Conversión de texto en habla multidominio basada en selección de unidades con ajuste subjetivo de pesos y marcado robusto de pitch

Francesc Alías Pujol

Universitat Ramon Llull

TESIS DOCTORAL

Conversión de texto en habla multidominio basada en selección de unidades con ajuste subjetivo de pesos y marcado robusto de *pitch*

Realizada por **Francesc Alías Pujol**

en el Centro **Enginyeria i Arquitectura La Salle**

y en el Departamento **Comunicaciones y Teoría de la Señal**

Dirigida por **Dr. Joan Claudi Socoró i Carrié**

Julio de 2006

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es

Per a tu, Piu

Agradecimientos

A Joan Claudi Socoró, mi director de tesis, por todas las discusiones y críticas, así como por su inestimable ayuda para formalizar las ideas de esta tesis. A Francesc Xavier Jové y Joan Lluís Pijoan, por aceptar ser mis directores de tesis durante los años 2000 y 2001.

A todos los compañeros del Área de Tecnologías del Habla, a los que han sido y a los que son. A Roger Guaus, por enseñarme a trabajar con rigor; a Josep Martí, por iniciarme en el apasionante mundo de las tecnologías del habla; y finalmente, a Ignasi Iriondo, por las innumerables discusiones sobre la línea de investigación planteada y por estar siempre dispuesto a colaborar en las pruebas subjetivas formales e informales, que han sido muchas.

A Antonio Bonafonte, por los consejos e ideas sugeridas, así como por la cesión de un corpus de voz en catalán para selección de unidades de la Universitat Politècnica de Catalunya junto al código del marcador de Entropic. Gracias por todo, Toni.

A Xavier Llorà, por ser, además de un gran amigo, el apoyo fundamental para desarrollar la investigación sobre el ajuste de pesos mediante algoritmos genéticos. Gracias por estar siempre que lo he necesitado y por tantas discusiones transoceánicas.

A Xevi Sevillano, por su amistad y por su colaboración en el desarrollo de este trabajo de investigación; sobretodo, gracias por ponerlo en duda.

A todos los proyectistas y colaboradores que han participado, de algún modo, en el desarrollo de esta tesis. A Íñigo Pérez, Javier Sanchis, Manuel Pablo Triviño, Jordi Domínguez, María Encarnación Navarro y Anna Navarro. Han sido muchos los corpus, las transcripciones fonéticas y las marcas de segmentación revisadas. A Pere Barnola, por ser mucho más que un proyectista, trabajando conmigo aún estando a muchos kilómetros de distancia. A Carlos Monzo y Xavi Gonzalvo, por su colaboración en la fase final de esta tesis. Han sido muchas pruebas y marcas de *pitch* compartidas. A David García, por contagiarme su entusiasmo y ganas de trabajar. Finalmente, quiero agradecer de forma especial el trabajo de Lluís Formiga durante estos años; gracias por intentar llegar siempre más allá de lo solicitado. Gracias a todos por vuestro esfuerzo pensando en mí. Ha sido un placer trabajar con vosotros y espero poder colaborar en vuestra investigación a partir de ahora.

A Àngel Rodríguez y Patrícia Lázaro del Departamento de Comunicación Audiovisual y Publicidad de la Universidad Autónoma de Barcelona, por el diseño y la grabación del corpus publicitario expresivo.

A Antònia Martí del Centro de Lingüística Computacional de la Universitat de Barce-

lona, por la cesión de un corpus de textos periodísticos en castellano.

A Tomoki Toda, por todas las charlas sobre ajuste de pesos compartidas de congreso en congreso.

A David Escudero y Quim Haro, por todos sus buenos consejos.

A Alain de Chevingé y Xuejing Sun por darme acceso a los algoritmos de estimación de la periodicidad YIN y SHRp, respectivamente.

A José Antonio Montero, por las discusiones algebraicas, por tantos cafés terapéuticos, y cómo no, por enseñarme el humor a tiempo real... al final hubo tesis en sín-tesis :)

A Rosa Alsina, por estar siempre que lo he necesitado.

A Javier Melenchón, por permitirme participar en el desarrollo del Locutor Virtual.

A David Miralles, por ser, además de un gran tipo, mi asesor personal de LaTeX.

Al resto de miembros del Departamento de Comunicaciones y Teoría de la Señal. En especial, a Elisa Martínez y David Badia por su apoyo y comprensión, y a Germán Cobo, Diego Torres, Santi Planet, Lourdes Meler y José Manuel Álvarez, por estar siempre dispuestos a echar una mano.

A todos los que han participado en las pruebas subjetivas, que han sido muchas. Muchas gracias. Sin vosotros esta tesis no habría sido.

A los compañeros del Departamento de Acústica encargados de las grabaciones de los corpus de voz, en especial a Ivana Rossell, Carles Vila y a Matthias Eibel.

A todos mis amigos, gracias por vuestro apoyo y los buenos momentos compartidos. Nunca olvidaré mi treinta cumpleaños.

A nivel personal, tengo que agradecer muy sinceramente el apoyo de mi familia, tanto la carnal como la política. A mi madre, por darme la oportunidad de estar donde estoy. A Maxi, por su apoyo incondicional, tanto moral como "informático". Iaia, Sílvia, Dolors, Carles, Jose, Montse, Emma y Clara, gracias por los buenos momentos que me habéis hecho vivir. En especial, a Rosa Maria, por estar siempre a mi lado, por su comprensión, y por animarme siempre a seguir en los momentos malos.

Finalmente, agradecer el apoyo que he recibido por parte del Departament d'Universitats Recerca i Societat de la Informació (DURSI) de la Generalitat de Catalunya mediante la beca de formación para personal investigador 2000FI-00679. Asimismo, parte de este trabajo se ha desarrollado en el marco de distintos proyectos de investigación financiados por el Ministerio Español de Ciencia y Tecnología (MCyT) mediante el Plan de Investigación Científica. Concretamente, se trata de los proyectos "*Locutor Virtual*"(FIT-150500-2002-410) e "*IntegraTV-4all*"(FIT-350301-2004-2). Asimismo, parte de la investigación desarrollada se ha llevado a cabo dentro del proyecto "*Personajes Virtuales*", financiado parcialmente por la Corporación Catalana de Radio y Televisión (CCRTV) y el Centre d'Innovació i Desenvolupament Empresarial (CIDEM), este segundo mediante el proyecto RDITSCON04-0005. Agradecer también a Enginyeria i Arquitectura La Salle el apoyo y la confianza prestada a lo largo de estos años.

Acrónimos y abreviaturas

AA Aprendizaje Artificial

aAGI Algoritmo genético interactivo activo

AE Algoritmos Evolutivos

AET *Average Execution Time*

AET *Average Execution Time*

AG Algoritmo genético

AGI Algoritmo genético interactivo

ASL *Average Segment Length*

CART *Classification and Regression Tree*

cGA *compact Genetic Algorithm*

CI Componente Independiente

COF *Coocurrence Frequency*

cPBIL *continuous* PBIL

cPL *cummulative Pattern Lenght*

CT Clasificación de Textos

CTH Conversión de Texto en Habla

CTH-DR CTH de Dominio Restringido

CTH-MD CTH Multidominio

CTH-PG CTH de Propósito General

CTH-SU CTH basada en Selección de Unidades

DTW *Dynamic Time Warping*

EGG Señal electroglotal

EI Extracción de Información

EM *Expectation-Maximization Algorithm*

GER *Gross Error Rate*

GPMER *Gross Pitch Marks Error Rate*

HMM *Hidden Markov Models*

HNM *Harmonic plus Noise Model*

ICA *Independent Component Analysis*

IDF *Inverse Document Frequency*

ITP *Interfície de Tractament de la Parla*

IWF *Inverse Word Frequency*

LPC *Linear Predictive Coding*

LSF *Line Spectral Frequencies*

LSI *Latent Semantic Indexing*

LSP *Line Spectral Pairs*

MAP Módulo de Ajuste del *Pitch*

MBROLA *Multi-Band Resynthesis Overlap and Add*

MEV Modelo de Espacio Vectorial

MFCC *Mel Frequency Cepstral Coefficients*

MLR *Multilinear Regression*

MOS *Mean Opinion Score*

NCOF *No Coocurrence Frequency*

NN *Nearest Neighbour*

PBIL *Problem Based Incremental Learning*

PDA *Pitch Detection Algorithm*

PDS Procesamiento Digital de la Señal

PESQ *Perceptual Evaluation of Speech Quality*

PL *Pattern Length*

PLN Procesamiento del Lenguaje Natural

PMA *Pitch Marking Algorithm*

PMFA *Pitch Marks Filtering Algorithm*

POS *Part-of-Speech (Tagging)*

PSOLA *Pitch Synchronous Overlap and Add*

RAH Reconocimiento Automático del Habla

RAPT *Robust Algorithm for Pitch Tracking*

RI Recuperación de Información

RRA Red Relacional Asociativa

RRA F RRA *Full*

RRA R RRA Reducida

RWS *Roulette Wheel Selection*

SAMPA *Speech Assessment Methods Phonetic Alphabet*

SAPI *Speech Application Program Interface*

SDR *Spoken Document Retrieval*

SHR *Subharmonic-to-Harmonic Ratio*

SLH Sistema de Lenguaje Hablado

SSML *Speech Synthesis Markup Language*

SSML-M SSML modificado

SVM *Support Vector Machines*

SYN *Lip Synchronization*

TD-PSOLA *Time Domain PSOLA*

TDT *Topic Detection and Tracking*

TF *Term Frequency*

UER *Unvoiced Error Rate*

VER *Voiced Error Rate*

WSS *Weight Space Search*

XTC *XML Time Code*

Resumen

El propósito final de la conversión de texto en habla (CTH) es la generación de habla sintética completamente natural a partir de un texto de entrada cualquiera. Históricamente, se han seguido dos estrategias para lograr este objetivo: la que prima la flexibilidad de la conversión ante la calidad de la síntesis, dando lugar a los sistemas de conversión de texto en habla de propósito general (CTH-PG); y la que antepone la naturalidad de la síntesis a la generalidad de la CTH, conocida como conversión de texto en habla de dominio restringido (CTH-DR). En la actualidad, la estrategia más utilizada para desarrollar los sistemas de CTH es la conversión de texto en habla basada en corpus o por selección de unidades (CTH-SU). Aunque la calidad de los sistemas de CTH-SU es bastante buena en general, todavía existen elementos que continúan siendo fuente de investigación.

En esta tesis se presentan distintas aportaciones en el contexto de la CTH-SU para mejorar, por un lado, la naturalidad de los sistemas de CTH-PG y, por otro, la flexibilidad de los sistemas de CTH-DR. Para abordar la primera cuestión, se presenta una técnica que permite incorporar de forma eficiente la percepción humana al proceso de selección de las unidades del corpus de voz mediante el ajuste subjetivo de los pesos de la función de coste que guía la selección de las unidades, controlando la fatiga y la consistencia del usuario. Asimismo, se presenta un método para mejorar la fiabilidad del proceso de etiquetado automático del corpus de voz, concretamente, de las marcas de *pitch* —cuestión fundamental en el contexto de los CTH basados en selección de unidades. En cuanto al segundo problema, y siguiendo la estrategia de CTH-DR, se presenta la conversión de texto en habla multidominio (CTH-MD), que persigue conseguir una calidad sintética equivalente a la de los sistemas de CTH-DR, aumentando su flexibilidad al considerar distintos dominios (estilos de locución, emociones, temáticas, etc.) para la síntesis. En este contexto, es necesario que el sistema de CTH-MD conozca, durante el proceso de conversión de texto en habla, qué dominio o dominios son los más adecuados para poder sintetizar el texto de entrada con la mayor naturalidad posible. En este caso, el sistema de CTH-MD incorpora un módulo de clasificación de textos a la arquitectura clásica de los sistemas de CTH adaptado a las necesidades que plantea la CTH-MD. Finalmente, todas las propuestas descritas se evalúan en términos objetivos —mediante el uso de medidas clásicas junto a nuevas propuestas— y/o subjetivos —mediante pruebas de percepción— para validar las mejoras conseguidas por los métodos desarrollados en el contexto de la CTH-SU en el camino hacia el desarrollo de nuevos sistemas de CTH de elevada calidad y flexibilidad.

Resum

El propòsit final de la conversió de text a parla (CTP) és la generació de parla sintètica completament natural a partir d'un text d'entrada qualsevol. Històricament, s'han seguit dues estratègies per a assolir aquest objectiu: la que prima la flexibilitat de la conversió davant la qualitat de la síntesi, donant lloc als sistemes de conversió de text a parla de propòsit general (CTP-PG); i la que anteposa la naturalitat de la síntesi a la generalitat de la CTP, coneguda com a conversió de text a parla de domini restringit (CTP-DR). En l'actualitat, l'estratègia més utilitzada per a desenvolupar els sistemes de CTP és la conversió de text a parla basada en corpus o per selecció d'unitats (CTP-SU). Tot i que la qualitat dels sistemes de CTP-SU és bastant bona en general, encara existeixen qüestions que continuen essent font d'investigació.

En aquesta tesi es presenten diverses aportacions en el context de la CTP-SU per a millorar, d'una banda, la naturalitat dels sistemes de CTP-PG i, per l'altra, la flexibilitat dels sistemes de CTP-DR. Per abordar la primera qüestió, es presenta una tècnica que permet incorporar de forma eficient la percepció humana al procés de selecció de les unitats del corpus de veu mitjançant l'ajust subjectiu dels pesos de la funció de cost que guia la selecció de les unitats, controlant la fatiga i la consistència de l'usuari. Així mateix, es presenta un mètode per a millorar la fiabilitat del procés d'etiquetatge automàtic del corpus de veu, concretament, de les marques de *pitch* —qüestió fonamental en el context dels CTP basats en selecció d'unitats. En quant al segon problema, i seguint l'estratègia de CTP-DR, es presenta la conversió de text a parla multidomini (CTP-MD), que persegueix aconseguir una qualitat sintètica equivalent a la dels sistemes de CTP-DR, augmentant la seva flexibilitat per considerar diferents dominis (estils de locució, emocions, temàtiques, etc.) per a la síntesi. En aquest context, és necessari que el sistema de CTP-MD conegui, durant el procés de conversió de text a parla, quin domini o dominis són els més adequats per a poder sintetitzar el text d'entrada amb la major naturalitat possible. En aquest cas, el sistema de CTP-MD incorpora un mòdul de classificació de textos a l'arquitectura clàssica dels sistemes de CTP adaptat a les necessitats que planteja la CTP-MD. Finalment, totes les propostes descrites s'avaluen en termes objectius —mitjançant l'ús de mesures clàssiques juntament amb noves propostes— i/o subjectius —mitjançant proves perceptives— per a validar les millores aconseguides pels mètodes desenvolupats en el context de la CTP-SU en el camí cap al desenvolupament de nous sistemes de CTP d'alta qualitat y flexibilitat.

