LA CONVERSIÓN DE TEXTO EN HABLA MULTIDOMINIO: PRINCIPIOS Y PORTABILIDAD

Francesc Alías y Joan Claudi Socoró

Departamento de Comunicaciones y Teoría de la Señal
Enginyeria i Arquitectura La Salle. Universitat Ramon Llull
Pg. Bonanova 8. 08022 Barcelona
{falias,jclaudi}@salle.url.edu

RESUMEN

La conversión de texto en habla (CTH) multidominio persigue conseguir una calidad sintética cercana a la de los sistemas de CTH diseñados para un determinado ámbito o aplicación, aumentando su flexibilidad al considerar distintos dominios (estilos de locución, emociones, temáticas, etc.) para la síntesis. En este trabajo, se presentan las motivaciones de esta estrategia desarrollada como evolución paralela a los sistemas orales multidominio, junto a distintas reflexiones sobre su flexibilidad y portabilidad para el diseño de nuevos sistemas de CTH a partir de las conclusiones obtenidas hasta el momento.

1. INTRODUCCIÓN

El propósito final de todo CTH es la generación de habla sintética completamente natural a partir de un texto de entrada cualquiera. Para lograr este objetivo, históricamente, la investigación en el ámbito de la CTH ha primado la capacidad del sistema de sintetizar *cualquier* mensaje sobre la naturalidad del mismo, es decir, la *flexibilidad* de la síntesis ante su *calidad* (*Aproximación A* en la figura 1) [1, 2]. Este enfoque se debe a que, ya desde sus inicios, los sistemas de síntesis fueron capaces de generar voz razonablemente inteligible a partir de una entrada de texto no restringida [2]. En el contexto de la CTH, este proceso se ha articulado, fundamentalmente, mediante el desarrollo de sistemas de conversión de texto en habla de propósito general (CTH-PG). Sin embargo, tiempo después, aparecieron nuevas aplicaciones de la CTH con un ámbito de funcionamiento controlado o restringido (p.ej. servicios de información meteorológica, de tráfico, etc.). En este contexto, se puede conseguir una elevada naturalidad de la señal sintética a cambio de reducir la generalidad de la síntesis, al utilizar textos pertenecientes al dominio. Esta filosofía constituye la segunda línea de investigación seguida en el camino hacia la consecución de una síntesis genérica *perfecta —Aproximación B* en la figura 1, introducida en [1] y posteriormente recogida en [2]. En este caso, el camino para llegar al objetivo marcado se articula, desde un punto de vista más práctico, a través de la consecución de productos de calidad notable

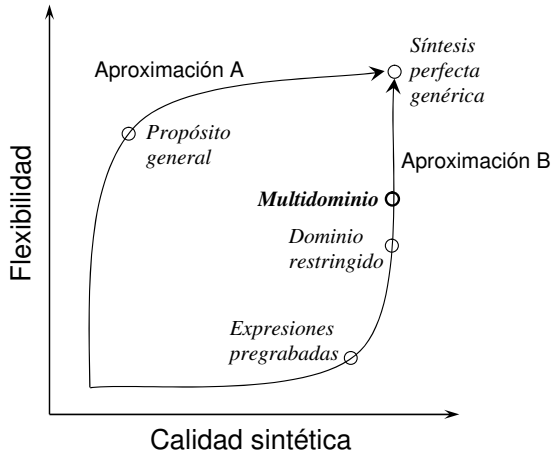

Figura 1. Aproximaciones seguidas en la investigación hacia la consecución de una CTH genérica perfecta, representadas según el binomio *calidad sintética - flexibilidad* del sistema — figura adaptada de [2].

(restringidos a aplicación) a medida que se avanza en la investigación —a diferencia de la *Aproximación A*, más teórica, donde, para obtener mejoras de calidad significativas, resulta necesario realizar grandes esfuerzos (p.ej. paso de la síntesis concatenativa basada en difonemas a síntesis basada en selección de unidades). Bajo esta segunda línea de investigación se ubican los CTH de dominio restringido (CTH-DR) (ver figura 1). Siguiendo esta segunda aproximación, se presentó la conversión de texto en habla multidominio (CTH-MD) en [3], con el objetivo de aumentar la flexibilidad de la síntesis a partir de la generalización de los resultados obtenidos mediante la síntesis de carácter restringido, manteniendo su calidad sintética. Aunque, inicialmente, esta aproximación sólo se diseñó para trabajar con subcorpus independientes bajo el marco de la CTH basada en corpus, posteriormente se observó que constituye un marco genérico para el desarrollo de cualquier tipo de sistema de CTH. En este trabajo se discute esta cuestión junto a las motivaciones de partida que dieron lugar al diseño de la nueva propuesta de filosofía de CTH desarrollada.

