Arquitectura para conversión texto-habla multidominio

Francesc Alías[*], Xavier Sevillano, Pere Barnola y Joan Claudi Socoró
Departamento de Comunicaciones y Teoría de la Señal
Enginyeria i Arquitectura La Salle, Universidad Ramon Llull
Pg. Bonanova 8, 08022 Barcelona
{falias, xavis, tm05122, jclaudi}@salleURL.edu

Resumen: este trabajo presenta una evolución en el diseño de la arquitectura para la conversión texto-habla multidominio (CTH-MD) basada en selección de unidades. Esta aproximación intenta conseguir una calidad sintética próxima a la de los sistemas de CTH de dominio limitado sin dejar de lado la síntesis de propósito general. La arquitectura multidominio implica disponer de un corpus de voz dividido en distintos dominios y estructurado jerárquicamente para optimizar el proceso de selección de unidades. En este trabajo, la jerarquización del corpus de voz se realiza mediante análisis en componentes independientes. Por otra parte, la CTH-MD necesita disponer de un módulo de clasificación de textos adaptado a sus necesidades.
Palabras clave: conversión texto-habla, síntesis multidominio, clasificación de textos, análisis en componentes independientes

Abstract: this paper presents an evolution in the design of multi-domain unit selection text-to-speech (MD-TTS) architecture. The goal of this approach is to achieve good synthetic quality like the limited domain TTS systems, without discarding a general purpose synthesis. The multi-domain architecture entails a speech corpus containing several domains. Moreover, it has been hierarchically structured by means of independent component analysis in order to optimize the unit selection process. On the other hand, MD-TTS demands a module capable to classify the texts in multiple domains, considering the system requirements.
Keywords: text-to-speech conversion, multi-domain synthesis, text classfication, independent component analysis

1. Introducción

En el ámbito de los sistemas de conversión texto-habla (CTH) basados en selección de unidades existen todavía muchos frentes abiertos para la obtención de una alta calidad de forma general (Black, 2002). Este trabajo pretende aportar un nuevo enfoque en el diseño de un sistema de síntesis de voz de máxima cobertura, flexibilidad y naturalidad: la conversión texto-habla multidominio (CTH-MD) (Alías y Llorà, 2003). Esta arquitectura permite realizar una síntesis de propósito general adaptada al dominio, reduciendo el espacio de búsqueda y logrando una calidad próxima a la de los sistemas de dominio limitado.

El sistema multidominio implica (i) redefinir la arquitectura y el contenido del corpus de voz y (ii) introducir un módulo que permita la selección del dominio más adecuado al texto a sintetizar (Breen y Jackson, 1998). El modelado de los textos y la herramienta utilizada para su clasificación deben ser capaces de codificar dos factores fundamentales de los que depende la calidad de la voz sintetizada en CTH: la continuidad y el estilo del texto.

En cuanto a la arquitectura del corpus multidominio, se presenta una estructuración jerárquica del mismo con el objetivo de optimizar el proceso de selección de unides. En este trabajo se ha optado por la aplicación de técnicas de análisis en componentes independientes, que agrupan los textos que componen el corpus en base a criterios de independencia estadística (Kaban y Girolami, 2000).

En la sección 2 del trabajo se presentan los fundamentos y la arquitectura del sistema de CTH-MD. En la sección 3 se describe el diseño del clasificador de textos. La sección 4 se dedica a la jerarquización del corpus. En la sección 5 se presentan los experimentos realizados sobre una colección de textos en catalán y, finalmente, se discuten las conclusiones del trabajo.

[*] Realizado con el apoyo del D.U.R.S.I. de la Generalitat de Catalunya mediante la beca 2000FI-00679

2. *Conversión texto-habla multidominio (CTH-MD)*

Este trabajo pretende aportar un punto de vista alternativo en el proceso de optimización de los sistemas CTH basados en selección de unidades (Black y Taylor, 1997). Estos sistemas convierten un texto de entrada en habla sintética, a partir de una secuencia de unidades sonoras extraídas de un corpus de voz pregrabado por un locutor en base a una colección de textos. En este trabajo se define una nueva arquitectura para el corpus de voz (o lo que es lo mismo, de la colección de textos en que se basa el corpus), que se sitúa en el espacio comprendido entre los sistemas de CTH de propósito general (CTH-PG) (Black y Taylor, 1997; Beutnagel et al., 1999) y los CTH para dominios limitados (CTH-DL) (Black y Lenzo, 2000; Montero et al., 2000).