Abstract

The final purpose of any Text-to-Speech (TTS) system is the generation of perfectly natural synthetic speech from any input text. Historically, two strategies have been followed in the quest for this goal: the general purpose TTS synthesis (GP-TTS), which strives the flexibility of the application at the expense of the achieved synthetic speech quality; and the limited domain TTS synthesis (LD-TTS), which prioritizes the development of high quality TTS systems by restricting the scope of the input text. At present, the most used strategy to develop TTS systems is the so called corpus-based text-to-speech or unit selection TTS (US-TTS) synthesis. Although the quality of US-TTS synthesis systems is quite good in general, there are still several open issues which are still being investigated.

This PhD thesis introduces different contributions for US-TTS systems in order to improve, by one hand, the naturalness of GP-TTS systems, and by the other hand, the flexibility of LD-TTS systems. To deal with the former problem, a new technique for efficiently incorporating human perception in the unit selection process by means of subjective weight tuning is introduced, which also allows controlling user fatigue and user consistency. Moreover, a new method for improving the reliability of automatic speech corpus labelling is described, particularly, a generic pitch marks filtering algorithm is introduced —an essential issue in corpus-based TTS systems. Moreover, the latter problem is addressed by multi-domain TTS (MD-TTS) synthesis, following the LD-TTS approach, which deals with achieving synthetic speech quality equivalent to that of LD-TTS systems, but improving TTS flexibility by considering different domains (speaking styles, emotions, topics, etc.) for conducting speech synthesis. In this context, the MD-TTS system needs to know, at run time, which domain or domains are the most suitable for synthesizing the input text with the highest synthetic speech quality. To that effect, the MD-TTS system incorporates a text classification module to classic TTS synthesis architecture adapted to the MD-TTS classification particularities. Finally, all the proposals are evaluated in terms of objective experiments —by means of classic or new measures— and/or subjective tests —perceptual tests— in order to validate the improvements achieved by the methods developed in the US-TTS framework, as a step further in our research towards developing high quality and flexible text-to-speech synthesis systems.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

Los métodos de acceso a la información actuales, generalmente basados en un teclado y/o un ratón, distan mucho de las vías de comunicación natural entre las personas y suponen, por ello, una barrera de acceso a un amplio colectivo poco familiarizado con las nuevas tecnologías o bien con ciertas discapacidades, la denominada *brecha digital*. En este contexto, las interfaces persona-máquina multimodales (p.ej. las cabezas parlantes) pueden jugar un papel fundamental para conseguir simular la interacción entre humanos en la comunicación con las máquinas, permitiendo un acceso más democrático a las nuevas tecnologías. Dado que el habla es uno de los medios de comunicación más utilizados entre las personas, este tipo de interfaces deberán ser capaces de tratar con la señal de voz, tanto para acceder a ellas (reconocimiento del habla), como para dar respuesta a las peticiones del usuario (síntesis del habla y sistemas de diálogo).

Una de las líneas de investigación relacionadas con la síntesis del habla es la conversión de texto en habla (CTH). Los sistemas de CTH se encargan de transformar un texto cualquiera en su mensaje oral correspondiente de la forma más natural posible (Yi y Glass, 1998; Taylor, 2000), es decir, deben ser capaces de emular la variabilidad del habla humana (idiomas, estilos de locución, emociones, etc.). Existen diversas aplicaciones donde los sistemas de CTH pueden ser aplicados, entre otros, en el acceso a información electrónica (p.ej. lectura de correo electrónico), para los servicios de información (p.ej. guías de museos o navegación en vehículos), los portales de voz, los juegos interactivos o la educación (p.ej. aprendizaje de segundas lenguas).

El trabajo de investigación que se ha desarrollado en esta tesis se centra en la mejora de la calidad y la flexibilidad de los sistemas de conversión de texto en habla, con el objetivo de dar un paso más hacia la consecución de unos sistemas de CTH capaces de generar voz sintética *perfecta*. Concretamente, las contribuciones de esta tesis se enmarcan en la CTH por concatenación basada en corpus o *selección de unidades* (Sagisaka et al., 1992; Black y Campbell, 1995; Hunt y Black, 1996). Esta estrategia de síntesis concatenativa se ha convertido durante la última década en la tecnología más utilizada en el ámbito de la CTH, debido al salto cualitativo que ha permitido en lo que se refiere a la calidad sintética obtenida (Chu et al., 2002).

1.1. Motivación y objetivos

La mejora en la calidad sintética obtenida por los sistemas de conversión de texto en habla basados en selección de unidades (CTH-SU) se ha basado, fundamentalmente, en el aumento del tamaño del corpus de voz gracias a la enorme mejora de la capacidad computacional de los ordenadores a lo largo de la pasada década. En este contexto, los sistemas de CTH-SU son capaces de generar voz con las características acústicas de la señal de voz grabada, ya que siguen la filosofía de seleccionar la mejor secuencia de unidades del corpus, minimizando la modificación de la señal (prosódica, puntos de concatenación, etc.) necesaria para construir la señal sintética (Balestri et al., 1999). De algún modo, la síntesis del habla se ha convertido en un problema de recopilación, etiquetado, indexado y recuperación de datos sobre grandes corpus de voz (Chu et al., 2001), dejando en un segundo plano el papel relevante que la investigación sobre técnicas de procesamiento digital de la señal había tenido hasta ese instante.

En la actualidad, parece claro que los sistemas de CTH-SU funcionan muy bien cuando trabajan sobre dominios limitados (Black y Lenzo, 2000; Black, 2003). No obstante, esta calidad se reduce cuando la CTH-SU se aplica a la síntesis de propósito general (Breuer y Abresch, 2004; Toda y Tokuda, 2005), ya que se trata de sistemas poco flexibles en lo que se refiere a generar señal sintética de elevada calidad fuera del estilo de locución utilizado en la grabación del corpus (Black, 2003). Si se pretende obtener una señal de voz con unas características distintas a las de la señal grabada (estilo de locución, emoción o calidad vocal), la modificación de la misma producirá una reducción de la calidad sintética (Yamagishi et al., 2003; Yamagishi et al., 2005). Es por ello que continúa siendo necesario investigar sobre nuevas técnicas de CTH que permitan controlar las características de la señal de voz generada (Toda, 2003; Miyanaga, Masuko y Kobayashi, 2004), de forma que la señal sintética generada tenga una elevada calidad sobre un contexto de mayor flexibilidad. En esta línea, recientemente, los sistemas de CTH basados en Modelos Ocultos de Markov (*Hidden Markov Models* o HMM, en inglés) han tomado una relevancia especial (Huang et al., 1997; Yamagishi et al., 2003; Yamagishi et al., 2005; Toda y Tokuda, 2005; Barros et al., 2005; Zen y Toda, 2005; Zhao et al., 2006). Esta tecnología pone por delante la flexibilidad del sistema de CTH ante la calidad de sintética conseguida, gracias a la parametrización de la señal de voz —en lugar de trabajar directamente con las muestras de voz, como hace CTH-SU.

No obstante, en el ámbito de los sistemas de CTH-SU todavía queda camino por recorrer con el objetivo de mejorar la calidad sintética obtenida (todavía existen situaciones donde la naturalidad sintética es pobre) y su flexibilidad (se trata de una estrategia fuertemente dependiente del corpus de voz disponible). El objetivo fundamental de esta tesis se basa en la definición de distintas estrategias que permitan aumentar la flexibilidad y la calidad de los sistemas de síntesis basados en selección de unidades. Para ello, el trabajo de investigación aborda tres elementos clave en el contexto de la CTH-SU:

- Mejorar la calidad sintética de la CTH-SU mediante la definición de un nuevo método de ajuste subjetivo de pesos eficiente. Dado que el receptor final de la señal sintética

es una persona, resulta fundamental incorporar la percepción humana en el proceso de selección de las unidades del corpus. Durante años, se han dedicado multitud de esfuerzos a intentar obtener alguna medida objetiva capaz de mapear la subjetividad humana de forma eficiente en el contexto de la CTH-SU. No obstante, se trata de una tarea muy complicada que no ha conseguido dar con una solución clara al problema. En este trabajo se cambia el enfoque al problema, ya que es el usuario el que guía de forma explícita el proceso de entrenamiento de los pesos de la función de coste de selección. Este cambio de enfoque provoca tener que abordar nuevos problemas, tal y como se detalla a lo largo de esta tesis.

- Aumentar la flexibilidad de la CTH-SU manteniendo la calidad sintética mediante la definición de una nueva estrategia de CTH capaz de sintetizar textos para distintos dominios denominada CTH multidominio o CTH-MD. De este modo, se abre la posibilidad de dotar a este tipo de sistemas la variabilidad sintética que tienen por definición los sistemas basados en HMM, en aras de conseguir unos sistemas de síntesis capaces de sintetizar señal de voz de alta calidad para distintos estilos de locución. Asimismo, la propuesta de CTH-MD está estrechamente ligada a la definición de un método automático de selección de dominio, en este caso, implementado mediante un clasificador de textos.
- Dotar de mayor calidad al proceso de etiquetado de los corpus de voz mediante un método de filtrado y ajuste robusto de las marcas de *pitch*. El corpus de voz es uno de los elementos clave de los sistemas de síntesis actuales, por lo que su caracterización robusta es fundamental. En el contexto de los CTH-SU el tamaño de estos corpus hace prácticamente inviable una revisión exhaustiva de los procesos de etiquetado necesarios para caracterizar y procesar las unidades acústicas que contienen. Es por estos motivos, que resulta fundamental disponer de herramientas que sean capaces de garantizar un etiquetado del corpus con un número mínimo de errores.

1.2. Marco del trabajo

El grupo de investigación en Tecnologías del Habla de Enginyeria i Arquitectura La Salle ha sido uno de los grupos pioneros en el campo de la síntesis del habla en España desde la década de los ochenta con los trabajos del Dr. Josep Martí (Martí, 1985; Martí, 1987; Martí, 1990). Se llevaron a cabo trabajos de investigación y desarrollo basados en estrategias de síntesis articulatoria, síntesis por formantes y síntesis basada en predicción lineal (LPC) (Alías y Iriondo, 2002). Estos sistemas se aplicaron, principalmente, a productos orientados a personas invidentes. En concreto, la colaboración que se mantuvo con la empresa CIBERVEU, S.A. del grupo ONCE se materializó en un conjunto de equipos de síntesis de voz, capaces de leer la información contenida en la pantalla o desde el teclado de un ordenador. En ese momento resultaba imprescindible que estos equipos fueran externos al ordenador personal, debido a la poca capacidad computacional y de memoria de la que los ordenadores disponían entonces. El producto más destacable de esa gama fue el CIBER232P, un equipo portátil que incorporaba el circuito integrado PCF8200 como módulo de síntesis. La

comunicación con el ordenador se realizaba vía serie y el equipo funcionaba como sistema de CTH para el catalán y el castellano (Llisterri et al., 1993).

Hasta aquel instante, la calidad de los sistemas de síntesis del habla sólo estaba condicionada por su grado de inteligibilidad, factor muy apreciado para un CTH orientado a personas con discapacidades visuales. El siguiente reto abordado pasó por incrementar la inteligibilidad de la señal sintética, motivo por el cual se centró el esfuerzo del grupo en la mejora de tres áreas que resultan críticas en lo referente a la calidad del resultado final de la síntesis:

- El análisis lingüístico del texto
- El modelado y la predicción automática de la prosodia (entonación, ritmo y energía)
- El procesamiento digital de la señal (PDS) para la genaración del habla sintética

Seguidamente, el progresivo avance tecnológico comportó una mayor capacidad de proceso y de memoria para los ordenadores personales, cosa que motivó el desarrollo de sistemas de CTH basados únicamente en *software*. Con el objetivo de mejorar el bloque de PDS, se optó por la síntesis concatenativa basada en difonemas y/o trifonemas [1]. Durante este periodo se implementó un sintetizador en catalán (Camps, Bailly y Martí, 1992; Guaus, Gudayol y Martí, 1996; Guaus et al., 1997), basado en la reciente publicación de la técnica PSOLA (*Pitch Synchronous Overlap and Add*) del trabajo de Moulines y Charpentier (1990), técnica que tuvo (y continua teniendo) una gran repercusión en la comunidad científica. Este CTH en catalán se convirtió en la base de los posteriores sistemas de síntesis desarrollados en nuestro grupo de investigación.

Investigación en síntesis por concatenación

Durante el año 1997, se inició el desarrollo de un nuevo sistema de CTH en catalán con el objetivo de obtener una alta inteligibilidad, partiendo de la experiencia acumulada durante los trabajos anteriores gracias a un proyecto financiado por Televisió de Catalunya, S.A. Se fijó el objetivo de mejorar los diferentes módulos que componían el CTH existente. En primer lugar, se mejoró el bloque de preprocesamiento del texto. A continuación, se desarrolló un lenguaje de reglas compiladas para llevar a cabo la conversión grafema-fonema (trascripción fonética) y se diseñó un módulo para determinar automáticamente la prosodia de las unidades.

En cuanto al módulo de síntesis (PDS) se llevaron a cabo mejoras en el proceso de concatenación de las unidades y se grabó, segmentó y etiquetó un nuevo corpus para el catalán de 1207 unidades, formado por 895 difonemas y 312 trifonemas (con una realización por

[1] Un difonema se define como la unidad sonora que empieza en la parte estable de un fonema y acaba en la parte estable del fonema contiguo, es decir, son dos medios fonemas o semifonemas, mientras que un trifonema incluye dos medios fonemas más un fonema central que normalmente tiene un carácter poco estable, p.ej. una /r/ —en este trabajo se utiliza la notación fonética SAMPA (Wells et al., 1992)

unidad). Con este proyecto se logró un sistema de CTH con una muy buena inteligibilidad y una naturalidad aceptable para ser integrado en distintas aplicaciones de conversión de texto en habla.

A continuación se describen algunos de los módulos más importantes obtenidos durante el desarrollo de este proyecto.

- **Editor de Mensajes Orales con Voz Sintética (EMOVS):** durante este periodo se diseñó e implementó una aplicación bajo entorno Windows® que permite la síntesis automática del habla a partir de un texto de entrada (ver figura 1.1). El objetivo inicial fue diseñar una herramienta que permitiera la modificación manual de la prosodia de la señal de voz, para ajustarla antes de ser empleada en un determinado canal de comunicación. Los parámetros editables son: la energía, la frecuencia fundamental, la duración de los fonemas y de las pausas y la trascripción fonética. La modificación prosódica se puede realizar a nivel de frase, de fonema o grupo de fonemas consecutivos. Esta herramienta ha sido empleada tanto para la mejora manual de las locuciones sintéticas, como para la extracción y el modelado de patrones prosódicos (Iriondo et al., 2000).