2. MOTIVACIÓN

En este apartado se describen los aspectos que han motivado el desarrollo de la CTH-MD respecto a la evolución de otros sistemas orales, así como las distintas propuestas de tipologías de corpus para síntesis existentes.

2.1. Sistemas orales multidominio

El desarrollo de sistemas multidominio es una de las nuevas líneas de investigación en el ámbito de los sistemas de lenguaje hablado (SLH) (o *spoken language systems*, en inglés) [4], entre los que destacan los sistemas de diálogo, los de traducción del habla y los de enrutamiento de llamadas [5]. La mayoría de los SLH, excluyendo los sistemas de dictado de propósito general, trabajan sobre un conjunto finito de dominios en los que el usuario puede hacer las consultas pertinentes a través del correspondiente sistema de reconocimiento automático del habla (RAH), p.ej. diferentes destinatarios en el enrutamiento de llamadas, distintas temáticas para los sistemas de traducción, o varios subdominios en los sistemas de diálogo complejos [6]. Conocer el dominio de la conversación permite mejorar el rendimiento y la eficacia de los módulos que conforman estos sistemas, por ejemplo, escogiendo el modelo del lenguaje más adecuado al dominio del reconocedor automático del habla, adaptando la estrategia de diálogo del gestor de diálogo si se produce un cambio de dominio dentro del discurso del usuario, o reduciendo los recursos utilizados por el SLH, al cargarlos dinámicamente según las necesidades particulares del diálogo en cada instante de la interacción (se adaptan al ámbito de la consulta o conversación) [4].

2.1.1. El RAH multidominio

En [7] se presentan dos aproximaciones distintas para la construcción de un reconocedor automático del habla multidominio (RAH-MD): *i*) trabajando con distintos RAH paralelos adaptados a dominio, seleccionando como resultado del reconocimiento aquél que aporte un mayor grado de confianza, o bien *ii*) se pueden combinar los datos de entrenamiento (léxicos, modelos de lenguaje, etc.) de los distintos reconocedores dependientes de dominio para construir una única red de búsqueda multidominio. Por un lado, la búsqueda paralela por dominio permite integrar distintos RAH ya existentes, una vez diseñados y optimizados para su dominio, restringiendo así el número de hipótesis por palabra a considerar. Sin embargo, la arquitectura en paralelo puede aumentar el coste computacional del proceso de reconocimiento del mensaje al realizar la búsqueda de la misma palabra en varias redes, complicándose la elección de la mejor hipótesis al tener que normalizar las medidas de confianza entre dominios. Por otro lado, el hecho de trabajar con una única red multidominio presenta las ventajas e inconvenientes complementarios a los de la arquitectura paralela, destacando el compromiso existente entre la reducción del coste computacional de la búsqueda y el aumento de la perplejidad en la resolución de la tarea del reconocimiento (ver [7] para más detalles).

Asimismo existe otra estrategia para el RAH-MD que consiste en dividir el proceso de reconocimiento del mensaje oral en dos fases: primero se aplica un sistema de reconocimiento *genérico* (independiente de dominio), y a continuación, se aplican los RAH entrenados sobre cada dominio (p.ej. ver [8], donde esta estrategia se emplea para sistemas de comprensión multidominio). En este contexto parece necesario incorporar un módulo de detección de dominio —generalmente, temático— para poder escoger, a partir del resultado del reconocedor genérico, el RAH dependiente de dominio más adecuado a la consulta del usuario [5].

2.1.2. Aproximaciones al corpus multidominio

Uno de los problemas fundamentales de los sistemas de CTH basados en selección de unidades (CTH-SU) es la pérdida de calidad sintética cuando el dominio del que proviene el texto de entrada no se ajusta al del corpus grabado [9], cuestión que perjudica la calidad tanto de los sistemas de CTH-PG [10, 11] como, evidentemente, la de los sistemas de dominio restringido [12]. En este contexto, se han presentado distintas aproximaciones con el objetivo de adaptar un CTH-PG a un determinado dominio, generalmente, mediante la incorporación de pequeños subcorpus dedicados a los dominios deseados [10, 11]. En estos trabajos se demuestra el efecto positivo que tiene la adecuación del corpus de voz al dominio deseado, con la consiguiente mejora de la calidad sintética obtenida.