Esta filosofía, denominada conversión texto-habla multidominio (CTH-MD) (Alías y Llorà, 2003), pretende conseguir una calidad sintética próxima a la de los sistemas de dominio limitado manteniendo la capacidad de generalización propia de los CTH-PG. Como se muestra en la figura 1, el hecho de definir esta nueva arquitectura para los sistemas de CTH implica introducir las siguientes modificaciones:

- Redefinir la estructura y el contenido de la base de datos.
- Introducir un nuevo módulo de clasificación de textos para diferentes dominios.

2.1. Corpus de voz multidominio

Normalmente, el diseño de un corpus de voz para CTH está estrechamente ligado al marco de trabajo del conversor de voz: (i) en CTH-PG, es el idioma de síntesis, ya que se pretende sintetizar *cualquier* texto, y (ii) en CTH-DL, es el dominio de funcionamiento (p.e., un reloj parlante (Black y Lenzo, 2000)). Por tanto, es la aplicación la que fija el planteamiento del corpus, lo que influye decisivamente en la calidad de la señal sintética obtenida. Así, la señal de voz sintética siempre reflejará el estilo y la cobertura del corpus (Breen y Jackson, 1998; Black, 2002).

Esta cuestión tiene una gran importancia en el funcionamiento de los sistemas de CTH-PG, ya que la calidad obtenida decrece de forma notable cuando el texto a sintetizar no

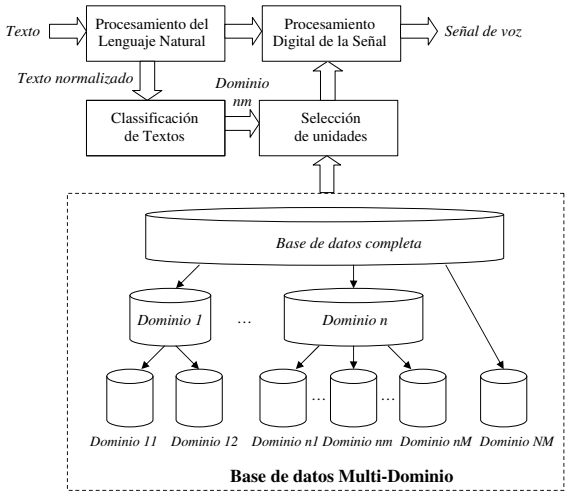

Figura 1: Diagrama de bloques de un sistema CTH-MD basado en selección de unidades. Donde n indica el índice del dominio y m el del subdomino.

se ajusta al estilo de la base de datos que se ha grabado (Black y Lenzo, 2000; Chu et al., 2002). Por otra parte, si se pretende incorporar varios dominios en un mismo corpus, el vocabulario contenido en cada uno de ellos debe tener una cobertura fonética y prosódica suficiente para el dominio tratado (Montero et al., 2000), minimizando el efecto de las palabras de fuera del vocabulario presente en sintetizadores basados en la concatenación de palabras (Möbius, 2001). A su vez, el módulo de selección de unidades deberá ser capaz de escoger entre los estilos contemplados (Breen y Jackson, 1998).

Esta problemática se aborda mediante la arquitectura multidominio que se presenta en este trabajo. Esta aproximación estructura jerárquicamente el corpus de voz en distintos dominios (ver figura 1) y los selecciona mediante un clasificador de textos. Cabe destacar que la estructura jerárquica del corpus de voz grabado, así como el número de niveles de la misma, dependerán del contenido de la colección de textos utilizados en su diseño.

2.2. Módulo de clasificación de textos para CTH-MD

Tal y como se desprende de la figura 1, el hecho de estructurar el corpus de voz como un conjunto de *subcorpus* de diferentes dominios, implica introducir un módulo que, a partir del texto de entrada, seleccione cuál es el dominio o conjunto de dominios sobre los

que se debe realizar la búsqueda de la cadena de unidades sonoras para la síntesis (selección de unidades).

