las reglas de transcripción fueran capaces de consultar información de más alto nivel que el contexto de letras (incluyendo dentro de las "letras" las pausas, y las separaciones entre sílabas y entre palabras) en el que se encuentra determinada letra. Entre estas informaciones que fue necesario añadir se encuentran:

- Consultas sobre la palabra en la que aparece determinada letra (si es una palabra concreta, si tiene una cierta categoría gramatical, si es un clítico, si empieza por una secuencia de caracteres determinada, ...).
- El mismo tipo de consultas sobre otras palabras que anteceden o siguen a la palabra en la que se encuentra la letra.

Uno de los problemas que queda pendiente de resolver es el caso de las transcripciones fonéticas "forzadas". Por ejemplo, cuando nos encontramos con una palabra que en unas ocasiones habría que pronunciar de acuerdo a las reglas de un idioma, y que en otras ocasiones habría que pronunciar de acuerdo a las reglas de otro idioma. Este es un caso relativamente frecuente con los nombres y apellidos de personas. Por ejemplo, puede haber personas apellidadas "González" que vivan en Cataluña y que pronuncien su apellido de acuerdo a la transcripción "española" [gonT'aleT], mientras que otras lo pronuncian de acuerdo a la transcripción "catalana" [gunz'al@z] (ambas transcripciones están representadas con el alfabeto SAMPA adoptado en el CTV). De momento, la transcripción fonética siempre se hace de acuerdo a las reglas del idioma seleccionado en el CTV multilingüe.

8. Generador de parámetros prosódicos

La tarea de este módulo consiste en asignar duración a cada uno de los alófonos generados en el módulo de conversión grafema-alófono (incluidas las pausas), y un contorno entonativo a cada grupo fónico.

El modelo de duraciones es un modelo multiplicativo, y el modelo de entonación caracteriza cada grupo fónico como concatenación de 3 zonas: inicial hasta la primera sílaba tónica, zona central, y zona final desde la última tónica hasta la pausa. Estas zonas se caracterizan por

el número de sílabas tónicas del grupo y el tipo de pausa que lo finaliza.

Hasta el momento, sólo se dispone de los parámetros de los modelos de duración y entonación para el idioma español. Simplemente se han trasladado los valores de estos parámetros a los otros idiomas. Este es uno de los aspectos que supone mayor merma de naturalidad en los idiomas distintos del español.

Para poder ajustar los parámetros de los modelos es necesario disponer de un banco de datos de voz con información prosódica. Este banco de datos se obtiene seleccionando un conjunto de textos que cubran todos los factores de los modelos (así como otros factores no contemplados que se deseen validar) y grabando a un locutor. A continuación se añade información prosódica (se segmentan los sonidos, se extrae el contorno entonativo, y se enriquece el texto con información de acentos y pausas) y se generan los parámetros de los modelos por métodos estadísticos.

Actualmente se ha abordado la generación de dichos bancos de datos para catalán, gallego y euskera. Estos bancos de datos, junto con la automatización parcial del proceso de personalización de la prosodia, permitirán realizar la caracterización prosódica de estos idiomas.

9. Bloque de síntesis

La tarea de este módulo es generar la voz sintética a partir de la información de alófonos y prosodia, y del inventario de unidades.

Este módulo es totalmente independiente del idioma. Maneja el conjunto de alófonos común a todos los idiomas, y la particularidad de cada uno queda recogida en su inventario de unidades, una tabla que, como todas las tablas propias de un idioma, se puede cargar, y descargar y sustituir por otra de manera dinámica.

Los parámetros acústicos (dependientes del modelo de síntesis empleado) de cada alófono quedan recogidos en el inventario. Sin embargo, la caracterización sonoro/sordo del alófono y su tratamiento por el modelo de síntesis (en el caso del modelo LPC) se hace por código. Así, mientras que en español sólo se tenían alófonos sonoros o sordos, al incluir el catalán aparecieron sonidos fricativos sonoros, que precisaban una caracterización mixta en el modelo de síntesis LPC.

Por otra parte, las peculiaridades del conjunto de alófonos de cada idioma es un factor que hemos de tener en cuenta, aunque haya quedado recogido en una tabla ajena al código. Al aumentar el número de alófonos, y sobre todo el número de vocales (se consideran 5 vocales en español y euskera, 7 en gallego y 8 en catalán), aumenta de manera importante el tamaño de dicha tabla, y puesto que el sistema tiene que funcionar con unos recursos limitados de memoria, esta característica puede repercutir en una merma de la calidad acústica de los inventarios con mayor número de alófonos, al ser necesario restringir las combinaciones recogidas, o bien aplicar una codificación más fuerte para reducir el tamaño final del inventario.

10. Conclusiones y asuntos pendientes

En esta comunicación se ha presentado la descripción de un CTV multilingüe para los idiomas español, catalán, gallego y vasco.