Con este mismo objetivo, se define el concepto de *corpus de voz multidominio* [3, 13]. La idea consiste en disponer de unidades de voz correspondientes a distintos dominios (estilos de locución, temáticas, emociones, etc.) coexistiendo dentro de un mismo corpus de voz. En este ámbito, se han presentado dos estrategias distintas para el diseño de este tipo de corpus: *i)* mezclar todos los dominios en un único corpus con un contenido predominante de unidades genéricas [10, 14, 11], o *ii)* disponer de tantos subcorpus como dominios [15, 16, 17]. Estos dos enfoques —o tipologías de corpus multidominio— han sido denominados *blending* y *tiering*, respectivamente, en la literatura de síntesis basada en corpus [9]. Según las características de los estilos de locución a considerar será más conveniente utilizar una u otra tipología de corpus. Asimismo, se puede observar que los sistemas de CTH-MD siguen un enfoque similar al de los sistemas de RAH-MD.

3. ARQUITECTURA DEL SISTEMA DE CTH-MD

La investigación en el ámbito de la CTH no había incorporado, hasta nuestro trabajo, la filosofía multidominio incorporada en otros sistemas orales de interacción persona-máquina, en su camino hacia mejorar la naturali-

dad y la usabilidad de los mismos. Esta situación ha sido motivada, fundamentalmente, por dos cuestiones. Primero, el hecho de que los CTH fueron capaces de abordar la síntesis de propósito general desde sus inicios, a diferencia de los sistemas de RAH, que tuvieron que restringir ya de entrada el dominio de funcionamiento para ser eficientes; y segundo, debido al papel secundario que ha tenido la CTH en el contexto de los SLH multidominio, donde la CTH podría tomar en consideración el dominio de la conversación para adaptar el mensaje de salida del sistema (selección de dominio supervisada), pero que, en general, ha sido utilizada como simple medio de transmisión de la información consultada por el usuario (síntesis genérica), por lo que tampoco se ha abordado la CTH multidominio.

A grandes rasgos, la introducción de esta nueva estrategia de síntesis multidominio en los sistemas de CTH implica, básicamente:

- Disponer de una arquitectura de CTH flexible, que permita incorporar, por un lado, cualquier estrategia de síntesis, y por otro, seleccionar la estructura y el contendido del corpus de voz según las necesidades del sistema o aplicación en la que el CTH-MD se enmarque.
- Incorporar información del dominio del texto de entrada mediante un módulo de clasificación de dominio, con el objetivo de mejorar la flexibilidad de la síntesis, sin perder calidad de los mensajes sintéticos (equivalente a la de los CTH-DR).

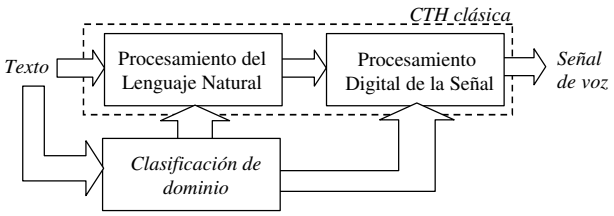

Figura 2. Diagrama de bloques de la arquitectura de un conversor de texto en habla multidominio con clasificación automática de dominio.

4. DESIGNACIÓN AUTOMÁTICA DE DOMINIO

La CTH-MD se asienta sobre la idea que la calidad de la CTH se puede mejorar si se conoce cuál es la *forma* más adecuada de pronunciar los textos a sintetizar [9] (p.ej. en [17] se escogen los textos que formarán parte del corpus de voz primando su facilidad para inducir la emoción deseada). Es decir, no todos los textos pueden ser pronunciados de cualquier forma (estilo, emoción, énfasis,...), ya que, por un lado, existen mensajes cuyo significado hace que sea inapropiado pronunciarlos de una determinada manera [18] (p.ej. órdenes militares *vs.* tristeza o miedo [15], mensajes conceptualmente complejos *vs.* voz de niño [9]), y por otro lado, existen mensajes que presentan una clara correlación con el modo de locución a utilizar [17] (p.ej. mensajes positivos o negativos *vs.* a unos patrones prosódicos determinados [14, 19], o frases más típicas de un niño o de una niña [9]). No obstante, no hay que olvidar que también existen mensajes que, según el contexto de la comunicación en el que se emitan, pueden cambiar de significado [16] (p.ej. "*Veo que hay mucha comida en la nevera*", en tono alegre o sarcástico). En este caso, escoger el modo de locución más apropiado para el mensaje pasa por disponer de información paralingüística (estado de humor del hablante, intencionalidad del mensaje, relación entre los interlocutores, ...), así como algunos parámetros extralinguísticos (edad, sexo, personalidad, ...), que también pueden afectar a la comunicación [20], cuestiones que quedan fuera del alcance del presente trabajo de investigación.