Esta necesidad se resuelve mediante la integración de un algoritmo de clasificación de textos (CT) en la arquitectura del conversor texto-habla. Entonces, dada una colección de documentos $\mathcal{D}$ y un conjunto de categorías $\mathcal{C}$, el algoritmo de CT asigna cada documento a la categoría (clasificación binaria) o grupo de categorías (*ranking* de clasificación) que más se adecúen al contenido del documento (Sebastiani, 2002).

Los dos características esenciales del algoritmo de CT deberán ser:

1. Adaptación al contexto y a las necesidades de la síntesis multidominio.
2. Optimización y simplicidad del diseño, con el fin de no ralentizar la síntesis.

A continuación se describe el diseño del módulo clasificador de textos, que adapta el modelado del texto y las medidas de semejanza escogidas a las necesidades de un CTH-MD. En este punto es importante destacar que el objetivo del algoritmo implementado no es tanto el de ser un excelente clasificador de documentos, sino que sólo pretende ser un elemento más de mejora de la calidad de síntesis dentro de la nueva arquitectura multidominio definida.

3. Diseño del clasificador de textos

La mayoría de sistemas de CT tratan el texto como una simple colección de palabras aisladas (bolsa de palabras o *bag-of-words*). De esta forma, cada texto se ve reducido sólo a sus términos constituyentes, ignorando su orden o las relaciones que puedan existir entre ellos. Por lo tanto, esta aproximación deja de codificar el efecto de la sinonimia y el de la polisemia. A pesar de estos efectos negativos, esta representación es una de las más utilizadas en CT, debido a que los algoritmos que se basan en ella han sido optimizados durante décadas (Sebastiani, 2002).

Por otro lado, en el contexto de la síntesis de voz, la aproximación de *bag-of-words* no permite representar dos de los factores fundamentales de los que depende la calidad de la voz generada:

- **Continuidad**: la señal de voz está constituida por una secuencia de unidades sonoras consecutivas (fonemas, difonemas, etc.) extraídas del corpus de voz. Si el texto se representa mediante palabras aisladas, se pierde su secuencialidad (p.e., se deja de considerar el efecto de la coarticulación entre palabras).
- **Estilo**: es una de las características fundamentales para la aproximación de CTH-MD, que pretende ajustar el estilo del texto de entrada al del dominio escogido. Esto se consigue al considerar la información contenida en las relaciones entre las palabras y su secuencialidad.

En este contexto es imprescindible hallar una estrategia que permita codificar estas dos informaciones. En este trabajo se ha escogido representar la información contenida en cada uno de los dominios mediante una *red relacional asociativa* (RRA), definida inicialmente para los sistemas de representación visual de documentos (Rennison, 1994). Este modelo se define como un *grafo* ponderado de nodos interconectados (ver figura 2) con tantos nodos como palabras a modelar, enlazados entre sí mediante conexiones ponderadas.

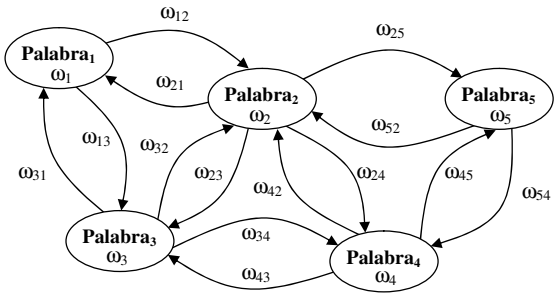

Figura 2: Ejemplo de una red relacional asociativa de palabras.

Entonces, tal y como se presenta en la figura 2, esta red contendrá:

- Información relacionada con cada una de las palabras del texto (ω_i).
- Información de las relaciones entre cada par de palabras del texto (ω_{ij}).
- El sentido de las conexiones entre las palabras a sintetizar ($\omega_{ij} \neq \omega_{ji}$).

3.1. Modelado del texto

Debido a la necesidad de contar con toda la información de la frase o párrafo a sintetizar, el sistema de clasificación de textos des-

arrollado no incluye ninguno de los preprocesamientos habituales en CT (eliminación de palabras vacías o *stop-listing* y lematización o *stemming*) (Sebastiani,2002). Simplemente, parte del texto normalizado procedente del módulo de procesamiento del lenguaje natural del CTH, que transforma en grafemas los números, acrónimos, etc. del texto de entrada (ver figura 1).