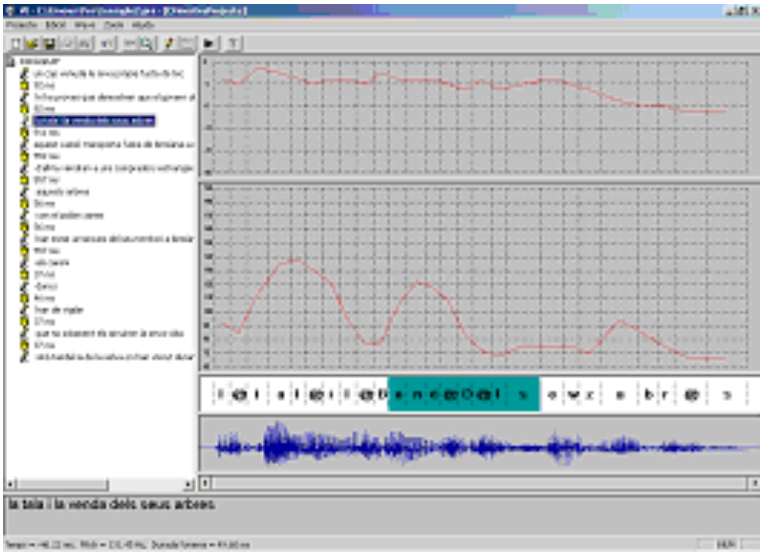


Figura 1.1: Ejemplo de EMOVS donde se puede observar el texto organizado en frases junto a la información fonética y prosódica asociada a la frase sintetizada.

- **Mejoras en la calidad segmental:** con el fin de mejorar la calidad de la síntesis del habla por concatenación de unidades, se abordó el estudio de los problemas surgidos de la unión entre difonemas y la modificación de su duración mediante la técnica PSOLA en el dominio temporal (en inglés, *Time Domain* PSOLA o TD-PSOLA) (Moulines y Charpentier, 1990). El hecho de trabajar con síntesis concatenativa comporta que

la evolución temporal del espectro de la señal generada sufra discontinuidades importantes en el centro de cada alófono. Estas discontinuidades provocan una degradación importante de la calidad del habla sintética. Por esta razón, se estudiaron diferentes técnicas para disminuir este efecto y mejorar la calidad global del sistema. Se aplicaron técnicas de interpolación entre tramas y se modificó el punto de concatenación de las unidades, permitiendo uniones en cualquier instante del fonema (Guaus et al., 1998). Hasta ese instante se utilizaban técnicas convencionales para modificar la duración de los fonemas: repetición de la señal en la parte estable del alófono, repetición *intercalada* de la señal, etc., cuestión que generaba una señal poco natural cuando las modificaciones de la duración eran importantes. La solución que se introdujo se basó en un sistema de transformación y combinación de las tramas de los segmentos adyacentes de voz proporcionando una calidad de voz superior al método anterior (Iriondo et al., 1999). De este modo se dio un paso adelante hacia la consecución de un sistema de CTH basado en difonemas de alta inteligibilidad.

- **Lenguaje de reglas (SINCAT2):** durante este periodo también se abordó el proceso de automatización de la generación de la prosodia a partir del texto de entrada, proceso por otro lado, nada sencillo. La solución que desarrolló el grupo se basó en la generación e implementación de un lenguaje para la interpretación de reglas, que permitiera convertir un texto en su correspondiente trascripción fonética y asignara a cada unidad su prosodia (entonación, ritmo y energía). Estas reglas se diseñaron para que fueran sencillas, claras y fácilmente modificables por el usuario. El nombre del lenguaje, SINCAT, es el acrónimo correspondiente a las palabras SÍNtesis en CATalán, y tal y como indica el título de este párrafo, el sistema de CTH actual trabaja con la segunda versión de este lenguaje de reglas compilado.

Investigación en CTH expresiva

Al mismo tiempo que se estaba trabajando en la mejora de la CTH en catalán, se empezó a trabajar en el contexto de la síntesis del habla expresiva, es decir, la transmisión del estado emocional del hablante a partir de la voz sintetizada. Aunque la mayoría de los sistemas actuales de síntesis se caracterizan por una gran inteligibilidad y una buena naturalidad, todavía queda camino por recorrer en lo que se refiere a la generación realista de habla expresiva. A partir de la colaboración con el Departamento de Comunicación Audiovisual y Publicidad (CAP) de la Universidad Autónoma de Barcelona (UAB), se inició una línea de investigación en el campo del modelado y la generación automática del habla expresiva o emocionada. Se partió de un modelo acústico de la expresión emocional (Rodríguez et al., 1999) que se modeló mediante un sistema de CTH (Iriondo et al., 2000). De este trabajo, se concluyó que para ciertas emociones se requiere de un módulo de procesamiento digital de la señal que permita grandes variaciones prosódicas (p.ej. la alegría). Por otro lado, cabe añadir a esta descripción que, recientemente, se ha llevado a cabo el primer paso para la automatización del modelado prosódico para síntesis expresiva (Iriondo et al., 2004), que próximamente será incluido dentro del sistema de CTH que se está desarrollando en la actualidad.

También se dedujo que la estrategia de síntesis empleada hasta el momento (TD-PSOLA (Moulines y Charpentier, 1990)) no tenía la suficiente versatilidad para lograr unos resultados lo suficientemente satisfactorios en lo que se refiere a naturalidad sintética. Por esta razón se estudiaron las diferentes alternativas de concatenación de la señal de voz que habían aparecido en los últimos años: desde MBROLA (*Multi-band Resynthesis PSOLA*) (Dutoit y Leich, 1996; Dutoit, 1997) hasta el modelo HNM (armónico-estocástico o *Harmonic + Noise Model* en inglés) de Stylianou (1996). Con el objetivo de incluir las nuevas ideas introducidas en el campo de la concatenación de tramas de voz, se decidió diseñar un modelo híbrido entre TD-PSOLA y los modelos armónicos-estocásticos (Stylianou, Dutoit y Schroeter, 1997; Stylianou, 2001) que permitiera variaciones prosódicas de mejor calidad (Iriondo, Alías y Melenchón, 2002). Actualmente, se está trabajando sobre un módulo de síntesis del habla basado en este modelo híbrido para ser incorporado en el sistema de CTH del grupo, cuyos primeros resultados se presentaron en (Iriondo et al., 2003).

Primeros pasos hacia la CTH basada en selección de unidades

Paralelamente, el grupo de Tecnologías del Habla se puso a trabajar sobre la CTH basada en selección de unidades, estudiando el diseño y el tamaño óptimos del corpus de voz (Guaus y Iriondo, 2000a; Guaus y Iriondo, 2000b), elemento crítico para el buen funcionamiento de este tipo de sistemas. Intuitivamente, parece que cuanto mayor sea el tamaño del corpus, mejor calidad sintética se obtendrá. No obstante, en estos trabajos se demostró que aunque esta intuición parece cierta, la relación entre el número de unidades del corpus y la calidad *esperada* de la señal sintética no es lineal —concretamente, es logarítmica—, por lo que se puede definir un determinado umbral a partir del cual el coste de aumentar el tamaño del corpus no compense la mejora de la calidad sintética obtenida. No obstante, estas consideraciones han quedado fuera del ámbito del presente trabajo de investigación.

1.3. Contribuciones del trabajo de investigación

El objetivo final de esta tesis es aportar nuevas estrategias para mejorar la calidad y la flexibilidad de los sistemas de conversión de texto en habla basados en selección de unidades. Para ello, el trabajo se fundamenta en la aplicación de diferentes técnicas que provienen de otras áreas de investigación con el objetivo de plantear nuevas soluciones a estas problemáticas. Por una parte, se estudia la aplicación de los algoritmos genéticos al proceso de ajuste subjetivo de los pesos de la función de coste de selección. Por otra parte, se incorporan técnicas del mundo de la clasificación de textos para poder considerar distintos *dominios* de síntesis dentro de un sistema de CTH multidominio. Finalmente, se define un método genérico para el marcado robusto del *pitch*, pensado fundamentalmente, para abordar el problema del etiquetado de los grandes corpus de voz para CTH-SU, pero aplicable a cualquier otra tarea relacionada con el marcado automático de la periodicidad de la señal de voz.

Para ello, las contribuciones fundamentales de esta tesis en cada una de estas tres líneas

de investigación han sido:

Sobre el ajuste de pesos de la función de coste de selección:

- Diseño de un método subjetivo de entrenamiento de los pesos de la función de coste del proceso de selección de unidades basado en un algoritmo genético interactivo activo (aAGI). Se consigue reducir la fatiga del usuario, mejorando la consistencia de las evaluaciones y obteniendo una calidad sintética mayor que los métodos de referencia (regresión lineal, algoritmo genético y algoritmo genético interactivo).
- Definición de una nueva medida para evaluar la consistencia de un usuario a lo largo de un proceso de consulta interactivo basado en aAGI.

Sobre la flexibilidad de los sistemas de conversión de texto en habla actuales:

- Propuesta de la nueva filosofía de CTH multidominio (CTH-MD) basada en selección de unidades, inspirada en los sistemas orales multidominio (sistemas de diálogo, de reconocimiento de voz, etc.), mejorando la flexibilidad de los sistemas de CTH-SU de alta calidad.
- Incorporación de un módulo encargado de asignar el dominio del texto de entrada (clasificador de textos) a la arquitectura clásica de los sistemas de CTH.
- Propuesta de un nuevo método de clasificación de textos basado en una red relacional asociativa (RRA) adaptado al contexto de la CTH-MD, que incorpora la clasificación temática y estilística de los textos.
- Propuesta de un modelo de RRA reducida (RRA R) capaz de minimizar la carga computacional del proceso de clasificación automática, sin afectar a la capacidad de clasificación (incluso es un método más robusto que el anterior respecto a la parametrización del texto y la calidad de los datos de entrenamiento).

Sobre el etiquetado robusto de marcas de *pitch*:

- Propuesta de un algoritmo de ajuste robusto de marcas de *pitch* (denominado PMFA) basado en la programación dinámica restringida, a partir de un algoritmo de detección o de marcado de *pitch* cualquiera, mejorando sus tasas de fiabilidad (para corpus de voz con distintas características vocales); con la particularidad de que se trata de un algoritmo cuya configuración es relativamente sencilla.
- Propuesta de la nueva medida de evaluación de la fiabilidad de los algoritmos de marcado de *pitch* con independencia del criterio local utilizado (GPMER), sin necesitar de un alineamiento temporal previo de las marcas y con una precisión mayor que las medidas utilizadas para evaluar los algoritmos de detección de *pitch*.

1.4. Organización de la tesis

El capítulo 2 se centra en el estudio y la propuesta de nuevos métodos de ajuste de los pesos que ponderan la función de coste utilizada para seleccionar las unidades del corpus de voz en el contexto de los sistemas de conversión de texto en habla basados en selección de unidades (CTH-SU). En primer lugar, una vez presentado, a grandes rasgos, el funcionamiento de los sistemas de CTH-SU, se describen las características fundamentales del proceso de selección de unidades de los sistemas de CTH basados en corpus, presentando los métodos clásicos de ajuste de pesos y los subcostes utilizados. A continuación, una vez comentadas las estrategias desarrolladas recientemente para el entrenamiento de pesos, se describe el camino seguido para definir el nuevo método de ajuste subjetivo de pesos desarrollado, pasando por la definición de un primer método objetivo de ajuste, ambos basados en la aplicación de los algoritmos genéticos al problema. Seguidamente se describe el método de ajuste subjetivo de pesos basado en algoritmos genéticos interactivos activos desarrollado, junto a la nueva medida de evaluación de la consistencia de los usuarios para controlar la fiabilidad de sus evaluaciones a lo largo del ajuste interactivo de pesos. Finalmente, se presentan los experimentos realizados que permiten validar la mejora conseguida con el método propuesto tanto en términos objetivos como subjetivos.

En el capítulo 3 se presenta la nueva estrategia de conversión de texto en habla multidominio (CTH-MD) desarrollada. Primero, se describen el funcionamiento de los sistemas multidominio existentes en el ámbito de la investigación en tecnologías del habla. Seguidamente, se resume el funcionamiento de los sistemas de CTH de propósito general y de dominio restringido y se presentan las dos técnicas más importantes para el diseño de corpus de voz multidominio para CTH-SU. Una vez definida la arquitectura del CTH-MD para CTH-SU, se procede a describir el sistema de clasificación de textos (CT) desarrollado para dar cobertura a las necesidades que plantea la CTH-MD, detallando las dos propuestas de clasificador de textos desarrolladas. Seguidamente, estas propuestas son evaluadas mediante un conjunto de pruebas exhaustivas, evaluando su eficiencia de clasificación así como el coste computacional que la tarea de clasificación implica. A continuación se presentan las pruebas subjetivas realizadas para evaluar el impacto del método de CT desarrollado sobre la calidad sintética obtenida. Finalmente, se discuten distintas cuestiones relacionadas con la propuesta a partir de las conclusiones extraídas de los experimentos, indicando también posibles líneas de trabajo futuras.

El capítulo 4 está dedicado a describir el método de marcado robusto de *pitch* desarrollado. Para empezar, se presenta una introducción que trata sobre el tipo de señal y los criterios de ubicación de las marcas utilizados habitualmente por los algoritmos de marcado de *pitch*. A continuación, se analizan las distintas propuestas de PMA existentes en la actualidad, haciendo hincapié en sus puntos críticos. A partir de las conclusiones obtenidas de este análisis, se propone un método genérico de marcado de *pitch* capaz de filtrar los errores de etiquetado de un algoritmo de detección o etiquetado de *pitch* de entrada cualquiera. Además, se presenta una nueva medida de evaluación capaz de evaluar y comparar PMAs con distintos criterios locales de marcado, sin tener que alinear previamente las marcas. Seguidamente, se analiza el funcionamiento de la propuesta en términos de la robustez de

marcado sobre dos corpus de voz distintos (uno de referencia y otro desarrollado en el marco de esta tesis), comparando su funcionamiento con tres métodos de etiquetado de referencia. Los resultados demuestran la gran mejora conseguida por el método propuesto a lo largo de las pruebas realizadas. Finalmente, se discuten distintos conceptos relacionados con la propuesta y se indican algunas líneas de mejora de la misma.

El capítulo 5 presenta las conclusiones más importantes de la investigación realizada, junto a distintas líneas de trabajo que quedan abiertas a partir de los resultados presentados en esta tesis.

Finalmente, esta tesis se acompaña de diversos anexos que tratan temas complementarios a la investigación descrita a lo largo de la memoria, destacando las interfaces y plataformas implementadas, junto a algunas de las aplicaciones en las que se ha utilizado parte del trabajo de investigación desarrollado.

Capítulo 2

Ajuste subjetivo de pesos eficiente

En este capítulo se presenta la primera de las aportaciones del trabajo de investigación. Se describe una estrategia robusta y eficiente para el ajuste subjetivo de los pesos de la función de coste involucrada en el módulo de selección de unidades de los sistemas de conversión de texto en habla basados en corpus (CTH-SU) —o *selección de unidades*. En este contexto, tanto el corpus de voz —que debe estar diseñado para conseguir una buena cobertura del idioma (o aplicación) y/o el estilo de locución deseados— como el módulo encargado de seleccionar las mejores unidades del corpus juegan un papel fundamental para conseguir una calidad sintética óptima. Concretamente, este módulo se fundamenta en dos elementos clave (Chu y Peng, 2001; Yi, 2003; Toda, 2003): *(i)* la *función de coste*, encargada de caracterizar la *bondad* de las unidades candidatas respecto a la secuencia de unidades objetivo y *(ii)* el proceso de *selección* de las unidades, encargado de escoger las unidades del corpus que *mejor* se ajusten a las especificaciones indicadas (trascripción fonética, prosodia, etc.). Por lo tanto, resulta fundamental disponer de un *buen* criterio de selección para conseguir una calidad sintética óptima en el contexto de los sistemas de CTH-SU (Chu y Peng, 2001; Yi, 2003; Campillo, 2005).