4.1. A partir de voz (en RAH-MD)

En el contexto de los sistemas multidominio guiados por voz, el módulo de detección y asignación automática de dominio (generalmente, una temática) a partir de las locuciones del usuario toma un papel relevante [4]. Conocer la temática de la consulta permite mejorar la precisión del sistema de reconocimiento, ya que permite escoger el modelo de lenguaje mejor adaptado a la tarea, con lo que se reduce tanto la perplejidad del modelo como la tasa de error de reconocimiento o WER [5].

La detección automática del dominio del mensaje en el ámbito de los SLH se realiza a partir de una selección de dominio *explícita*, mediante un conjunto predefinido de comandos (palabras clave), o bien, deduciendo los cambios de dominio de forma *implícita*, a partir de las hipótesis de reconocimiento recogidas durante la interacción del usuario con el sistema [7, 4, 5]. Cabe añadir que, a diferencia de los sistemas de detección de temática que trabajan con grandes volúmenes de datos (p.ej. artículos periodísticos, noticias o transcripciones de las mismas), en el contexto de interacción persona-máquina, normalmente se trabaja con un volumen mucho más reducido de datos [5].

4.2. A partir de texto (en CTH-MD)

De forma paralela, en el contexto de la CTH-MD también resulta necesario disponer de algún módulo que indique al CTH el dominio más adecuado sobre el que llevar a cabo el proceso de síntesis del texto de entrada. Este módulo puede ser externo al CTH, mediante una selección manual [16] o supervisada, como por ejemplo, en sistemas de diálogo —suelen trabajar con un conversor de concepto en habla (*concept-to-speech*, en inglés) [2]— o algunas aplicaciones audiovisuales con mensajes sintéticos controlados [15].

Asimismo, este proceso puede formar parte del sistema de CTH-MD mediante la inclusión de un módulo de clasificación automática de textos, como se propuso en [3]. Las particularidades de la tarea —textos cortos y bajo coste computacional— han guiado el desarrollo de

una nueva propuesta de sistema de clasificación de textos adaptada al problema de la CTH-MD (ver [21, 22] para más información). Fundamentalmente, la propuesta incorpora información temática y estilística de los textos (p.ej. las coocurrencias entre las palabras, los signos de puntuación, etc.) con el objetivo de modelar completamente el texto a sintetizar (típicamente de longitud muy corta, p.ej. 1 frase). Esta propuesta está claramente influenciada por la estrategia de síntesis utilizada, en este caso, la síntesis basada en corpus, siendo además el corpus multidominio implementado según la estrategia *tiering*. En el caso de trabajar con estrategias de síntesis más flexibles (p.ej. basadas en HMM), no resultaría tan crítico que el sistema de clasificación tomara en consideración todas y cada una de las palabras del texto.

5. PORTABILIDAD DE LA PROPUESTA

Aunque, inicialmente, la filosofía de CTH-MD se definió sólo pensando en el ámbito de la síntesis sobre distintos dominios independientes [3, 13], posteriormente se observó que constituye un marco genérico para el desarrollo de cualquier tipo de sistema de CTH. A continuación se presentan algunas ideas que describen la portabilidad de la propuesta más allá de lo presentado en los trabajos desarrollados hasta el momento.

5.1. Arquitectura flexible

La flexibilidad de esta arquitectura permite abordar desde la CTH-PG (un único dominio genérico), pasando por la CTH-DR (un único dominio restringido), hasta la CTH para distintos dominios (estructurados o no jerárquicamente, según su contenido y/o sus características acústicas). Estos dominios pueden ser incorporados explícitamente como subcorpus independientes (p.ej. [17, 9]), o como pequeños apéndices acompañando a un corpus genérico (p.ej. [10, 11, 14, 23]), así como, para dividir un mismo corpus (con la misma calidad vocal) en distintas temáticas (p.ej. periodísticas: *política, sociedad, cultura, deportes, . . .*, entre otros). Por otro lado, esta arquitectura también permite la coexistencia de filosofías de síntesis distintas, desde la síntesis basada en corpus (donde esta estructura se puede replicar y/o profundizar tanto como se quiera, siempre que el subdominio tenga suficientes unidades para la síntesis), pasando por la síntesis basada en modelos ocultos de Markov (con modelos adaptados a cada dominio), hasta soluciones híbridas (p.ej. síntesis genérica más subdominio adaptado a tarea, síntesis genérica más transformación de voz, entre otras). Simplemente, el CTH deberá tener las herramientas necesarias para abordar de forma eficiente la gestión de los datos con los que se diseñe.