3.1.1. Parametrización del texto

En cuanto a la parametrización del texto de entrada al CT, existen diversas aproximaciones como la booleana o la probabilística, entre otras (Aas y Eikvil, 1999). Sin embargo, todas ellas se suelen basar en el modelo de *bag-of-words*, con los inconvenientes que ello supone para CTH, ya que no parametriza la continuidad y estilo del texto. Así, un CT en el contexto de la síntesis del habla debe considerar parámetros que incluyan esta información, como son:

Frecuencia de los términos: número de apariciones de cada palabra en el texto (*term frequency*, tf). Es uno de los parámetros básicos para cualquier CT y se incluye en cada uno de los nodos de la RRA (figura 2), para ponderar la importancia de aquella palabra en el texto.

Frecuencia de coocurrencia: número de veces que dos palabras aparecen consecutivamente en el texto (*co-occurrence frequency*, cof). Este parámetro codifica la importancia de las relaciones entre las palabras (Rennison, 1994). En la RRA se incluye como el peso entre las conexiones de los nodos del *grafo* (figura 2).

Longitud del patrón: número de palabras consecutivas del texto a sintetizar que aparecen de forma sucesiva en la RRA de un dominio (*pattern length*, pl), normalizado respecto al número total de palabras del texto de entrada $[0 \leq \mathrm{pl} \leq 1]$. Parametriza el grado de pertenencia del texto de entrada respecto a cada uno de los dominios. De esta forma se evalúa la secuencialidad de las palabras, a diferencia de Rennison (1994), donde sólo se considera el número de veces que las dos palabras aparecen simultáneamente en el documento.

3.1.2. Representación del texto

Toda la información obtenida de la parametrización del texto debe representarse de algún modo. En este trabajo se ha optado por seguir el modelo de espacio vectorial (MEV) (Salton, 1989), como en muchos otros sistemas de CT. En este modelo, los textos son representados como vectores en un espacio multidimensional. Por tanto, debe definirse un vector que contenga el $\mathrm{tf_i}$ de cada palabra y su relación con el resto de palabras del texto mediante $\mathrm{cof_{ij}}$ (ecuación 1).

$$\vec{v} = (\mathrm{tf}_1, \mathrm{cof}_{11}, .., \mathrm{tf_i}, \mathrm{cof}_{i1}, .., \mathrm{cof_{ij}}, ..) \qquad (1)$$

En esta representación vectorial ($\vec{v}$) no se incorpora el parámetro pl. Esto es debido a que éste aporta información global del grado de semejanza entre las RRA que modelan el texto de entrada y cada uno de los dominios. De todos modos, este parámetro deberá formar parte de las medidas de similitud que se pasan a describir a continuación.

3.2. Medidas de similitud

Es necesario definir una función que evalúe el grado de pertenencia del texto de entrada a sintetizar (t) a cada uno de los modelos de los dominios (D). Por el hecho de trabajar con el MEV, la primera medida utilizada fue la *distancia del coseno* (Salton, 1989), pero los resultados obtenidos no fueron muy satisfactorios (Alías y Llorà, 2003). A partir de esta distancia, se define una nueva medida (ecuación 2) que incluye la longitud del patrón (pl) ponderando la distancia del coseno. De esta forma, se considera la información lingüística del texto codificada por este parámetro.

$$S_1(t, D) = \mathrm{pl} \cdot cos(t, D) \qquad (2)$$

A partir de estas dos medidas de similitud se estudió otro tipo de distancia, distinta de las utilizadas en las aproximaciones clásicas de CT, basada en una función *sigmoidea* (la función logística), utilizada como función de activación en redes neuronales (ecuación 3).

$$\mathrm{sigm}(x) = \frac{1}{1 + e^{-\left(\frac{x-\mu}{\sigma}\right)}} \qquad (3)$$

donde μ indica el umbral de activación de la función y σ define la pendiente de la misma. En (Alías y Llorà, 2003) se describe su ajuste en el marco de CT para CTH-MD. Se trata de una función de recorrido $[0, 1]$, lo que resuelve el problema de la normalización de la medida de similitud.