En general, los sistemas de CTH-SU han conseguido durante la pasada década dar un salto cualitativo importante en el ámbito de síntesis del habla. Gracias a esta tecnología, la síntesis pasó de conseguir señales con una buena *inteligibilidad* a obtener señales de voz de muy alta calidad (sobre todo cuando se recuperan expresiones *casi-enteras* del corpus). No obstante, todavía existen situaciones donde esta calidad no se mantiene a lo largo de toda la señal sintetizada (por ejemplo, en concatenaciones de naturalidad pobre) (Breuer y Abresch, 2004; Toda, Kawai y Tsuzaki, 2004; Toda y Tokuda, 2005), por lo que todavía queda espacio para continuar mejorando esta tecnología. Esto es debido, fundamentalmente, al propio enfoque de estos sistemas, donde el proceso de selección de unidades debe ser capaz de recuperar las unidades del corpus que conseguirán una mejor calidad sintética (criterio subjetivo), cuestión que debe estar reflejada en el proceso de ajuste de los parámetros considerados durante el proceso de selección (subcostes y pesos) y los elementos que intervienen en el proceso de selección (algoritmo de búsqueda, unidades consideradas, etc.) (Black, 2002; Yi, 2003).

El ajuste de pesos de la función de coste de selección es uno de los problemas que aún no han sido resueltos favorablemente en el contexto de la conversión de texto en habla basada en corpus (Campillo, 2005), debido a que se trata de una de las tareas más complejas relacionadas con el entrenamiento de los sistemas de CTH-SU (Hunt y Black, 1996; Lee, Lopresti y Olive, 2001; Toda, Kawai y Tsuzaki, 2004; Campillo, 2005; Zhao et al., 2006). Esto es debido a que el *criterio* utilizado para seleccionar las mejores unidades del corpus (función de coste) debe estar correlado con características *perceptuales*, para poder generar así una a señal sintética de máxima calidad (Black y Lenzo, 2001b; Lee, Lopresti y Olive, 2001; Peng, Zhao y Chu, 2002; Toda et al., 2002; Campillo, 2005; Toda, Kawai y Tsuzaki, 2004; Zhao et al., 2006). Es decir, de algún modo, la función de coste debería ser capaz de seleccionar la secuencia de unidades que, dadas sus particularidades fonéticas, prosódicas, lingüísticas, etc. consigan una mejor calidad sintética. En este contexto, el entrenamiento eficiente de los pesos de la función de coste de selección es uno de los elementos clave para conseguir una elevada calidad sintética.

Este capítulo del trabajo de investigación se centra en el estudio y la propuesta de una nueva técnica para el ajuste eficiente de los pesos de la función de coste de selección, utilizando criterios subjetivos. Para ello, se parte de un enfoque distinto al habitual para este problema, típicamente centrado en encontrar una medida objetiva capaz de mapear con garantías los criterios subjetivos de los usuarios. En este caso, en lugar de intentar reemplazar al usuario a través de un modelo matemático más o menos bien definido, la técnica desarrollada utiliza al propio usuario durante el proceso de entrenamiento de los pesos para validar la calidad sintética conseguida por una u otra configuración de pesos. En este contexto, se evita abordar el problema de definir una medida objetiva (no resuelto satisfactoriamente hasta el momento) a cambio de afrontar nuevos problemas, fundamentalmente: *(i)* conseguir una buena robustez estadística del entrenamiento de los pesos evitando alargar en exceso el tiempo del proceso y *(ii)* controlar la robustez del criterio del usuario en la toma de decisiones. Para ello, se presenta una estrategia de ajuste de los pesos basada en técnicas de aprendizaje artificial que es capaz de incorporar la subjetividad humana en el proceso de entrenamiento de los pesos de la función de coste de forma *eficiente*, evitando la fatiga del usuario y controlando su consistencia a la hora de guiar el proceso encargado de determinar los pesos.

A continuación, se pasan a describir todos los elementos involucrados en los sistemas de CTH concatenativos basados en selección de unidades. Como primer punto, se realiza una breve introducción al campo de la síntesis del habla, haciendo énfasis en la técnica de conversión de texto en habla basada en *selección de unidades*. Esta introducción recoge la descripción del módulo de selección de unidades, junto a las estrategias presentadas hasta el momento en la literatura que trata sobre el ajuste los pesos involucrados en la función de coste de este módulo. A continuación se describe el camino seguido hasta llegar a la definición de la estrategia de ajuste de pesos desarrollada, describiendo los distintos pasos realizados, junto a los motivos que los han guiado. Finalmente, se presentan los experimentos y resultados obtenidos mediante esta técnica, junto a los conseguidos a lo largo de la investigación desarrollada en el contexto del ajuste eficiente de los pesos de la función de coste del módulo de selección de unidades.

2.1. Introducción

2.1.1. La conversión de texto en habla

Un sistema para la conversión de texto en habla (CTH) es una herramienta (generalmente, un programa informático) desarrollada con el objetivo de transformar un texto *cualquiera* (introducido desde teclado, obtenido de un fichero, de un reconocedor óptico de caracteres, etc.) en la señal de voz correspondiente, con el objetivo de conseguir la máxima inteligibilidad y naturalidad posibles. Entre muchas otras aplicaciones, los sistemas de CTH pueden utilizarse en interfaces persona-máquina, aplicaciones multimedia, aplicaciones telefónicas, sistemas de diálogo, ayuda a la navegación de páginas *web*, ayuda a discapacitados o para el aprendizaje de segundas lenguas.

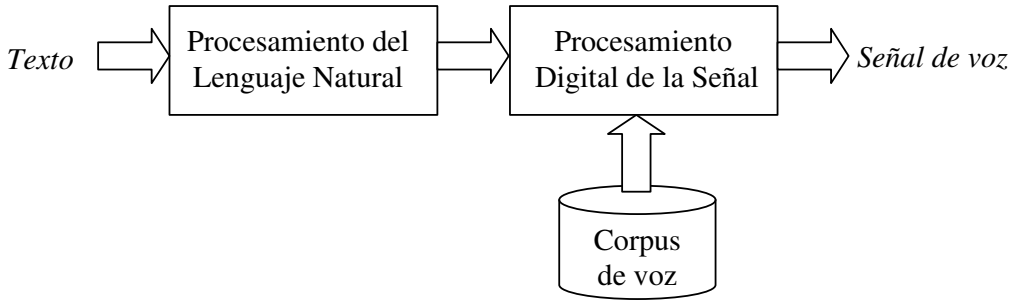


Figura 2.1: Diagrama de bloques simplificado de un conversor de texto en habla.

Según el enfoque presentado en (Dutoit, 1997), un sistema de CTH puede dividirse en dos grandes bloques (ver figura 2.1):

1. **El bloque de Procesamiento del Lenguaje Natural (PLN):** este bloque es el encargado de analizar el texto de entrada para obtener información sobre su trascripción fonética y sus características lingüísticas y prosódicas para la síntesis. El primer módulo del PLN es el preprocesador, que se encarga de normalizar el texto, pasando, por ejemplo, de números a letras, desambiguando los acrónimos, controlando las excepciones, etc. A continuación el texto pasa por el analizador morfosintáctico, que se encarga de subdividir el texto en grupos morfosintácticos. Para ello, el analizador morfológico examina cada una de las palabras del texto, asignándoles todas sus posibles interpretaciones morfológicas y extrayendo su lema. A continuación el módulo desambiguador selecciona la categoría más adecuada de cada palabra según el contexto sintáctico en el que se encuentra. Finalmente, el analizador sintáctico estructura el texto en sintagmas simples organizados jerárquicamente, a partir de la gramática asociada al idioma de trabajo del CTH. Acto seguido, se aplica el proceso de conversión de grafema a fonema mediante un fonetizador o transcriptor fonético[1]. Finalmente,

[1]En castellano y catalán, normalmente se trata de un sistema basado en reglas, mientras que para lenguas más complejas fonéticamente, como por ejemplo el inglés, se suele hacer uso de diccionarios.

el módulo prosódico asigna la energía, la duración y el tono más adecuados a las unidades fonéticas del texto de entrada para llevar a cabo la síntesis a partir de un determinado modelo aprendido a priori.

2. **El bloque de Procesamiento Digital de la Señal (PDS):** este módulo es el encargado de obtener las muestras de la señal de voz sintética a partir de la información (trascripción fonética más prosodia) obtenida del bloque de PLN. De las distintas estrategias que existen para llevar a cabo la síntesis de la señal, cabe destacar:

   - **Síntesis paramétrica o basada en reglas:** se basa en el modelado del proceso natural de generación de la señal de voz. Entre las estrategias más conocidas se encuentra la síntesis articulatoria, que parametriza los movimientos del tracto vocal por sus formantes, modelando las resonancias del tracto mediante filtros, o mediante un modelo matemático (p.ej. mediante coeficientes de predicción lineal), entre otros.
   - **Síntesis concatenativa:** se basa en la grabación y posterior unión de unidades de voz reales, normalmente pronunciadas por un locutor profesional. Existen dos aproximaciones fundamentales en los sistemas de CTH basados en síntesis concatenativa: la que se basa en el almacenamiento de una sola realización por unidad y la que trabaja con muchas realizaciones por unidad (varias copias de la misma unidad en contextos prosódicos y lingüísticos distintos), denominada síntesis basada en corpus o por *selección de unidades* (Black y Campbell, 1995; Hunt y Black, 1996).

De entre las distintas técnicas que se pueden utilizar dentro del bloque de PDS del CTH, una de las más empleadas por la comunidad científica hasta el momento, ha sido la síntesis concatenativa. Durante años, los sistemas de CTH han trabajado con corpus con una única realización por unidad, con el objetivo de conseguir una buena inteligibilidad del habla generada. Las unidades utilizadas pueden ser los fonemas, los alófonos, las semisílabas, los difonemas, etc. del idioma de trabajo. De entre las distintas opciones, las unidades más empleadas en los sistemas de CTH basados en síntesis concatenativa fueron los difonemas. Un difonema se define como la unidad sonora que empieza en la parte estable de un fonema y acaba en la parte estable del fonema contiguo, es decir, son dos medios fonemas o semifonemas. Así pues, el difonema contiene la zona de transición entre fonemas (punto de concatenación) que es uno de los elementos más complicados de modelar del proceso natural de la generación de la síntesis, permitiendo su unión con el difonema vecino por su parte estable. De este modo se consigue evitar la concatenación de las unidades por zonas inestables de la señal, cuestión que provoca una calidad de síntesis pobre.

2.1.2. Conversión de texto en habla basada en selección de unidades

Durante años, la tecnología de síntesis basada en la concatenación de difonemas (y por extensión de trifonemas), fue la más empleada para desarrollar sistemas de CTH. El problema fundamental de estos sistemas recae en utilizar un corpus que contiene una única

realización por unidad, es decir, sólo hay grabada una versión de cada una de las unidades del idioma de síntesis. Normalmente, estas unidades se graban utilizando frases portadoras, de las que se selecciona la unidad deseada, o mediante palabras vacías o logatomos (Black y Lenzo, 2001b). Los problemas fundamentales de esta tecnología son dos:

1. **Modificación prosódica:** la señal de voz grabada debe adaptarse a los requisitos prosódicos indicados por el bloque de PLN. Si éstos no difieren mucho de la prosodia real (grabada) de la señal, la calidad sintética será bastante buena. En cambio, si ésto no es así (debido a que las técnicas de etiquetado y modificación prosódica no son perfectas o bien por la ubicación de la unidad en un contexto muy distinto al que se grabó), los importantes cambios prosódicos que tienen que sufrir las unidades pueden provocar un descenso muy importante de la naturalidad de la señal sintética generada. De ahí que sea interesante disponer de una gran variedad de realizaciones de cada unidad, en diferentes contextos prosódicos, lingüísticos, fonéticos, etc., así como de la optimización de las técnicas de etiquetado y modelado prosódicos (ver capítulo 4).

2. **Concatenación de unidades:** durante el proceso de síntesis (bloque de PDS), una vez modificadas prosódicamente, se procede a unir las unidades entre sí. Debido a que sólo se dispone de una realización por unidad, existirán tantos puntos de concatenación como número de unidades a sintetizar menos una (p.ej. si se trata de tres unidades, existen dos puntos de unión). Asimismo, estas uniones serán artificiales puesto que las unidades se han grabado individualmente en el corpus de voz. Por muy bueno que sea el algoritmo de concatenación (LPC, TD-PSOLA, MBROLA, HNM, etc.) o la técnica de concatenación utilizada (Conkie y Isard, 1996; Guaus et al., 1998), las uniones de las tramas de las señales de voz de las unidades nunca conseguirán ser tan naturales como el proceso fisiológico que las genera secuencialmente, provocando la aparición de discontinuidades espectrales.

No obstante, a pesar de las múltiples mejoras introducidas en el proceso de concatenación de unidades a partir la reducción de las discontinuidades espectrales en el punto de concatenación mediante un diseño refinado del corpus de voz, los CTH basados en difonemas no consiguieron obtener la naturalidad sintética deseada, debido al elevado número de puntos de concatenación existentes (Möbius, 2000). Por este motivo, y gracias al aumento de la capacidad de memoria y de potencia en el cálculo de los ordenadores, durante la pasada década, la investigación en síntesis del habla dio un paso hacia delante. Concretamente, se empezó a trabajar con corpus de voz de mayor tamaño que los utilizados en CTH basada en difonemas, incorporando múltiples realizaciones de las unidades básicas de síntesis. En éste ámbito, los primeros trabajos fueron desarrollados por el grupo de ATR (*Advanced Telecommunications Research Institute International*), presentando el primer sistema de síntesis basado en unidades de tamaño variable (no uniformes) a principios de los 90 (Sagisaka, 1988; Takeda, Katsuo y Sagisaka, 1990; Sagisaka et al., 1992). En la misma institución, y siguiendo una línea de trabajo paralela, se desarrolló el trabajo que dio nombre a la nueva estrategia de CTH: la CTH basada en *selección de unidades* o CTH-SU (Black y Campbell,

1995; Hunt y Black, 1996). Uno de los primeros sistemas basados en esta estrategia fue el CHATR (Black y Taylor, 1994), que se basa en la concatenación directa de las unidades disponibles en el corpus, sin incorporar modificación de la señal (prosódica ni en el punto de concatenación), enfoque utilizado por otros sistemas posteriores[2], p.ej. en (Chu et al., 2001; Peng, Zhao y Chu, 2002), y hace un par de años, se presentó una nueva versión del sistema denominado XIMERA en el trabajo de (Kawai et al., 2004). A partir de estos trabajos, se definieron las características fundamentales de la síntesis basada en selección de unidades, que son:

1. Disponer de un corpus de voz con un gran número de realizaciones para cada una de las unidades mínimas consideradas (fonemas, difonemas, etc.), con la consecuente diversidad prosódica y lingüística de las unidades.

2. Seleccionar la secuencia de unidades del corpus que mejor se ajuste a las características prosódicas y lingüísticas de la secuencia de unidades a sintetizar (información obtenida del bloque de PLN) en tiempo de ejecución.