Asimismo, por el momento, la designación de dominio provoca elegir un subcorpus u otro dentro del corpus multidominio (estrategia *tiering*). Sin embargo, como se acaba de comentar, la arquitectura diseñada permite disponer de corpus multidominio mixtos en los que se mezcle o complemente un corpus con un dominio u otro de forma que, para un determinado texto de entrada, el módulo de selección escoja la mejor secuencia de unidades considerando toda la información contenida en el corpus (estrategia *blending*). En este contexto, se puede realizar el proceso de selección de unidades mediante la incorporación de pesos relacionados con el grado de pertenencia del texto a sintetizar respecto a cada dominio —con una filosofía similar a la descrita en [14, 23]—, flexibilizando la preselección de las unidades que implica la selección de dominio según la estrategia *tiering*.

5.2. Aplicabilidad de la detección de dominio

Por el momento, la detección automática de dominio ha sido sólo utilizada para escoger el modelo prosódico y el subcorpus correspondientes al dominio identificado. No obstante, el hecho de poder determinar el dominio del texto de entrada permite mejorar la manera de sintetizarlo desde otros módulos del sistema de CTH, como por ejemplo: *i)* ayudando a desambiguar el mensaje en la fase de normalización del texto (p.ej. si se determina que el texto de entrada pertenece al dominio *matemático*, el texto $3/4$ se deberá convertir en "*tres cuartos*", evitando transcribirlo como "*tres de abril*"); *ii)* considerando varios perfiles prosódicos en la búsqueda (p.ej. como en [24]); *iii)* guiando el proceso de selección de unidades mediante los pesos pertinentes (en una estrategia de CTH-SU con un corpus genérico más varios subcorpus ad hoc a dominio, p.ej. [14]); *iv)* controlando las modificaciones a realizar por el módulo de procesamiento digital de la señal según el dominio del texto de entrada (p.ej. existen estilos de locución con calidades vocales particulares, donde grandes modificaciones de la señal pueden empeorar claramente la calidad sintética [25]); *v)* activando el módulo de transformación de voz para generar el estilo del dominio detectado (si el CTH dispone de un corpus genérico más un postprocesamiento mediante conversión de voz), etc. Según las características y particularidades del CTH sobre el que se incorpore el módulo de clasificación de dominio, éste tendrá un impacto mayor o menor en el funcionamiento del sistema de síntesis.

5.3. Optimización del proceso de selección

La propuesta de CTH-MD permite abordar el problema de la optimización del coste computacional de los sistemas de CTH basados en selección de unidades desde un punto de vista diferente al convencional en el contexto. Según la implementación de la CTH-MD *tiering*, al mismo tiempo que se escoge el dominio (subcorpus) más adecuado para realizar la síntesis, se consigue reducir el coste computacional del proceso de selección [3]. Esto es debido a que, una vez escogido uno de los subcorpus del corpus multidominio *tiering*, el conjunto de unidades sobre el que se realiza la búsqueda es de tamaño considerablemente menor respecto al total de unidades presentes

en el corpus —el coste computacional del proceso de clasificación es mínimo, como se demuestra en [21].