A partir de esta función, se diseña una medida de evaluación (ecuación 4) como la

media geométrica de las diferencias entre los parámetros tf y cof de t y D según la sigmoidea, y la longitud del patrón respecto al modelo (pl).

$$S_2(t,D) = \Bigg(\mathrm{pl} \cdot \mathrm{sigm}\Big(\sum_i (\mathrm{tf}_i^t - \mathrm{tf}_i^D) \Big) \cdot \mathrm{sigm}\Big(\sum_{i,j} (\mathrm{cof}_{ij}^t - \mathrm{cof}_{ij}^D) \Big) \Bigg)^{1/3} \quad (4)$$

4. Generación de los modelos de los dominios

Los dominios del corpus de voz se generan a partir de la colección de textos sobre la que se construye el corpus de voz, siguiendo un proceso supervisado o no supervisado. El primero obtendrá los modelos a partir de los textos etiquetados manualmente, es decir, asignados al dominio correspondiente por un experto. En cambio, el segundo generará los dominios de forma no supervisada, buscando las categorías de clasificación, o sea los dominios, a partir de los textos de la colección (algoritmo de agrupamiento o *clustering*).

Una vez obtenidos estos modelos básicos (D_{nm} en la 1) se puede proceder a la agrupación de los dominios a un nivel superior. A partir de un cierto grado de similitud entre los dominios básicos, se desarrolla un proceso de asociación de aquellos dominios que se consideren semejantes, obteniendo un agrupamiento jerárquico de dominios (D_n en la figura 1). Del mismo modo, se puede proceder a una subdivisión de los dominios D_{nm} en subdominios. Este proceso permitirá obtener un corpus de voz multidominio estructurado jerárquicamente (ver figura 1). En el nivel superior de esta estructura se halla la base de datos completa, lo que aproximadamente correspondería a un corpus para un CTH-PG. Las propiedades más importantes de un corpus jerarquizado son:

1. Disponibilidad de niveles intermedios de clasificación.
2. Estructuración del contenido del corpus.
3. Arquitectura modular y configurable.

Gracias a esta estructura, cuando el CT no halla suficiente similitud entre el texto t y los modelos básicos o bien cuando t presenta grados de pertenencia muy similares respecto varios modelos, se puede repetir la búsqueda en un nivel superior de la estructura multidominio. De este modo, se optimiza el funcionamiento del módulo de selección de unidades, ya que la búsqueda se realiza sobre el región del corpus más adecuada.

A continuación se describe la técnica utilizada para la generación automática de la estructura jerárquica del corpus de voz. Una vez finalizado este proceso, cada grupo (o *cluster*) de textos se modelará mediante una RRA.

4.1. Jerarquización mediante ICA

El análisis en componentes independientes (*Independent Component Analysis* o ICA) es una técnica estadística de propósito general fundamentada en un modelo generativo de variables latentes (Hyvärinen y Oja, 2001). El modelo ICA representa n variables aleatorias observadas como una combinación lineal de n variables ocultas, y asume la independencia estadística de estas últimas, denominadas componentes independientes (CI). Así pues, los algoritmos ICA hallarán las CI, así como los coeficientes de su combinación lineal, partiendo únicamente de las observaciones disponibles.

4.1.1. ICA aplicado a CT

En el ámbito de la clasificación de textos, la aplicación de ICA se basa en la asunción de un modelo generativo de documentos como combinación de *ámbitos semánticos* (Isbell y Viola, 1999; Kaban y Girolami, 2000). Es decir, un documento se debe a la interacción de un conjunto de variables ocultas independientes que lo generan. Estadísticamente, un ámbito semántico es una distribución probabilística centrada en los términos (o palabras) más usuales de la temática tratada.

El uso de ICA como CT está muy vinculado al *análisis de semántica latente* (ASL) (Deerwester et al., 1990). Esta técnica proyecta los documentos en un espacio ortogonal de dimensionalidad reducida, extrayendo las K direcciones principales de los datos. La aplicación de ICA sobre el espacio ASL permite descubrir los K temáticas independientes que generaron los documentos, lo que permite su clasificación (Kaban y Girolami, 2000).