Durante el desarrollo de este sistema hemos podido comprobar la importancia fundamental que tiene el conocimiento lingüístico de los diferentes idiomas en la conversión texto-voz. Tal es esta importancia, que puede llegar a influir de manera decisiva en el diseño de la arquitectura del conversor, y de las estructuras de datos manejadas. Esto nos confirma en la opinión de que la conversión texto-voz es una disciplina altamente dependiente de las características del idioma, y que no es posible alcanzar resultados de calidad sin contar con un conocimiento del mismo.

Tanto o más importante que el conocimiento del idioma es saber enfocar ese conocimiento con criterios prácticos. Normalmente es muy difícil capturar toda la riqueza que un idioma puede presentar, y hay que decidirse por opciones que pueden no ser exhaustivas o no ser correctas en todos los casos.

También se ha comprobado que es prácticamente imposible intentar abordar todos los problemas a la vez y darles solución. Nos parece que el enfoque más adecuado para el trabajo en conversión texto-voz es el de construir progresivamente, introduciendo mejoras y nuevas funciones de manera incremental, y haciendo un continuo proceso de revisión sobre las funciones ya existentes. Así es como se han incorporado nuevos idiomas al CTV.

Entre los principales asuntos pendientes que quedan en el conversor multilingüe, y que esperamos abordar en el futuro, se encuentran los siguientes:

- Mejoras en la prosodia. Actualmente sólo se ha trabajado con profundidad en la prosodia del idioma español. Las últimas tareas realizadas han consistido en diseñar unos modelos de duraciones y de entonación (F0) más completos que los anteriores, y ajustar los parámetros de estos modelos a la voz de un locutor determinado mediante un análisis estadístico. Se pretende obtener una prosodia personalizada para otros locutores en español, y para los locutores de los otros idiomas. Para ello se cuenta con unos bancos de datos con información prosódica de estos idiomas. En el caso del euskera será necesario completar previamente la categorización y el pausado, pues son requisitos previos para obtener la información manejada por los modelos de prosodia. Al ser el euskera una lengua declinada y con una sintaxis bastante diferente a la de español, catalán y gallego, se prevé que la adaptación de la categorización y el pausado supondrá un esfuerzo mayor.

  Los modelos actuales de prosodia no pueden darse por definitivos. Será necesario revisar estos modelos y enriquecerlos para que puedan manejar más información. Sería deseable profundizar más en la estructura sintáctica e incorporar información de tipo semántico y pragmático.
- Voz femenina. Hasta ahora, sólo se dispone de un inventario de unidades para voz femenina en español (locutor femenino) en el CTV multilingüe. No se dispone de ningún locutor femenino para los otros idiomas. La calidad de la voz sintética femenina obtenida es actualmente inferior a la de la voz sintética masculina.

  Un primer paso para mejorar la calidad de la voz sintética femenina es realizar el proceso de personalización de la prosodia, pues hasta ahora la prosodia femenina se obtiene simplemente mediante un escalado de los valores de la prosodia masculina.

  También es posible que sea necesario modificar los algoritmos de análisis, codificación y síntesis de voz de manera que permitan considerar y procesar mejor las características de la voz femenina, y así mejorar la calidad acústica de la síntesis.

11. Reconocimientos

En el desarrollo del CTV multilingüe han intervenido numerosas personas que no aparecen como autores de esta comunicación, pero a las que queremos expresar nuestro reconocimiento. Sin su colaboración no se habría podido realizar el CTV multilingüe.

Gracias a Luis Monzón, Sonia Herranz, Julia Giménez, Lourdes Aguilar, Montserrat Riera, Carme de la Mota, Juan María Garrido, Antonio Mestre, Jordi Renom, Eduardo Rodríguez, Javier Fernández, Elisa Fernández, Manuel González, e Inmaculada Hernáez.

12. Referencias

[Coile93] B. van Coile, "On the Development of Pronunciation Rules for Text-to-Speech Synthesis", Proceedings Eurospeech, 1993

[Lindström93] A. Lindström, "A Modular Architecture Supporting Multiple Hypotheses for Conversion of Text to Phonetics and Linguistic Entities", Proceedings Eurospeech, 1993

[Mañas87] J.A. Mañas, "Word Division in Spanish", Communications of the ACM, vol 30, nº 7 1987

[Olive90] J.P. Olive, "A New Algorythm for a Concatenative Speech Synthesis System Using an Augmented Acoustic Inventory os Speech Sounds", Proceedings of the ESCA Workshop on Speech Synthesis, 1990

[Rodríguez93] M.Á. Rodríguez y otros, "Amigo: un conversor texto-voz para español", Boletín nº 13 de la SEPLN, febrero 1993

[Rodríguez96] M.Á. Rodríguez y otros, "On the Use of a Sinusoidal Model for Speech Synthesis in Text-to-Speech", Progress in Speech Synthesis, Ed. Springer-Verlag, 1996

[Sproat94] R. Sproat, "A Modular Architecture for Multilingual Text-to-Speech", Conference Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis", 1994

[Talkin90] D. Talkin, ""Pitch-Synchronous Analysis and Synthesis for TTS Systems", Proceedings of the ESCA Workshop on Speech Synthesis, 1990