5.4. Enriquecimiento del análisis del texto de entrada

En el mismo camino en el que se encuentra la filosofía de CTH-MD descrita, han aparecido, recientemente, distintos trabajos en el ámbito de la investigación en tecnologías del habla que, mediante un análisis del texto más allá del típico para la CTH (del que se encarga el módulo de procesamiento del lenguaje natural —ver figura 2), pretenden dotar de mayor información al sistema de síntesis con el objetivo de mejorar la calidad sintética de salida. En este ámbito destacan los trabajos que pretenden estimar, a partir del texto a sintetizar, la actitud o postura del autor [19] o la emoción subyacente en el mensaje [26, 23], con el objetivo de mejorar la naturalidad de la síntesis. En [19] se demuestra la correlación entre las variaciones prosódicas y la aparición de adjetivos que expresan una actitud positiva o negativa del mensaje con mayor o menor intensidad —regulada mediante los adverbios que los acompañan. En [23] se diseña un corpus con tres emociones (neutra, alegre y enfado) a partir de 400 frases extraídas de textos periodísticos, agrupadas en un único corpus (estrategia *blending*). Mediante un diccionario (*Dictionary of Affect*) se determina el grado de emotividad de cada palabra y se ajusta la función de coste para realizar la búsqueda de unidades según el tipo y/o grado de emotividad de la palabra. Como conclusión del trabajo, se demuestra la viabilidad subjetiva de la propuesta y se indica que cuanto mayor sea el número de unidades emotivas presentes en la frase sintetizada, mayor percepción de esa emoción tendrá el usuario.

En el ámbito de los CTH dependientes de aplicación, destaca la investigación en el ámbito de la lectura de cuentos infantiles [26, 27]. En estos trabajos se pretende determinar la emoción más adecuada a cada pasaje del cuento a partir de las palabras y la estructura del texto (p.ej. longitud de la frase, análisis de las dependencias, puntuación de la frase, etc.), incorporando al análisis del texto conocimiento externo al mismo (p.ej. WordNet). Por otro lado, se encuentran [28] y [29], que incorporan una red semántica y una red de conocimiento genérico (*common sense*) al problema, respectivamente —ver [27], para más detalles sobre la detección de emociones a partir de texto.

6. DISCUSIÓN

No hay que olvidar que el objetivo final de la propuesta es obtener voz sintética de alta calidad. Hasta el momento, las pruebas realizadas (a nivel de frase) para estudiar la propuesta de CTH-MD se han sustentado en un elemento clave: la correlación entre el dominio del texto y el estilo de locución utilizado [30]. Debido a las características vocales de los dominios del corpus de voz utilizado (organizados según la estrategia *tiering)*, se ha escogido en tiempo de ejecución un dominio u otro para llevar a cabo la síntesis (evitando mezclar unidades en la síntesis con calidades vocales muy distintas), seleccionando la secuencia de unidades con menor número de concatenaciones a partir de una función de coste muy simple. Esta relación es fundamental para que la tarea del clasificador de textos propuesto seleccione automáticamente uno u otro estilo de locución a partir del texto de entrada. Por lo tanto, cabe diferenciar entre los textos que pueden mapearse directamente a un estilo de locución determinado, de los textos que no tienen una relación tan directa, como se ha comentado en la introducción del presente capítulo.

Tomando esta cuestión en consideración, en [30] se observa que cuando la CTH-MD funciona adecuadamente (es decir, su calidad es equivalente a la de un CTH-DR), los resultados obtenidos demuestran la mejora de la flexibilidad manteniendo una elevada calidad en la síntesis, gracias a adaptar el dominio de la selección de unidades —junto a la prosodia— al dominio del corpus al que mejor se ajusta el texto de entrada. Estos resultados se encuentran en la línea de otros trabajos, como los de [10, 11, 14], en los que la síntesis obtenida al adaptar la CTH al dominio del texto de entrada mejora claramente la obtenida a partir de sistemas de CTH-PG. Por otro lado, cuando el clasificador de textos asigna el texto de entrada a un dominio distinto al que fue grabado (error de clasificación), los resultados obtenidos muestran la dificultad de la clasificación de estas frases. Concretamente, los evaluadores muestran un criterio de selección mucho más vago que en el primer caso, ya que, en este caso, se está comparando el resultado de dos síntesis de dominio restringido, por lo que la selección se basa en lo apropiado que es el estilo al mensaje a transmitir —cuestión que pertenece más a un plano cognitivo que a un ámbito de calidad sintética.

En un futuro se pretende realizar un nuevo conjunto de pruebas que permitan comparar la calidad obtenida por la estrategia actual con la conseguida al trabajar con todas las unidades en un mismo corpus multidominio (estrategia *blending*), seleccionando las unidades incorporando información sobre el dominio en la función de coste considerada —al estilo de lo descrito en [23].