4.1.2. Descripción del método

Un asunto clave en la aplicación de ICA como CT es la elección de la dimensionalidad del problema (K). Según la relación entre K y el número de categorías, $|\mathcal{C}|$, en que un experto ha dividido la colección de forma manual, el algoritmo ICA aporta distintas soluciones:

- Si $K = |\mathcal{C}|$, se obtienen tantos *clusters* como categorías, clasificando los documentos en los ámbitos semánticos correspondientes a las categorías.
- Si $K < |\mathcal{C}|$, se produce una agrupación de categorías en macrocategorías. Estos grupos de jerarquía superior contienen las categorías más dependientes estadísticamente entre sí.
- Si $K > |\mathcal{C}|$, los textos pertenecientes a cada una de las categorías se separan en subcategorías. Conviene resaltar que el número de temáticas presentes en una colección de documentos puede ser mayor que el considerado por un etiquetador experto (Kaban y Girolami, 2000). Es decir, dentro de una temática principal pueden existir diversas subtemáticas.

Estas ideas entroncan con el concepto de corpus de voz multidominio de un CTH-MD, permitiendo la estructuración jerárquica de los textos que lo componen.

Es lógico pensar que cuanto más homogénea sea la temática de un determinado dominio del corpus, sus textos tenderán a agruparse en un único grupo, y éste será descubierto para valores pequeños de K. En cambio, para un dominio de temática más *variada*, los textos que contiene se distribuirán en un mayor número de *clusters*, que sólo podrán ser hallados cuando se incremente la dimensionalidad del problema.

Para determinar la estructura jerárquica del corpus se aplica sucesivamente el algoritmo ICA para valores crecientes de K, analizando la agrupación de los textos pertenecientes a cada dominio.

Se ha establecido experimentalmente que los *clusters* significativos de cada dominio son aquellos que contengan, como mínimo, la mitad del número de textos asignados al grupo predominante. Por último, el número óptimo de *clusters* ($\mathrm{NC_{opt}}$) será aquel que aporte una mayor tasa de clasificación para cada dominio.

5. Experimentos

A continuación se analiza el funcionamiento del CT diseñado para CTH-MD, así como la estructura jerárquica del corpus obtenida. Los experimentos se han llevado a cabo sobre una colección de artículos extraídos del periódico catalán AVUI, recopilados durante distintos periodos de los años 2000 y 2003 (Alías y Llorà, 2003). Este corpus está formado por 200 documentos (2600 términos) divididos en 4 dominios: D={*política*(60 documentos), *sociedad*(60), *música*(40), *teatro*(40)}.

A continuación se presentan los experimentos desarrollados. En el primero, se analiza la capacidad de la RRA como CT en el contexto de la CTH-MD. En el segundo, se presentan los resultados obtenidos de la jerarquización del corpus mediante ICA.

5.1. Clasificación mediante RRA

En el cuadro 1 se muestran las tasas de clasificación obtenidas por el CT diseñado con las dos distancias de similitud estudiadas para distintos porcentajes de test (sobre el total de documentos). Como se puede apreciar las dos medidas consiguen tasas de clasificación idénticas excepto para un 25 % de test, donde S_1 presenta un comportamiento no monótono y S_2 un valor muy bajo, debido a que el proceso de entrenamiento parte de pocos textos lo que provoca un pobre modelado de las RRA de los dominios.

% Test	10	15	20	25
S_1	.695	.636	.628	.654
S_2	.695	.636	.628	.577

Cuadro 1: Tasas de clasificación con RRA según la medida de similitud.

Para demostrar la necesidad de la adaptación del sistema de CT al entorno de la CTH, donde no se dispone de grandes corpus de textos, se ha analizado el comportamiento de uno de los algoritmos de CT de mejores prestaciones: *Support Vector Machines* (SVM) (Sebastiani, 2002). Para ello se ha utilizado el programa SVM^{light} de Joachims (2000), para clasificar el colecciones de textos en catalán del que se dispone, sin ningún preprocesamiento (ni *stop-list* ni *stemming*).

Este sistema de CT presenta un bajo rendimiento al trabajar con un número reducido de textos, tanto para el entrenamiento co-

mo para el *test* (Tang, 2001). Además, en el mismo trabajo se demuestra que la falta de preprocesamiento también empeora las tasas de acierto. En el corpus estudiado se producen ambas situaciones, lo que provoca un pobre rendimiento de clasificación de SVM para CTH-MD.