3. Minimizar el número de puntos de concatenación y la necesidad de modificación prosódica de la señal, reduciendo la sensación de artificialidad de la señal sintética.

De ahí que la arquitectura de un CTH-SU disponga de un corpus de voz de mayor tamaño que la de un CTH basado en difonemas, y, en consecuencia, incorpore un módulo encargado de seleccionar la cadena *óptima* de unidades en tiempo de ejecución (ver figura 2.2). En este contexto la selección de las unidades más adecuadas para la síntesis pasa de ser un proceso *off-line* (en la técnica basada en difonemas se escogen las mejores realizaciones que deben conformar el corpus durante su diseño), a un proceso *on-line* (en síntesis basada en corpus, la mejor secuencia de unidades se escoge en tiempo de ejecución).

Resumen de la evolución de los sistemas de CTH-SU

La mayoría de los sistemas de conversión de texto en habla actuales utilizan una estrategia de síntesis concatenativa basada en la selección de las mejores unidades del corpus de voz según unos determinados parámetros de búsqueda. Durante la última década han aparecido distintas aproximaciones que abordan el problema de la síntesis basada en corpus. Los trabajos iniciados en ATR definieron las bases de esta tecnología (Sagisaka, 1988; Takeda, Katsuo y Sagisaka, 1990; Sagisaka et al., 1992), y fueron pronto seguidos por la mayor parte de la comunidad científica. Entre otros, destacan los trabajos desarrollados en el CSTR (*Centre for Speech Technology Research*) de la Universidad de Edinburgo (Black y Taylor, 1997a; Hofer, Richmond y Clark, 2005; Clark, Richmond y King, 2005), el LTI (*Language Technologies Institute*) de la Universidad Carnegie Mellon (Rudnicky et al., 2000; Black y Lenzo, 2001b; Black, 2002; Black, 2003; Black y Lenzo, 2004; Black y Tokuda, 2005), el MIT

[2]Este enfoque parece viable si las unidades recuperadas del corpus son muy cercanas a la secuencia objetivo, cuestión que sólo se consigue cuando la cobertura del corpus es óptima para la secuencia de unidades deseada.

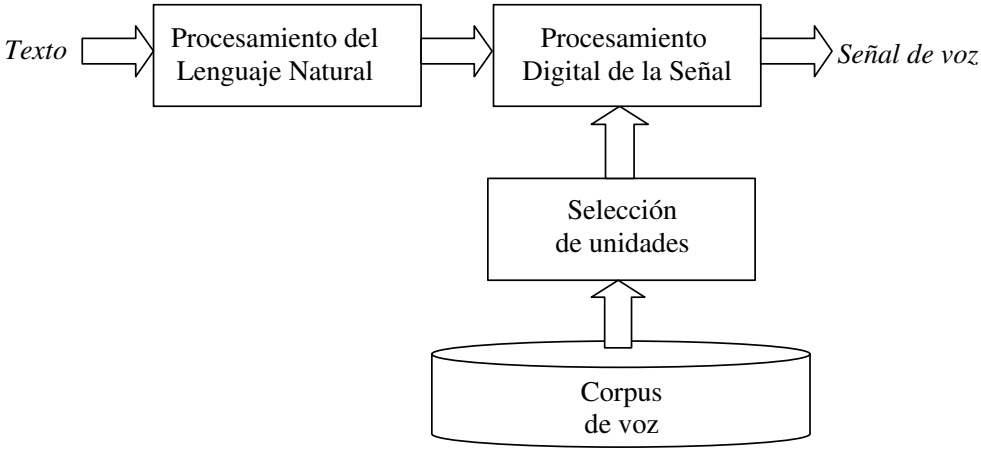


Figura 2.2: Diagrama de bloques de un conversor de texto en habla basado en selección de unidades.

(*Massachussets Institute of Technology*) (Yi y Glass, 1998; Yi y Glass, 2002; Yi, 2003), o el Laboratorio TCTS (*Théorie des Circuits et Traitement du Signal*) de la Faculté Polytechnique de Mons (Dutoit, 1997; Dutoit y Stylianou, 1997), entre muchos otros.

A nivel nacional[3], cabe destacar el grupo TALP (Centre de Tecnologies i Aplicacions del Llenguatge i la Parla) de la Universitat Politècnica de Catalunya (Bonafonte et al., 1998; Febrer y Bonafonte, 2000; Febrer, 2001) que ha desarrollado sistemas de CTH-SU para el castellano y el catalán, así como el Departamento de Teoría de la Señal y Comunicaciones de la Universidad de Vigo, que partiendo del trabajo del TALP ha continuado su línea de investigación con el sistema Cotovía (Campillo y Rodríguez Banga, 2002; Rodríguez Banga et al., 2002; Campillo, 2005),

Nota aparte merece el *Festival Speech Synthesis System* (Black y Taylor, 1997b; Taylor, Black y Caley, 1998; Taylor, Black y Caley, 2000-2003). Esta plataforma ofrece un conjunto de herramientas para la creación de corpus de voz, el desarrollo y mejora de los módulos de un CTH, etc. Sobre esta plataforma se han implementado varios CTH para distintos idiomas, entre otros: italiano (Cosi et al., 2001), alemán (*http://www.ims.uni-stuttgart.de/phonetik/synthesis/*), e incluso, Tegulú (Vepa, Ayachitam y Reddy, 2002). Posteriormente se desarrolló una versión de Festival reducida para sistemas portables, denominada Flite (*Festival lite*) (Black y Lenzo, 2001a) (*http://www.speech.cs.cmu.edu/flite/*), y más recientemente se ha presentado una nueva versión de Festival, más orientada a la selección de unidades (Clark, Richmond y King, 2004). Toda la información referente a esta plataforma se puede encontrar en la dirección *http://festvox.org/*.

Además de la investigación en el ámbito académico, el mundo empresarial también tomó la tecnología de la selección de unidades para desarrollar sus nuevos CTH. A con-

[3]Existen multitud de aportaciones de la comunidad científica nacional en el ámbito de la síntesis del habla, pero en este trabajo el autor se centra, fundamentalmente, en las aportaciones a la CTH-SU.

tinuación, y a modo de resumen (sin querer ser un repaso exhaustivo al estado de las tecnologías del habla en el ámbito empresarial) se presentan algunos de las aportaciones comerciales más significativas para el desarrollo de los sistemas de CTH-SU desde el ámbito empresarial. De entre los múltiples sistemas comerciales desarrollados en la última década, cabe destacar el CTH-SU de AT&T (Natural Voices) descrito inicialmente en los trabajos de (Beutnagel, Conkie y Syrdal, 1998; Beutnagel et al., 1999; Beutnagel, Mohri y Riley, 1999; Beutnagel y Conkie, 1999; Conkie, 1999), actualizado en (Conkie et al., 2000; Syrdal et al., 2000) y adaptado posteriormente al alemán en (Jilka y Syrdal, 2002). Asimismo, empresas como IBM también han apostado por esta tecnología (Donovan y Eide, 1998; Donovan et al., 2001; Hamza y Donovan, 2002; Eide et al., 2003; Hamza et al., 2004), así como British Telecom con el sistema Laureate (Breen y Jackson, 1998), Microsoft con WHISTLER (*Windows Highly Intelligent STochastic taLkER*) (Huang et al., 1996; Huang et al., 1997; Hon et al., 1998; Chu et al., 2001; Huang, Acero y Hon, 2001), el sistema de RealSpeak de ScanSoft (antigua Lernout & Hauspie) (Rutten et al., 2000; Coorman et al., 2000), o el sistema multilingüe ACTOR® de la empresa Loquendo (Quazza et al., 2001; Tesser et al., 2005) (surgida del grupo CSELT (*Centro Studi e Laboratori Telecomunicazioni*) (Balestri et al., 1999)), que el octubre de 2002, en colaboración con el Departamento de Filología Española de la Universidad Autónoma de Barcelona (UAB) presentó a 'Montserrat', un CTH basado en selección de unidades desarrollado para el catalán de gran naturalidad (posteriormente acompañado por 'Jordi', la voz masculina en catalán). Cabe añadir la empresa Cepstral y el grupo Acapela, consorcio formado por Babel Technologies, Infovox y Elan Speech, uno de los más importantes a nivel europeo, ya que aporta soluciones en tecnologías del habla para 23 idiomas. Por otro lado, a nivel nacional destacar la empresa Telefónica I+D, con CTH para todos los idiomas oficiales del estado, junto a la empresa ATLAS (*Applied Technologies on Language and Speech*) con el sistema Verbio para catalán, castellano y portugués. Añadir que recientemente se ha establecido un marco de competición para la evaluación de la calidad de los sistemas de conversión de texto en habla denominado *Blizzard Challenge* cuya primera edición se realizó recientemente (Black y Tokuda, 2005).

Elección del tipo de unidad

La elección de la unidad básica del sistema de síntesis concatenativa es uno de los elementos importantes en el diseño de los sistemas de CTH-SU. Tanto trabajar con unidades básicas de pequeño o gran tamaño (p.ej. estados de un modelo de Markov, semifonemas, fonemas, difonemas, sílabas, palabras,...) presenta puntos a favor y en contra (Chu et al., 2001). Por un lado, resulta *sencillo* diseñar un corpus que contenga una buena cobertura prosódica y espectral de las unidades de tamaño pequeño, mientras que esto se convierte en una tarea mucha más complicada —o prácticamente imposible— cuando se trabaja con unidades de mayor tamaño, como por ejemplo palabras. Por lo contrario, si se trabaja con unidades de tamaño pequeño, el número de concatenaciones en la señal sintética aumenta respecto a trabajar con unidades de tamaño mayor, con una degradación potencial mayor (Chu et al., 2001; Black, 2002).

Como ya se ha comentado, el sistema de CTH para el catalán desarrollado en el área de

Tecnologías del Habla durante los años 90 estaba basado en la concatenación de difonemas y trifonemas (1207 unidades en total), como la mayoría de sistemas de síntesis de la época. En cambio, la introducción de la CTH-SU volvió a abrir el debate sobre el tipo de unidad mínima a considerar durante el proceso de grabación del corpus de voz. Aunque la síntesis por selección de unidades permite encontrar cadenas de unidades largas (sílabas, palabras, o incluso frases enteras), es importante determinar cuál es la entidad mínima que compone la secuencia objetivo a buscar dentro del corpus. Esta elección influirá decisivamente en el tamaño del corpus de voz a grabar: cuánto más larga sea la unidad escogida, mayor tendrá que ser el tamaño del corpus para obtener una buena variabilidad (prosódica y lingüística) de la unidad escogida (Beutnagel et al., 1999). Por ejemplo, si se trabaja a nivel de fonema, resultará necesario grabar menos expresiones que si se trabaja a nivel de palabra (todas las del idioma, con todas las variantes deseadas).

De ahí que los primeros sistemas de CTH-SU optaran por trabajar con el *fonema* como unidad mínima (Black y Campbell, 1995; Black y Taylor, 1997a). De este modo, se reduce considerablemente el espacio de búsqueda (p.ej. para el catalán, 37 fonemas), permitiendo que el corpus de voz contenga muchas variantes de las unidades mínimas (robustez estadística). Pero, por otro lado, este tipo de unidad tiene el problema que los límites de la unidad (inicio y fin del fonema) presentan variaciones bruscas en la forma de onda, cosa que dificulta su concatenación, aspecto decisivo en la calidad final de la señal sintética.

Para evitar los problemas de la concatenación a nivel de fonemas (que provocan inconsistencias en la calidad de los sistemas de CTH-SU), se optó por trabajar con medios fonemas (del inglés *half-phones*) (Conkie, 1999; Beutnagel et al., 1999) o también denominados semifonemas (Febrer, 2001; Campillo, 2005) o demi-fonemas (Beutnagel et al., 1999). En este caso el fonema se divide en dos medios fonemas: la parte izquierda, que va desde el inicio del fonema hasta su parte central, o zona estable, y la parte derecha, que contiene el resto del fonema. Por ejemplo, un fonema cualquiera $/x/$ quedará dividido en $/x_i/$ (izquierda) y $/x_d/$ (derecha). Por lo tanto, para cada fonema aparecen dos medios fonemas, doblándose el número de unidades a considerar en el corpus de voz. Sin embargo, se añade la posibilidad de realizar uniones *suaves* por la zona más estable de la señal de voz (la mitad del fonema), aunque sin evitar las uniones por zonas de transición.

Existen otras aproximaciones que trabajan con unidades de síntesis a nivel subfonético, como los estados de un modelo oculto de Markov que modelan la señal de voz, sistema empleado por IBM (Donovan et al., 2001), y denominados como *senones* por Microsoft (Huang et al., 1997).

El problema de todas estas unidades se encuentra en el hecho que se pueden continuar produciendo concatenaciones *duras* (en zonas poco estables de la señal) durante la generación de la señal sintética, aun cuando se tomen en consideración procesos de selección del punto óptimo de la unión (Conkie y Isard, 1996). Por este motivo, nuestro grupo empezó a trabajar en un sistema de selección de unidades basado en difonemas y trifonemas (Guaus y Iriondo, 2000a; Guaus y Iriondo, 2000b), opción también contemplada en otros trabajos (Breen y Jackson, 1998; Beutnagel, Conkie y Syrdal, 1998; Coorman et al., 2000; Toda et al., 2002; Clark, Richmond y King, 2005). De este modo, el sistema de CTH basado en

selección de unidades, presentará, como mínimo, la misma calidad en la concatenación que la de un sistema de CTH basado en difonemas y trifonemas. Esto sucederá sólo cuando el proceso de búsqueda de unidades no sea capaz de encontrar dos unidades (difonemas o trifonemas) consecutivas en el corpus para la síntesis.

No obstante, sea cual sea la unidad mínima escogida, resulta necesario definir un corpus de voz que contenga todas las unidades del idioma de trabajo, introduciendo la variabilidad necesaria para dar respuesta a las particularidades de la estrategia de síntesis basada en selección de unidades. Con este objetivo en mente, en el seno del grupo se decidió diseñar el corpus de voz dividiéndolo en dos partes (Guaus y Iriondo, 2000a) distintas, en la línea de otros sistemas, como por ejemplo (Blouin et al., 2002; Campillo, 2005), pero incluyendo algunas variantes. Concretamente, para el caso de trabajar con los difonemas y trifonemas como unidades básicas, el corpus se estructuró de la siguiente forma:

- **Unidades básicas:** provienen del concepto de sistema de síntesis basado en difonemas y trifonemas: serán 895 difonemas y 312 trifonemas (1207 unidades). Así, el sistema de síntesis siempre encontrará, como mínimo, una realización de la unidad (difonema o trifonema) a sintetizar.

- **Realizaciones de las unidades básicas:** formada por las variantes prosódicas y lingüísticas de las unidades básicas. Al tratarse de difonemas y trifonemas, asegurar una buena cobertura de sus variantes provoca un aumento considerable del tamaño del corpus respecto al de unidades de menor tamaño. Así, pues, el módulo de síntesis, a partir de la indexación sobre la lista de unidades básicas buscará en la base de datos la mejor secuencia de unidades para ser sintetizada.