7. CONCLUSIONES

De algún modo, se puede argumentar que la CTH-MD permite dar un paso más en la convergencia de los sistemas de RAH y los de CTH, que se ha venido observando en la última década [31], en este caso en el desarrollo de sistemas orales multidominio. En este trabajo, una vez detalladas las motivaciones de la propuesta, se ha descrito cómo la filosofía de CTH-MD permite ajustar el sistema de CTH a las características de los datos disponibles, las necesidades del sistema de síntesis (orientado a aplicación, con varios dominios, con enfoque de propósito general más adaptación a un dominio, etc.) y las estrategias de síntesis más adecuadas, en lugar de tener que crear un nuevo sistema de CTH para cada nueva aplicación.

De este modo, la CTH-MD consigue integrar en un mismo CTH dominios que deben estar explícitamente representados en el corpus para ser reproducidos fidedignamente [9] (p.ej. determinadas emociones como la alegría, según [32] o estilos de locución con características vocales propias, según [25, 20]), junto a otros dominios que pueden ser generados sintéticamente de forma *bastante* realista (p.ej. tristeza desde un corpus neutro mediante modificaciones prosódicas [32], o incorporando pequeños subcorpus a un corpus de propósito general de tamaño mayor, p.ej. buenas o malas noticias [14]).

Por lo tanto, pensamos que la CTH-MD puede constituir un marco genérico para el desarrollo de futuras aplicaciones donde intervenga la síntesis de voz, incorporando o no (según el caso) la detección automática de dominio, adaptada a las características de los datos de trabajo.

8. BIBLIOGRAFÍA

[1] J. Yi y J. Glass, "Natural-sounding speech synthesis using variable-length units," in *Proceedings of ICSLP*, Sydney, Australia, 1998, pp. 1167–1170.

[2] P. Taylor, "Concept-to-Speech synthesis by phonological structure matching," *Philosophical Transactions of the Royal Society, Series A*, vol. 356, no. 1769, pp. 1403–1416, 2000.

[3] F. Alías, I. Iriondo, y P. Barnola, "Multi-domain text classification for unit selection Text-to-Speech Synthesis," in *The 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, 2003, pp. 2341–2344.

[4] K. Rüggenmann y I. Gurevych, "Assigning domains to speech recognition hypotheses," in *Proceedings of HLT-NAACL Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing*, Srinivas Bangalore y Hong-Kwang Jeff Kuo, Eds., Boston, Massachusetts, USA, 2004, pp. 70–77, ACL.

[5] I.R. Lane, T. Kawahara, T. Matsui, y S.Ñakamura, "Dialogue Speech Recognition by Combining Hierarchical Topic Classification and Language Model Switching," *IEICE Transactions on Information and Systems*, vol. E88D, no. 3, pp. 446–454, 2005.

[6] I.R. Lane, T. Kawahara, T. Matsui, y S.Ñakamura, "Out-of-domain detection based on confidence measures from multiple topic classification," in *Proceedings of ICASSP*, Montreal, Canadá, 2004, vol. 1, pp. 757–760.

[7] T.J. Hazen, I.L. Hetherington, y A. Park, "FST-Based Recognition Techniques for Multi-Lingual and Multi-Domain Spontaneous Speech," in *Proceedings of EuroSpeech*, Aalborg, Dinamarca, 2001, vol. 2, pp. 1591–1594.

[8] Grace Chung, *Towards multi-domain speech understanding with exible and dynamic vocabulary*, Ph.D. thesis, Massachusetts Institute of Technology, Junio 2001.

[9] A.W. Black, "Unit Selection and Emotional Speech," in *Proceedings of EuroSpeech*, Geneve, Suiza, 2003, pp. 1649–1652.

[10] M. Chu, C. Li, P. Hu, y E. Cahng, "Domain adaptation for TTS systems," in *Proceedings of ICASSP*, Orlando, USA, 2002, vol. 1, pp. 453–456.

[11] V. Fischer, J. Botella, y S. Kunzmann, "Domain Adaptation Methods in The IBM trainable Text-To-Speech System," in *Proceedings of ICSLP*, Jeju Island, Corea del Sur, 2004, pp. 1165–1168.

[12] W. Hamza y J.F. Pitrelli, "Combining the flexibility of speech synthesis with the naturalness of pre-recorded audio: a comparison of two approaches to phrase-splicing TTS," in *Proceedings of InterSpeech*, Lisboa, Portugal, 2005, pp. 2585–2588.

[13] F. Alías, X. Sevillano, P. Barnola, y J.C. Socoró, "Arquitectura para conversión texto-habla multidominio," *Procesamiento del Lenguaje Natural*, vol. 31, pp. 83–90, Septiembre 2003.