5.2. Jerarquización del corpus mediante ICA

Este experimento analiza la capacidad del algoritmo ICA para jerarquizar los documentos del corpus, aplicándolo para distintos valores de K. Se ha utilizado un algoritmo de punto fijo que maximiza el momento de tercer orden de los datos (Kaban y Girolami, 2000). Como la colección se divide en cuatro dominios, la primera prueba consiste en hallar $K = 4$ *clusters*, resultando las tasas de clasificación del cuadro 2.

Dominio	% clasificación
Política	70.5
Sociedad	75.4
Música	93.2
Teatro	100

Cuadro 2: Tasas de clasificación con $K = 4$.

El siguiente paso del experimento analiza la capacidad de clasificación y agrupamiento del algoritmo conforme aumenta la dimensionalidad del problema. En la figura 3 se muestra la evolución de las tasas de clasificación de los textos de cada dominio para diversos valores de $K > 4$. Como puede apreciarse, la tasa de clasificación máxima en el dominio de *política* se logra para $K = 7$ y es del 96,7 %, mejorando notablemente los resultados obtenidos para este dominio con $K = 4$. El número de *clusters* óptimo en que se agrupan sus textos es $\mathrm{NC}_{\mathrm{opt}} = 4$.

El dominio de *sociedad* se clasifica con un acierto del 75,4 % cuando $K = 4$ ($\mathrm{NC}_{\mathrm{opt}} = 1$) y $K = 10$ ($\mathrm{NC}_{\mathrm{opt}} = 3$). A igual tasa de clasificación se prefiere considerar tres grupos, pues ésto permitirá un mejor modelado de los textos, al aplicarse una RRA para cada uno.

Los textos de *música* y *teatro* se agrupan en un solo *cluster* aunque la dimensionalidad del problema crezca. Esto es lógico si pensamos que se trata de textos de una temática homogénea y muy específica. Las máximas tasas de clasificación para estos dominios se

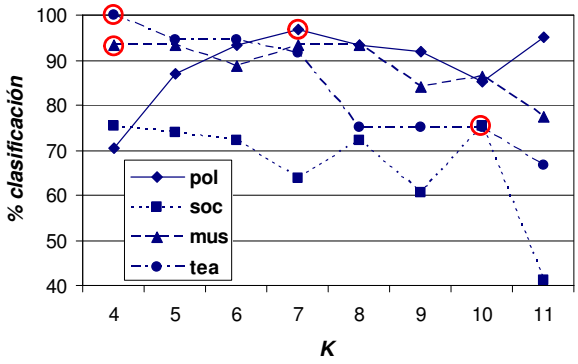

Figura 3: Tasas de clasificación con ICA y $K \in [4, 11]$. $\mathrm{NC}_{\mathrm{opt}}$ indicados con un círculo.

obtienen para $K = 4$ y son las que se muestran en el cuadro 2. El experimento finaliza aplicando el algoritmo ICA con valores de $K < 4$ para generar niveles de jerarquía superior (macrodominios y superdominios). Tal y como se muestra en la figura 4, el análisis para $K = 3$ muestra un agrupamiento de los textos de *música* y *teatro* en un macrodominio que podríamos denominar *cultura*, mientras para $K = 2$ los textos de *política* y *sociedad* forman un superdominio propio.

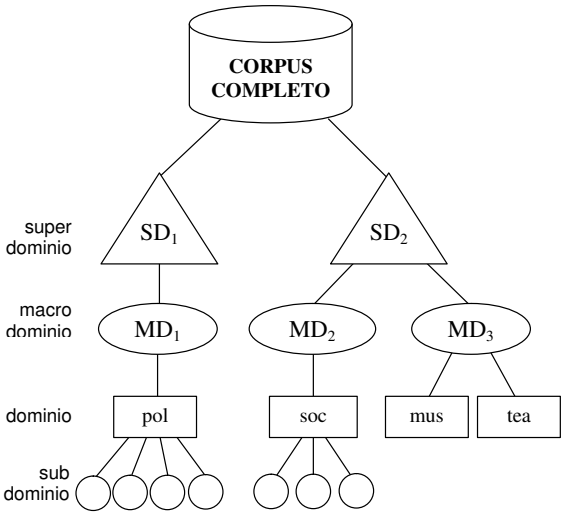

Figura 4: Estructura jerárquica de la colección de textos obtenida mediante ICA.