Dimensionado del corpus de voz

El dimensionado del corpus de voz es uno de los factores clave en el diseño de un corpus para síntesis basada en selección de unidades. Es uno de los factores que rige la calidad alcanzada por la señal de voz sintética. Así pues, ha sido y continúa siendo una de las líneas de investigación más importantes para este tipo de sistemas. Existen distintas aproximaciones en el diseño de los corpus de voz para síntesis basada en selección de unidades. Estos corpus deben estar diseñados con el fin de incluir todas las variantes más relevantes de las realizaciones del idioma o el ámbito de aplicación (*dominio*). Las primeras aproximaciones trabajaban con corpus de duración inferior a la hora, como los de IBM (Donovan y Eide, 1998) y el sistema CHATR, introducido por (Black y Taylor, 1994), de unos 45 min de duración (Möbius, 2000). Posteriormente, se optó por corpus de duraciones superiores (varias horas) como en el caso del corpus de AT&T (Conkie, 1999; Beutnagel et al., 1999), uno de los primeros sistemas completos de síntesis de voz con las mejores prestaciones del momento, la propia IBM con un corpus de unas 3 horas (Donovan et al., 2001). Posteriormente, se ha trabajado con 15 horas para el chino en Microsoft (Chu et al., 2001) o para el inglés en IBM (Fischer, Botella y Kunzmann, 2004), e incluso en ATR se ha llegado a las 60 o 100 horas para el japonés en el CTH XIMERA (Kawai et al., 2004), dimensiones ya habituales

para los sistemas comerciales (Campbell, 2005). Bajo este enfoque, cuanta más voz se haya almacenado, mejor calidad se podrá obtener. Desde otro punto de vista, también se ha trabajado en busca de una dimensión *óptima* del corpus de voz. Es decir, se defiende que no porque el corpus contenga un mayor número de unidades, la síntesis será mejor, en términos de coste computacional y en estabilidad de la calidad de la señal generada. Existen distintos estudios sobre el tema, pero hasta el momento no se ha llegado a ninguna conclusión firme. Estudios como los de AT&T (ver (Möbius, 2000) para un resumen) indican que el corpus tendría que ser lo mayor posible, pero el problema recae en la imposibilidad de conseguir la cobertura suficiente para cubrir todos los fenómenos lingüísticos de un idioma (Black, 2002). Por otra parte, existen otros trabajos (Guaus y Iriondo, 2000a; Kawai et al., 2004) donde se argumenta que no es imprescindible disponer de corpus de tamaños enormes para sistemas de CTH basados en esta estrategia de síntesis, ya que se llega a la saturación en la mejora de la calidad esperada o la variabilidad de las unidades recuperadas mediante la función de coste. Así pues, por un lado, corpus relativamente pequeños son más tratables computacionalmente y, por el otro, parece claro que la calidad de la síntesis, aunque crece al aumentar el tamaño del corpus, no lo hace siguiendo una proporción lineal. Por lo tanto, puede ser interesante determinar un umbral por encima del cual el aumento del coste computacional provocado por el incremento del número de unidades de la base no compense la mejora en la calidad final de la síntesis.

Por otra parte, el diseño del corpus de voz es otra de las cuestiones críticas a considerar, como se describe en (van Santen, 1997; Batůšek, 2001; Black, 2002), ya que la selección óptima del conjunto de textos a grabar resulta fundamental (Iida y Campbell, 2001). En el trabajo de van Santen (1997), se generó un vector de parámetros para caracterizar las posibles variantes de cada unidad (en este caso, el difonema) considerada en su CTH a partir de su módulo de PLN. El trabajo se realizó sobre un gran corpus de textos periodísticos, analizado mediante un conjunto de test de textos genéricos. A pesar de caracterizar cada unidad con un número reducido de parámetros y de posibles valores de estos parámetros, el estudio obtuvo un bajo índice de cobertura (variabilidad) del corpus ante los textos de test según el vector de parámetros definido. Según el autor, este problema aumenta cuando el género (o *estilo*) del texto de test no se ajusta al considerado en el diseño del corpus. Por otra parte, en el estudio de Batůšek (2001) se discute qué parámetros hay que considerar al caracterizar las unidades en el proceso de diseño del corpus, con el fin de obtener una buena cobertura del corpus (según un determinado umbral). En este trabajo se defiende el hecho de que se deben generar vectores lo más completos posibles (con información prosódica y fonética) cosa que parece contradecir las conclusiones del trabajo de van Santen (1997).

Recientemente, se ha desarrollado un corpus de voz de referencia, siguiendo la filosofía del proyecto Festival (Black y Taylor, 1997b; Taylor, Black y Caley, 2000-2003). Este corpus, denominado ARCTIC, se ha desarrollado en la Unversidad Carnegie Mellon (Kominek y Black, 2003; Kominek y Black, 2004) con el propósito de empezar a disponer de corpus comunes para que la comunidad científica pueda realizar comparaciones válidas entre sistemas de síntesis del habla. Hasta el momento, el corpus solo dispone de voces en inglés, cada una de ellas de duración alrededor de la hora (Kominek y Black, 2003). Siguiendo la filosofía de competición de DARPA para los sistemas de reconocimiento, durante el año 2005 se

ha llevado a cabo la primera competición de sistemas de síntesis (*The Blizzard Challenge)* (Black y Tokuda, 2005) sobre el corpus CMU ARCTIC. Después de presentar los resultados de los sistemas a competición, se ha decidido aumentar el tamaño del corpus y empezar a trabajar en la línea de incluir nuevos idiomas a medio plazo.

Así pues, parece claro que las cuestiones que se refieren al tamaño y al diseño del corpus de voz continúan siendo uno de los puntos clave en este ámbito de investigación, ya que se trata de uno de los puntos críticos de todo sistema de síntesis basado en corpus.

Optimización del proceso de selección

Otra de las líneas de investigación en el ámbito de los sistemas de conversión de texto en habla basados en selección de unidades ha centrado en la optimización del proceso de selección propiamente dicho. Fundamentalmente, se ha trabajado en la reducción del coste computacional del proceso de búsqueda de unidades, con el objetivo de acelerar el proceso pero sin perder calidad en la síntesis (Conkie et al., 2000). Existen distintas estrategias para la reducción de este coste computacional. A continuación se describen algunas de ellas de forma breve:

- Limitar el número de unidades candidatas que intervienen en el proceso de selección. Estas unidades se agrupan en distintos conjuntos mediante árboles de decisión, de forma que la selección de unidad se puede realizar de forma jerárquica y la selección dinámica de unidades se restringe al subconjunto elegido, reduciendo el coste del proceso de selección (Black y Taylor, 1997a; Macon, Cronk y Wouters, 1998; Donovan, 2000).

- Precalcular parte de los costes de concatenación antes de realizar el proceso de búsqueda (Beutnagel, Mohri y Riley, 1999; Conkie et al., 2000). Según este método, se genera una memoria *caché* de las uniones más habituales y así se acelera el proceso de selección, manteniendo una calidad similar a la obtenida de la búsqueda global.

- Dejar de considerar posibles candidatos dentro del espacio de búsqueda mediante la poda de caminos (*pruning*, en inglés) (Black y Taylor, 1997a; Febrer, 2001), p.ej. por ser demasiado similares.

- Eliminar redundancias y *outliers* en cada grupo de unidades (Febrer, 2001; Kim, Lee y Hirose, 2001; Hamza y Donovan, 2002). De esta manera, se dejan de considerar aquellas unidades que no añaden variabilidad al corpus (redundancias), o bien, que tienen valores muy alejados de la media (*outliers*). Este segundo caso es más crítico, porque es necesario vigilar de no eliminar aquellas unidades que aparecen en contextos poco habituales y que están correctamente etiquetadas.

2.1.3. Módulo de selección de unidades

En el ámbito de los sistemas de CTH-SU, el proceso de selección de la mejor unidad pasa de ser un proceso *off-line* (previo a la conversión de texto en habla) a ser *on-line* (durante la síntesis). Es decir, en los CTH basados en difonemas se tenía que escoger, antes de su grabación, las características que tenían que presentar las unidades (normalmente grabadas en un tono constante), y tras varias grabaciones, se escogían las mejores. Aun así, se podía dar la situación de tener que volver a grabar algunas de las unidades si, una vez realizadas las primeras pruebas de síntesis, provocaban errores en la señal sintética. En cambio, en el caso de la estrategia de selección de unidades, a pesar de tener que diseñar también el corpus según unos determinados criterios de cobertura prosódica y lingüística del idioma o dominio considerado, la tarea de seleccionar las mejores unidades acústicas se realiza durante el propio proceso de síntesis. De esta tarea se encarga el *nuevo* módulo que se incorpora en la arquitectura del CTH entre el bloque de PDS y el corpus de voz, denominado en la figura 2.2 como módulo de "*Selección de unidades*".

Este módulo desarrolla un papel fundamental dentro de la estrategia de síntesis concatenativa basada en corpus, puesto que es el encargado de seleccionar las unidades del corpus que mejor se ajusten a las características del texto a sintetizar: la secuencia objetivo (en inglés, *target sequence*). Dada la trascripción fonética del texto a sintetizar y los parámetros prosódicos asociados (definidos por el módulo de PLN), el algoritmo de selección debe escoger la mejor realización de cada unidad de la secuencia a sintetizar. En este caso, las unidades del corpus se representan como una red de estados de transición con unos costes de ocupación de estado y de transición asociados, al estilo de un sistema de reconocimiento de voz basado en modelos ocultos de Markov (HMM) (Hunt y Black, 1996). Por este motivo, también se aplicará una búsqueda dinámica del *camino óptimo* a lo largo de la red; concretamente se suele aplicar el algoritmo de Viterbi (1967), adaptado para reconocimiento del habla por (Rabiner y Juang, 1993). La diferencia fundamental de este proceso reside en que, para reconocimiento, se decide la secuencia de estados de mayor probabilidad según unos modelos acústicos de unidades junto a la gramática utilizada por el sistema de reconocimiento. En cambio, para selección de unidades, se busca la secuencia de unidades del corpus que presente un menor coste, según la función de coste de selección. Esta función toma en consideración el grado de similitud entre la unidad candidata (u_i) y la unidad objetivo (*target*, t_i) mediante el *coste de unidad* (coste de ocupación de estado en la red) y el grado de continuidad entre las unidades consecutivas (u_{i-1}, u_i) a través del *coste de concatenación* (coste de transición en la red) (Hunt y Black, 1996).

En la figura 2.3 se presenta un ejemplo de la red de posibles caminos a considerar en el proceso de búsqueda de las unidades candidatas del corpus de voz para la síntesis de un determinado texto de entrada. En ella se indica el coste de unidad entre la unidad objetivo t_i y la unidad candidata u_i (C^t, definido en la ecuación (2.1)) y el de concatenación entre las unidades consecutivas u_{i-1} y u_i (C^c, en la ecuación (2.2)), según lo descrito en (Hunt y Black, 1996).

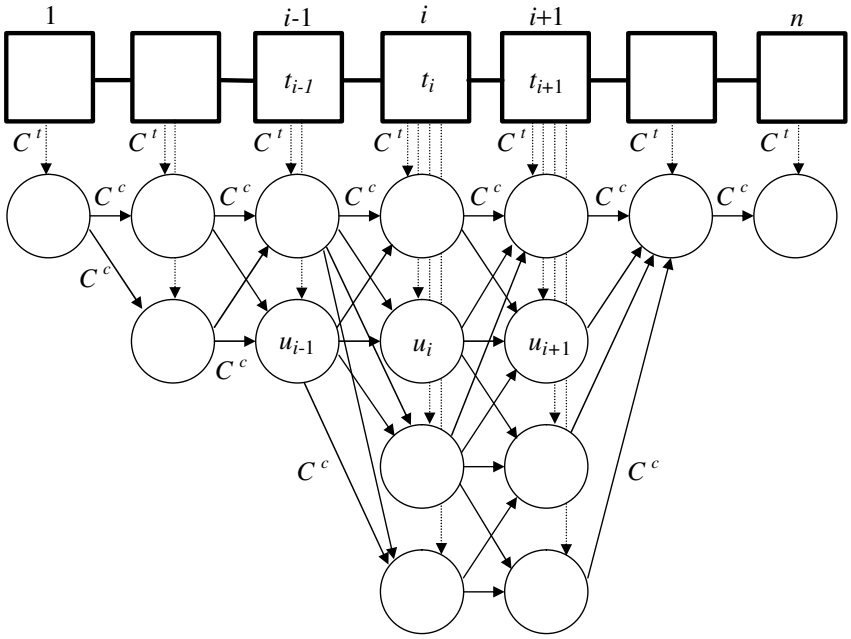


Figura 2.3: Búsqueda de la secuencia objetivo (t) de n unidades dentro de la red de unidades candidatas (u). Las flechas discontinuas representan el coste de unidad (C^t) y las continuas, el coste de concatenación (C^c).

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^u(t_i, u_i) \tag{2.1}$$

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i) \tag{2.2}$$

El coste de unidad y el coste de concatenación se definen como una suma ponderada de p y q subcostes, ecuaciones (2.1) y (2.2) respectivamente. Estas medidas ponderarán la importancia de las diferencias entre la secuencia de unidades objetivo y unidades candidatas en términos de información prosódica (duración, energía y frecuencia fundamental), fonética (trascripción y contexto fonético), fonológica (posición en la frase, palabra, acento, etc.), entre otras. Estos requisitos son suministrados por el módulo de PLN del CTH. Así pues, el algoritmo de selección buscará minimizar la función de coste que se obtiene de la combinación lineal de C^t y C^c para las n unidades que forman la secuencia a sintetizar, definida según la ecuación (2.3) (Hunt y Black, 1996).

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i) \tag{2.3}$$

donde t_1^n representa la secuencia de uniades objetivo $\{t_1, t_2, \ldots, t_n\}$ y u_1^n representa la secuencia de uniades candidatas $\{u_1, u_2, \ldots, u_n\}$.