[14] W. Hamza, R. Bakis, E. M. Eide, M. A. Picheny, y J. F. Pitrelli, "The IBM Expressive Speech Synthesis System," in *Proceedings of of ICSLP*, Jeju Island, Corea del Sur, 2004, pp. 2577–2580.

[15] W.L. Johnson, S.Ñarayanan, R. Whitney, R. Das, M. Bulut, y C. LaBore, "Limited domain synthesis of expressive military speech for animated characters," in *Proceedings of IEEE Workshop on Speech Synthesis*, Santa Monica, USA, 2002, pp. 163 – 166.

[16] N. Campbell, "What type of inputs will we need for Expressive Speech Synthesis?," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, USA, 2002.

[17] A. Iida, N. Campbell, F. Higuchi, y M. Yasumura, "A corpus-based Speech Synthesis System with Emotion," *Speech Communication*, vol. 40, no. 1,2, pp. 161–187, 2003.

[18] J. Yamagishi, K. Onishi, T. Masuko, y T. Kobayashi, "Acoustic Modelling of Speaking Styles and Emotional Expressions in HMM-based Speech Synthesis," *IEICE Transactions on Information and Systems*, vol. E88D, no. 3, pp. 502–509, 2005.

[19] Y. Sagisaka, T. Yamashita, y Y. Kokenawa, "Generation and perception of F_0 markedness for communicative speech synthesis," *Speech Communication*, vol. 46, no. 1, pp. 376–384, 2005.

[20] N. Campbell, "Developments in Corpus-Based Speech Synthesis: Approaching Natural Conversational Speech," *IEICE Transactions on Information and Systems (Invited paper)*, vol. E88D, no. 3, pp. 376–383, Marzo 2005.

[21] F. Alías, X. Gonzalvo, X. Sevillano, J.C. Socoró, J.A. Montero, y D. García, "Clasificación de Textos Adaptada para Conversión de Texto en Habla Multidominio," *Procesamiento del Lenguaje Natural*, vol. 37, pp. 267–274, Setiembre 2006.

[22] F. Alías, X. Sevillano, y J.C. Socoró, "Text Classification based on Associative Relational Networks for Multi-Domain Text-to-Speech Synthesis," in *SIGIR-2006 Workshop on Stylistics for Text Retrieval in Practice*, Seattle, USA, Agosto 2006.

[23] G. Hofer, K. Richmond, y R.A.J. Clark, "Informed blending of databases for emotional speech synthesis," in *Proceedings of InterSpeech*, Lisboa, Portugal, 2005, pp. 501–504.

[24] F. Campillo y E. Rodríguez Banga, "Combined prosody and candidate unit selections for corpus-based Text-to-Speech systems," in *Proceedings of ICSLP*, Denver, USA, 2002, vol. 1, pp. 141–144.

[25] O. Turk, M. Schröder, B. Bozkurt, y L.M. Arslan, "Voice quality interpolation for emotional Text-to-Speech synthesis," in *Proceedings of InterSpeech*, Lisboa, Portugal, 2005, pp. 797–800.

[26] C. Ovesdotter, D. Roth, y R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of HLT/EMNLP*, Vancouver, Canadá, 2005, pp. 579–586.

[27] V. Francisco y P. Gervás, "Análisis de dependencias para la marcación de cuentos con emociones," *Procesamiento del Lenguaje Natural*, vol. 37, pp. 137–144, Setiembre 2006.

[28] Z.-J. Chuang y C.-H. Wu, "Emotion recognition from textual input using an emotional semantic network," in *Proceedings of ICSLP*, Denver, USA, 2002, pp. 2033–2036.

[29] H. Liu, H. Lieberman, y T. Selker, "A model of textual affect sensing using real world knowledge," in *Proceedings of ICSLP*, Miami, USA, 2003, p. 125132.

[30] F. Alías, J.C. Socoró, X. Sevillano, I. Iriondo, y X. Gonzalvo, "Multi-domain Text-to-Speech Synthesis by Automatic Text Classification," in *Proc. of InterSpeech*, Pittsburgh, USA, 2006.

[31] M. Ostendorf y I. Bulyko, "The impact of speech recognition on speech synthesis," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, USA, 2002.

[32] I. Iriondo, F. Alías, J. Melenchón, y M. A. Llorca, "Modeling and Synthesizing Emotional Speech for Catalan Text-to-Speech Synthesis," *Tutorial and Research Workshop on Affective Dialog Systems*.