6. *Conclusiones y trabajo futuro*

Este trabajo surge como continuación del trabajo presentado en (Alías y Llorà, 2003) donde se presentó un nuevo enfoque para la síntesis de voz: la conversión texto-habla mutidominio (CTH-MD). En esta comunicación se

han descrito dos elementos clave de la arquitectura para CTH-MD (i) un módulo de clasificación de textos basado en redes relacionales asociativas que modelan los dominios y los textos de entrada y (ii) un método, basado en el análisis en componentes independientes, para estructurar jerárquicamente la colección de textos que dará lugar al corpus de voz del CTH-MD.

Los experimentos desarrollados prueban la viabilidad del CT basado en RRA en el contexto de la CTH-MD frente a aproximaciones clásicas, como SVM. Además, se ha comprobado que ICA es una herramienta útil para la estructuración jerárquica del corpus.

Como principal línea de futuro se plantea la integración de estos bloques en un sistema CTH-MD real, aplicando el modelado por RRA sobre la estructura jerárquica obtenida.

Bibliografía

Aas, K. y L. Eikvil. 1999. Text categorisation: A survey. Technical Report Nr. 941, (ISBN 82-539-0425-8), Norwegian Computing Center.

Alías, F. y X. Llorà. 2003. Evolutionary weight tuning based on diphone pairs for unit selection speech synthesis. En *EuroSpeech*, a celebrarse en Geneve, Switzerland.

Beutnagel, M., A. Conkie, J. Schroeter, Y. Stylianou, y A. Syrdal. 1999. The AT&T Next-Gen TTS system. En *Joint Meeting of ASA, EAA, and DAGA2*, páginas 18–24, Berlin, Germany.

Black, A.W. 2002. Perfect Synthesis for all of the people all of the time. En *IEEE TTS Workshop 2002 (Keynote)*, Santa Monica, USA.

Black, A.W. y K. Lenzo. 2000. Limited Domain Synthesis. En *ICSLP*, Beijing, China.

Black, A.W. y P. Taylor. 1997. Automatically clustering similar units for unit selection in speech synthesis. En *EuroSpeech*, páginas 601–604, Rodes, Greece.

Breen, A. y P. Jackson. 1998. Non-uniform unit selection and the similarity metric within BT's LAUREATE TTS system. En *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, NSW, Australia.

Chu, M., C. Li, P. Hu, y E. Cahng. 2002. Domain adaption for TTS Systems . En *ICASSP*, Orlando, USA.

Deerwester, S., S.-T. Dumais, G.-W. Furnas, T.-K. Landauer, y R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal American Society Information Science*, 6(41):391–407.

Hyvärinen, A., Karhunen J. y E. Oja. 2001. *Independent Component Analysis*. John Wiley and Sons.

Isbell, C.-L. y P. Viola. 1999. Restructuring Sparse High Dimensional Data for Effective Retrieval. *Advances in Neural Information Processing Systems*, (11):480–486.

Joachims, T. 2000. SVMlight v3.50. http://ais.gmd.de/~ thorsten/svm_light/.

Kaban, A. y M. Girolami. 2000. Unsupervised Topic Separation and Keyword Identification in Document Collections: A Projection Approach. Technical Report Nr. 10, Dept. of Computing and Information Systems, University of Paisley.

Möbius, B. 2001. Rare events and closed domains: two delicate concepts in Speech Synthesis. En *Fourth ISCA Workshop on Speech Synthesis*, páginas 41–46, Perthshire, Scotland.

Montero, J.M., R. Córdoba, J.A. Vallejo, J. Gutiérrez-Arriola, E. Enríquez, y J.M. Pardo. 2000. Restricted-domain female-voice synthesis in Spanish: from database design to ANN prosodic modelling. En *ICSLP*, páginas 621 – 624, Beijing, China.

Rennison, E. 1994. Galaxy of News: An Approach to Visualizing and Understanding Expansive News Landscapes. En *ACM Symposium on User Interface Software and Technology*, páginas 3–12.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.

Sebastiani, F. 2002. Machine learning in automated text categorisation. *ACM Computing Surveys*, 34(1):1–47.

Tang, N. 2001. Text Categorisation using Support Vector Machines. Informe técnico, Department of Computer Science, University of Sheffield, United Kingdom.