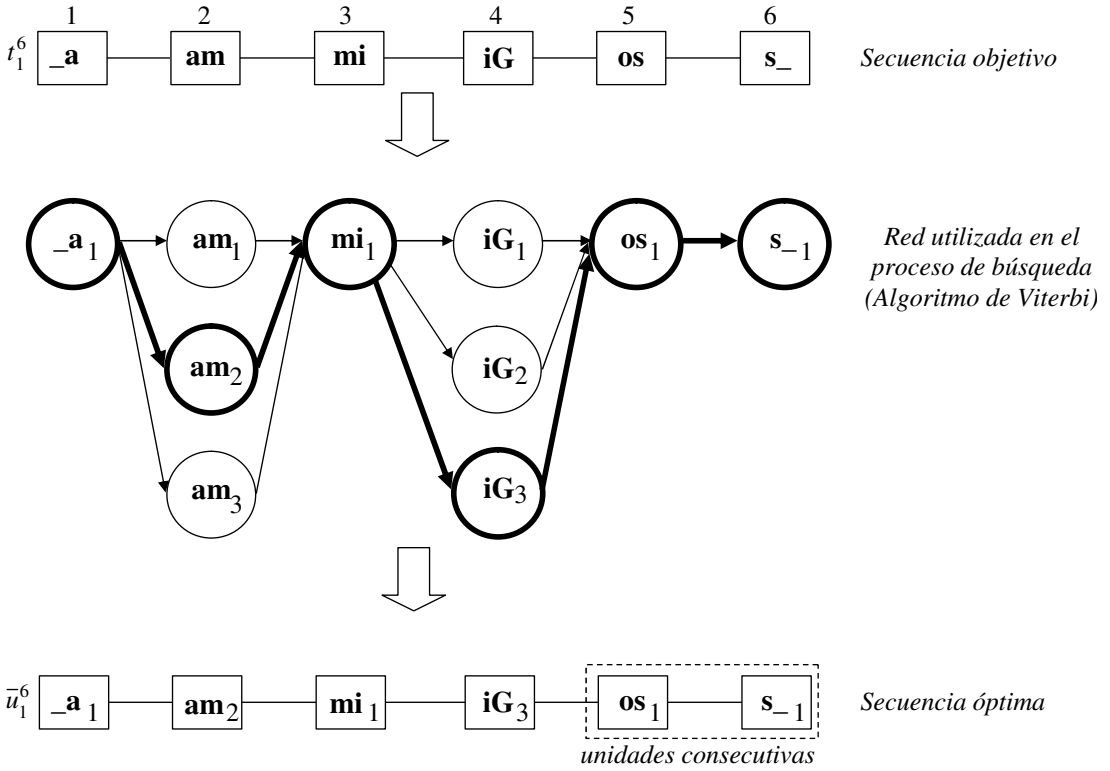


Figura 2.4: Ejemplo del proceso de selección para la palabra *amigos*, fonéticamente transcrita por /_amiGos_/ (notación SAMPA). En el ejemplo, algunas unidades —en este caso, difonemas— presentan una sola realización en el corpus, mientras que otras tienen tres.

La figura 2.4 presenta el proceso de selección de unidades sobre una palabra desglosada en difonemas. El resultado es la secuencia *óptima* de unidades del corpus (u_1^{n*}) que minimiza el coste acumulado $C(t_1^n, u_1^n)$ (ecuación (2.3)), a lo largo de toda la red (unidades t_i y u_i en la figura 2.3). El principal objetivo de la selección de unidades pasa por escoger la secuencia de unidades con el mínimo coste de unidad y de concatenación respecto a la secuencia de unidades objetivo t_1^n (Jilka y Syrdal, 2002), a partir de la ecuación (2.4) (Hunt y Black, 1996).

$$u_1^{n*} = \overset{Argmin}{u_1, \ldots, u_n} C(t_1^n, u_1^n) \tag{2.4}$$

En este pequeño ejemplo (con muy pocas realizaciones por unidad), se supone que aparece una concatenación natural (*unidades consecutivas* en la figura 2.4). Este es uno de los aspectos que debe potenciar el proceso de selección, con el objetivo de minimizar el número de puntos de concatenación en la señal sintética. A continuación, se describen

las técnicas utilizadas para el entrenamiento y ajuste de los subcostes de selección y de los pesos que los ponderan, elementos fundamentales en el proceso de selección, y por lo tanto, elementos críticos para obtener una señal sintética de máxima calidad (naturalidad).

2.1.4. Subcostes de selección

La definición de los subcostes a utilizar en la función de coste (ecuaciones (2.1) y (2.2)), así como los parámetros que éstos consideran (información que se extrae de las unidades del corpus mediante un vector de características multidimensional (Peng, Schuurmans y Wang, 2003)), constituyen otra de las líneas de investigación más importantes en el ámbito de la CTH basada en selección de unidades (Black, 2002). Existen muchas variantes sobre las medidas a utilizar: desde considerar sólo información simbólica mediante árboles fonéticos (Breen y Jackson, 1998; Black y Taylor, 1997a) o estructuras fonológicas (Taylor y Black, 1999), pasando por utilizar distancias simbólicas, escalares o vectoriales (Coorman et al., 2000) o funciones matemáticas más complejas, como la sigmoidea (Febrer, 2001; Toda, Kawai y Tsuzaki, 2004), e incluso modelando los subcostes mediante técnicas de aprendizaje artificial (Campillo, 2005). No obstante, sea cual sea el tipo de función de coste utilizado, resulta necesario ponderar la importancia de unos subcostes frente a otros en el proceso de selección de unidades.

Según el vector de parámetros o representación escogida para describir las unidades, existen diferentes tipos de medidas para evaluar los subcostes (de unidad o de concatenación) descritos en las ecuaciones (2.1) y (2.2). De este modo se determina el grado de semejanza entre la unidad objetivo (t_i) y la unidad candidata (u_i) o bien, el grado de similitud entre unidades consecutivas de la red (u_{i-1} y u_i). Según (Coorman et al., 2000), se pueden definir tres tipos de distancias, según los parámetros que se quiera comparar:

1. **Simbólica:** determina el grado de semejanza entre parámetros difícilmente cuantificables del vector de características de las unidades. Por ejemplo, si se está comparando el *acento* de la unidad (C_t), la distancia sólo podrá ser '0', si ambas unidades están acentuadas o no, o '1' si no coinciden en esta característica (Febrer, 2001; Campillo, 2005). No obstante, una medida simbólica no tiene porqué ser sólo binaria: puede dar valores intermedios, como al evaluar el contexto fonético de la unidad (C_c) (p.ej. (Toda et al., 2002)). Por ejemplo, si se desea encontrar una unidad con un fonema oclusivo sordo como vecino, se debe puntuar con un coste menor aquella unidad candidata que tenga una oclusiva sonora que la que presente una fricativa sorda, siguiendo una medida que codifique el grado de parecido sonoro entre unidades. Así pues, este tipo de distancia está íntimamente ligada al conocimiento experto de la característica a evaluar.

2. **Escalar:** son medidas que comparan parámetros cuantificables obtenidos de las unidades. Por ejemplo, la diferencia en el tono de las unidades, en la duración o en la energía, entre otras. Se pueden calcular como la diferencia absoluta entre los parámetros, cuadrática, etc., o bien, penalizando la diferencia en un cierto sentido (p.ej. en

la duración se puede penalizar más que la unidad candidata presente una duración menor a la deseada). El trabajo de (Febrer, 2001) presenta una aportación para la medida de las distancias mediante funciones escalares continuas y discretas (definidas por tramos).

3. **Vectorial:** es una distancia definida entre informaciones generalmente cuantificables de las unidades representadas en forma de vectores multidimensionales. Por ejemplo, en el cálculo del C^c para evaluar la continuidad espectral se suelen modelar las tramas extremas de las unidades mediante parámetros cepstrales en la escala Mel (*Mel Frequency Cepstrum Coefficients* o MFCC), *Line Spectral Frequencies* (LSF) o *Line Spectral Pairs* (LSP), entre otros.

Uno de los principales inconvenientes a la hora definir el coste de concatenación es la falta de una medida objetiva que se corresponda con la sensación subjetiva de la continuidad en el habla. En esta línea, se han llevado a cabo muchos estudios y aproximaciones con el fin de buscar una distancia acústica que tome en consideración la subjetividad humana. Se han aplicado distancias como la Euclídea o la de Mahalanobis (Donovan, 2001), la de Kullback-Leiber (Veldhuis y Klabbers, 2003), la de Itakura-Saito (Rabiner y Juang, 1993), entre otras. También se han tomado en consideración distintos parámetros dentro de estas distancias, como los cepstrum (normalmente MFCC) (Black y Campbell, 1995; Tsuzaki y Hisashi, 2002; Campillo y Rodríguez Banga, 2002), los LPC (Macon, Cronk y Wouters, 1998), información de los formantes (Ding y Campbell, 1997), LSF (Vepa, King y Taylor, 2002) o LSP (Wu y Chen, 2001), etc. Además, se han llevado a cabo muchos trabajos estudiando la correlación de estas medidas objetivas con la respuesta subjetiva de los usuarios. Destacan los experimentos de (Klabbers y Veldhuis, 1998; Macon, Cronk y Wouters, 1998) actualizados en (Klabbers y Veldhuis, 2001; Wouters y Macon, 2001), los de (Donovan, 2001; Stylianou y Syrdal, 2001) o, posteriormente, los de (Vepa, King y Taylor, 2002; Peng, Zhao y Chu, 2002; Tsuzaki y Hisashi, 2002). Más recientemente se encuentran nuevos trabajos, donde se extienden estas investigaciones aplicando técnicas de modelado estadístico que incorporan distintos indicadores de las discontinuidades acústicas (se incluyen también los sonidos sordos) (Syrdal y Conkie, 2004; Syrdal y Conkie, 2005) o aplicando técnicas no lineales para modelar las transiciones (Pantazis, Stylianou y Klabbers, 2005). En todos ellos se consideran distintos parámetros y técnicas, llegando a conclusiones heterogéneas, prevaleciendo el hecho de que no hay ninguna medida que mantenga una gran consistencia a lo largo de los estudios (Vepa, King y Taylor, 2002; Campillo, 2005). Aún así, en general, la distancia simétrica de Kullback-Leiber (SKL) —recientemente utilizada también para modelar el coste de unidad en (Zhao et al., 2006)—, junto con la distancia basada en MFCC, son de las que presentan mejores resultados en la literatura (Stylianou y Syrdal, 2001; Donovan, 2001; Tsuzaki y Hisashi, 2002). No obstante, en otro trabajo reciente, se demuestra como la aplicación de una técnica basada en el discriminante Lineal de Fisher mejora los resultados obtenidos utilizando SKL (Pantazis, Stylianou y Klabbers, 2005).

Finalmente, cabe añadir que estos subcostes pueden integrarse en la función de coste global siguiendo lo indicado en la ecuación (2.3), o aplicando algún mapeo más complejo, como los descritos en (Toda, 2003; Toda, Kawai y Tsuzaki, 2004), donde además de estudiar

el subcoste más adecuado para cada tipo de parámetro también se analiza la mejor manera, en términos de correlación subjetiva, de integrar estos subcostes. No obstante, este análisis queda fuera del alcance del presente trabajo de investigación que se centra en la definición clásica de la función de coste como el sumatorio de subcostes ponderados (ver ecuación (2.3)).

Subcostes utilizados

El trabajo de investigación que se describe en este documento se centra en el diseño y optimización de un método que permita el entrenamiento eficiente de los pesos de la función de selección. Sin embargo, no se debe olvidar que estos pesos ponderan unos subcostes que conforman dicha función de coste (ecuación (2.3)) y que también hay que diseñar adecuadamente —ver entre otros, (Febrer, 2001; Toda, 2003; Toda, Kawai y Tsuzaki, 2004; Campillo, 2005) para un estudio detallado de este tipo de subcostes. No obstante, dado que no son el objetivo de este trabajo de investigación, su diseño se ha simplificado por el momento. Concretamente, la función de coste de selección utilizada sólo toma en consideración tres parámetros prosódicos: la duración, la frecuencia fundamental y la energía, junto con los Mel Cepstrum (MFCC) (que modelan el espectro de la señal). A partir de estos parámetros se diseñan los subcostes utilizados en los experimentos desarrollados para el ajuste de pesos y que se pasan a describir a continuación.

Como se acaba de comentar, existen diversos métodos y aproximaciones, más o menos sofisticadas, para el diseño de los subcostes de selección. Durante este trabajo de investigación, se han probado distintas configuraciones para su diseño. Fundamentalmente, se ha escogido trabajar con medidas escalares y vectoriales (por las particularidades de los parámetros escogidos), siguiendo una filosofía similar a la presentada en (Febrer, 2001).

En una primera aproximación, se optó por trabajar con unas medidas sencillas, basadas en las diferencias entre los valores de los parámetros considerados (P_j en las siguientes ecuaciones). Concretamente, los subcostes de unidad (ecuación (2.1)) se miden a partir de las diferencias medias de frecuencia fundamental (*pitch*), energía y duración de las unidades comparadas. En cambio, los subcostes de concatenación (ecuación (2.2)) toman en consideración las diferencias locales (en el punto de concatenación, indicadas como L - *left* - y R - *right* - en las ecuaciones) del *pitch*, la energía y los coeficientes MFCC (cálculo vectorial). Los subcostes C_j^c se calcularán sobre la última trama de la unidad anterior (u_{i-1}) y la primera trama de la unidad actual (u_i). Estos datos se extraen de la información almacenada en el corpus de voz, mediante los ficheros correspondientes. Así pues, los subcostes considerados son:

- **PIT T**: subcoste de *pitch* de unidad (o *target*, T). Permite comparar la similitud de frecuencias fundamentales entre la unidad objetivo y la candidata (promediadas a lo largo de cada unidad)[4].

[4]Debido a que la unidad de trabajo es el difonema, debe considerarse la aparición de parejas de simifonemas sonoros y sordos. En este caso, sólo se considerará el valor de *pitch* del semifonema sonoro en el cálculo del coste. La frecuencia fundamental se obtiene a partir de las marcas de *pitch* del corpus —ver capítulo 4

- **ENE T**: subcoste de energía de unidad[5]. Codifica la similitud de energía media de la unidad candidata respecto a la objetivo.
- **DUR T**: subcoste de duración de unidad[6]. Determina la similitud entre las duraciones de la unidad objetivo y la candidata.
- **PIT C**: subcoste de *pitch* de concatenación. Analiza la similitud de las frecuencias fundamentales de las unidades en el punto de concatenación[7].
- **ENE C**: subcoste de energía de concatenación. Codifica la diferencia de nivel energético de las unidades a concatenar.
- **MFC C**: subcoste espectral de concatenación. Determina cómo es de buena la unión entre las unidades a nivel espectral. Su cálculo se basa en la estimación del espectro mediante su parametrización cepstral en la escala Mel (en inglés, *Mel Frequency Cepstral Coefficients*, o MFCC, de ahí su nombre). Se utilizan 24 coeficientes cepstrales más sus derivadas calculadas sobre una ventana de $20ms$ en el punto de concatenación (hacia atrás o hacia delante, según se trate de u_{i-1} o u_i, respectivamente).

Para acotar los subcostes dentro del mismo rango de valores y evitar un sesgo en el ajuste de los pesos (cada subcoste trabaja con unidades diferentes: Hz, milisegundos, etc.), es necesario introducir algún tipo de normalización de las medidas utilizadas. Esta normalización deberá considerar parámetros estadísticos que modelen la distribución de los subcostes dentro del corpus de voz, según el enfoque escogido para el entrenamiento de los pesos (a nivel de unidad, por grupos de unidades o para todo el corpus). Por ejemplo, si se ajustan los pesos para todo el corpus, se considerará la distribución estadística de los subcostes para todas las unidades grabadas. Se calcula la media (μ en las ecuaciones) del subcoste, la desviación estándar (*std* o σ en las ecuaciones), el valor máximo (*Max* en las ecuaciones) y el valor mínimo (*min* en las ecuaciones) según el *contexto de ajuste* de los pesos considerado (por unidad, por grupo de unidades o para todo el corpus). Entonces, según el criterio de normalización escogido, las ecuaciones para el cálculo de los subcostes pueden ser:

- **Normalización *std*:** consiste en calcular las diferencias entre los parámetros evaluados respecto a la desviación media estándar del subcoste para el contexto de ajuste.

$$C_j^t(t_i, u_i) = \frac{\left|\overline{P_j}(t_i) - \overline{P_j}(u_i)\right|}{\sigma_{P_j}} \tag{2.5}$$

$$C_j^c(u_{i-1}, u_i) = \frac{\left|P_j^R(u_{i-1}) - P_j^L(u_i)\right|}{\sigma_{P_j}} \tag{2.6}$$

para más detalles.

[5]Calculada como la raíz cuadrada del sumatorio de muestras al cuadrado normalizado por el numero de muestras.

[6]La duración de las unidades se determina a partir de las marcas de segmentación del corpus.

[7]Si el semifonema en el punto de concatenación es sordo, este coste no será considerado, es decir, será nulo.

donde $\overline{P_j}(\cdot)$ indica el valor medio del subcoste para la unidad analizada, mientras que σ_{P_j} hace referencia a la desviación de este subcoste para las unidades analizadas.

- **Normalización *mean-std*:** cálculo de las diferencias de los parámetros normalizadas respecto a la media y la desviación del subcoste calculado según el contexto de ajuste (esta normalización también se conoce como *z-score* (Toda, 2003)):

$$\begin{aligned} X^t &= \left|\overline{P_j}(t_i) - \overline{P_j}(u_i)\right|, \\ C_j^t(t_i, u_i) &= \frac{X^t - \mu_{X^t}}{\sigma_{X^t}} \end{aligned} \tag{2.7}$$

$$\begin{aligned} X^c &= \left|P_j^R(u_{i-1}) - P_j^L(u_i)\right|, \\ C_j^c(u_{i-1}, u_i) &= \frac{X^c - \mu_{X^c}}{\sigma_{X^c}} \end{aligned} \tag{2.8}$$

donde μ_{P_j} representa el valor medio del subcoste para el conjunto de unidades analizadas.

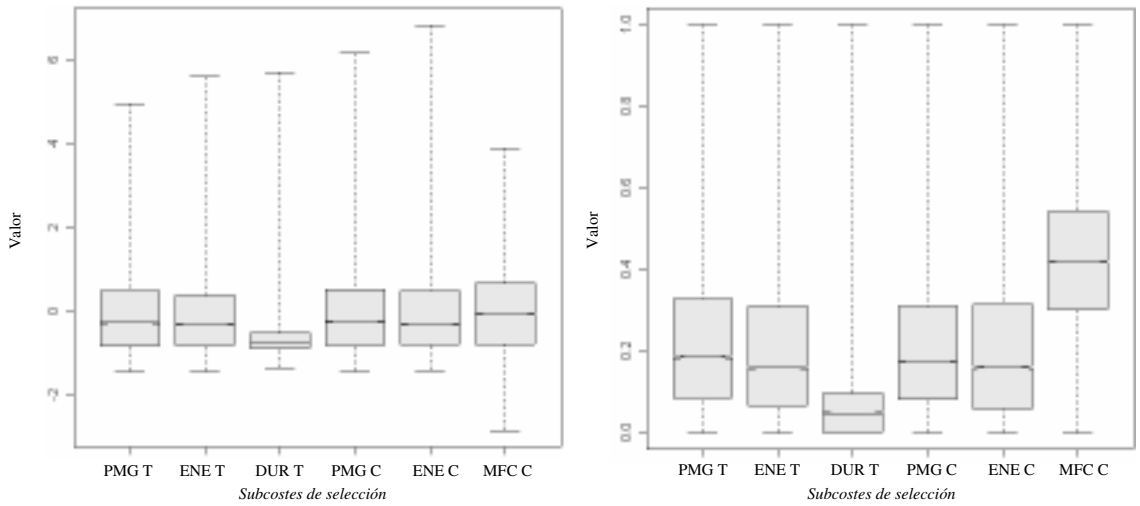
- **Normalización *Max-min*:** esta normalización permite trabajar con unas medidas de similitud (subcostes) acotados dentro del rango $[0, 1]$, evitando el sesgo que el valor de los costes puede causar para el entrenamiento de los pesos. Las ecuaciones que definen esta normalización son:

$$\begin{aligned} X^t &= \left|\overline{P_j}(t_i) - \overline{P_j}(u_i)\right|, \\ C_j^t(t_i, u_i) &= \frac{X^t - min(X^t)}{Max(X^t) - min(X^t)} \end{aligned} \tag{2.9}$$

$$\begin{aligned} X^c &= \sum_{1}^{N} \left|P_j^R(u_{i-1}) - P_j^L(u_i)\right|, \\ C_j^c(u_{i-1}, u_i) &= \frac{X^c - min(X^c)}{Max(X^c) - min(X^c)} \end{aligned} \tag{2.10}$$

donde $Max(\cdot)$ y $min(\cdot)$ representan el valor máximo y mínimo del subcoste, respectivamente, para el conjunto de unidades analizadas.

Por ejemplo, en la figura 2.5 se presenta el resultado del cálculo de los 6 subcostes considerados para todas las unidades del corpus a partir de las normalizaciones *Mean-std* y *Max-min*. De la observación de la distribución de los subcostes diseñados, se puede deducir que, aunque la mayor parte de los subcostes tienen un distribución de valores parecida, el subcoste



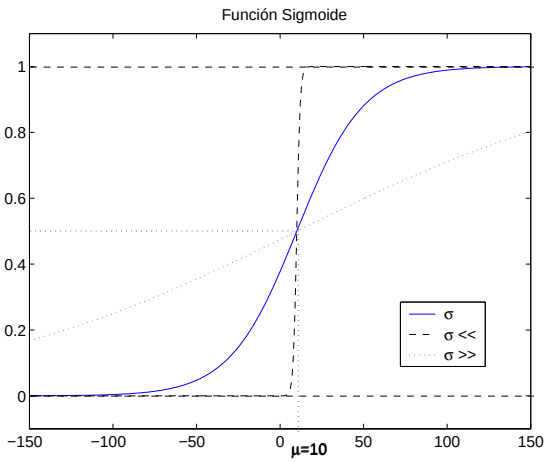
(a) Normalización *mean-std* (b) Normalización *Max-min*

Figura 2.5: Distribución de los seis subcostes diseñados a lo largo de todas las unidades del corpus normalizados según las estadísticas globales.

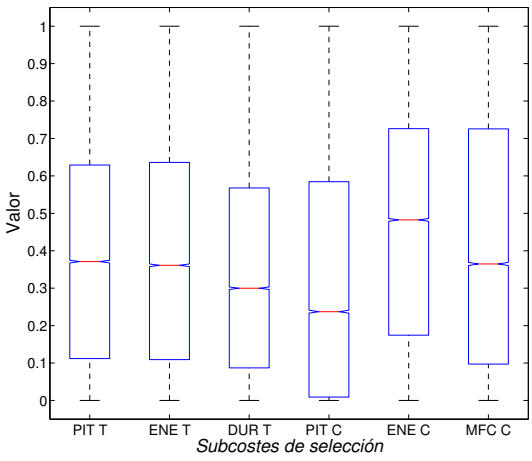
de duración de unidad (`DUR T`) presenta un comportamiento particular (mucho más concentrada en un conjunto de valores que el resto de subcostes), sea cuál sea la normalización escogida. Fundamentalmente esto se debe a haber normalizado con las estadísticas globales a todas las unidades del corpus y a los valores espurios debidos a errores de etiquetado —en este caso, existe una unidad que presenta dos realizaciones con duraciones muy distintas entre sí, provocando esta distribución del subcoste. Se trata, pues, de un parámetro muy sensible en el ajuste de pesos que, como se verá en el apartado 2.4.1, provocará el sesgo de los resultados en los experimentos de ajuste de pesos realizados inicialmente, obteniendo la mayoría de configuraciones de pesos un peso para `DUR T` más elevado que el resto (ver los histogramas de la figura 2.25). Por lo tanto, vuelve a demostrarse la necesidad de un buen diseño del corpus para disponer de un buen entrenamiento de los parámetros que intervienen en el algoritmo de selección de un CTH-SU.

A partir de las conclusiones obtenidas de estos experimentos iniciales, se optó por cambiar la forma de normalizar los subcostes de la función de coste. En este caso se utiliza:

- **Normalización *sigmoidea*:** cálculo de las diferencias de los parámetros normalizadas respecto a la media y la desviación del subcoste sobre una función sigmoidea (Febrer, 2001) (ver ejemplos en la figura 2.6(a)). En la figura 2.6(b) se presenta el resultado del cálculo de los subcostes para todas las unidades del corpus utilizando las siguientes ecuaciones:



(a) Ejemplos de función sigmoidea.



(b) Distribución de los seis subcostes considerados a lo largo de todas las unidades del corpus normalizados según una función sigmoidea.

Figura 2.6: Subcostes normalizados según una función sigmoidea.

$$X^t = \left|\overline{P_j}(t_i) - \overline{P_j}(u_i)\right|,$$
$$C_j^t(t_i, u_i) = 1 - e^{-\frac{(X^t - \mu(X^t))^2}{\sigma(X^t)^2}} \tag{2.11}$$

$$X^c = \sum_1^N \left|P_j^R(u_{i-1}) - P_j^L(u_i)\right|,$$
$$C_j^c(u_{i-1}, u_i) = 1 - e^{-\frac{(X^c - \mu(X^c))^2}{\sigma(X^c)^2}} \tag{2.12}$$

Asimismo, las estadísticas utilizadas (media y desviación) del proceso de normalización se obtienen para cada una de las unidades analizadas, evitando el problema que se acaba de comentar sobre el impacto de una normalización global de los subcostes.

2.1.5. Primeras técnicas para el ajuste de pesos

Como se puede observar de las ecuaciones (2.1) y (2.2), los subcostes que constituyen la función de coste están ponderados por unos pesos (w_j^t y w_j^c) que definen la importancia que debe tomar cada uno de los subcostes en la selección de la secuencia de unidades del corpus. De lo que se deduce que el entrenamiento eficiente de estos pesos es fundamental

para escoger las mejores unidades del corpus de voz, según los parámetros indicados por la secuencia objetivo (Black, 2002).

El entrenamiento de estos pesos es uno de los procesos más complejos en la fase de diseño de la función de coste de selección (Hunt y Black, 1996; Lee, Lopresti y Olive, 2001; Toda, Kawai y Tsuzaki, 2004; Campillo, 2005; Zhao et al., 2006). El objetivo de cualquier proceso de entrenamiento de estos pesos pasa por determinar el conjunto de valores $\mathcal{W} = (w_1^t, \ldots, w_p^t, w_1^c, \ldots, w_q^c)$ para obtener la máxima calidad de la señal sintética, según las restricciones indicadas en las ecuaciones (2.13) y (2.14) (Park, Kim y Kim, 2003), restringiendo así el rango de exploración. Este será el vector de pesos óptimo: $\mathcal{W}^*$. El concepto de optimización está fuertemente ligado al ámbito de la percepción humana, ya que el vector $\mathcal{W}^*$ será aquel que permita conseguir generar señales sintéticas con la máxima inteligibilidad y naturalidad posibles.

$$\sum_{i=1}^{p+q} w_i = 1, \tag{2.13}$$

$$w_i \geq 0 \tag{2.14}$$

Las primeras técnicas propuestas para el ajuste de pesos se basaron en utilizar una medida objetiva encargada de determinar la mejor selección de unidades en términos de la calidad subjetiva percibida por el usuario (Black y Campbell, 1995). En ese caso, la distancia objetiva utilizada para entrenar los pesos de la función de coste debería reflejar la similitud perceptual de las unidades acústicas comparadas (Hunt y Black, 1996). Para ello, típicamente, se utiliza la distancia Euclídea media entre vectores cepstrales (alineados temporalmente) (Black y Campbell, 1995; Hunt y Black, 1996; Meron y Hirose, 1999), aunque se observa posteriormente que este tipo de medida objetiva no permite evaluar todos los parámetros que caracterizan las unidades acústicas (Campillo, 2005).

Por otro lado, como ya se ha comentado, el entrenamiento de pesos se puede realizar a tres niveles: por unidad, por grupo de unidades o para todo el corpus (Hunt y Black, 1996; Black y Taylor, 1997a), utilizando como secuencia objetivo expresiones naturales extraídas del corpus de voz (Peng, Schuurmans y Wang, 2003). Por lo tanto, el ajuste de los pesos dependerá del corpus de voz utilizado, según su cobertura fonética y prosódica (Campillo, 2005). Asimismo, estos pesos pueden depender o no de la técnica de modificación prosódica de la señal (p.ej. TD-PSOLA (Moulines y Charpentier, 1990)). Existen trabajos que incorporan la modificación prosódica de la señal dentro del proceso de ajuste de los pesos, como por ejemplo (Meron y Hirose, 1999) (así como los métodos de ajuste manual), mientras que en otros, se trata de independizar los efectos de la modificación de la señal para obtener un entrenamiento independiente del método de modificación de la señal utilizado, por ejemplo (Hunt y Black, 1996; Toda et al., 2002).

A continuación, se presenta un resumen de las estrategias más habituales empleadas en el complicado proceso de ajuste de los pesos involucrados en la función de coste (ecuación (2.3)).

- **Ajuste manual:** como primera aproximación, los pesos pueden ajustarse de forma

manual, mediante un proceso supervisado perceptualmente (Coorman et al., 2000; Febrer, 2001; Chu et al., 2001; Blouin et al., 2002; Peng, Zhao y Chu, 2002; Meng et al., 2002; Toda, 2003; Campillo, 2005). El proceso parte, normalmente, de la elección previa de un conjunto de valores finito para los pesos de la función de coste (definida en la ecuación (2.3)). Entonces, se sintetizan un conjunto de frases o segmentos de voz con cada configuración de pesos y se presentan los resultados al evaluador (habitualmente, un experto en tecnologías del habla). Éste escoge, para cada prueba y para cada grupo de resultados, la realización sonora que subjetivamente más le guste (máxima naturalidad). Una vez finalizadas las pruebas, se obtiene la configuración manual de los pesos de la función de coste.

A modo de ejemplo, algunos de los problemas principales que se pueden presentar siguiendo esta aproximación son:

1. Es necesario seleccionar un grupo de pesos finito, cosa que implica la necesidad de discretización del espacio de búsqueda siguiendo un determinado criterio, sesgando los resultados, por lo que resulta interesante disponer de un *sistema automático* que vaya guiando el proceso de selección de los pesos considerados en lugar de escogerlos a priori.
2. Los evaluadores, aún siendo expertos, pueden seguir criterios distintos al evaluar las realizaciones sonoras, sobre todo cuando tienen que analizar muchas unidades o comparar realizaciones muy distintas entre sí (Black y Campbell, 1995). Esto puede provocar que los resultados puedan presentar ciertas inconsistencias, si no se controla la *consistencia* del criterio de los usuarios.
3. Los criterios que toma en consideración un experto en tecnologías del habla pueden ser distintos a los criterios que un usuario no acostumbrado a escuchar voz sintética pueda aplicar, por lo que resultaría interesante estudiar los resultados obtenidos por distintos *perfiles de usuario*.
4. El número de evaluaciones necesario para realizar un ajuste fiable desde un punto de vista estadístico provoca que la duración de las pruebas sea elevada, causando la *fatiga* del usuario y que los resultados sean poco fiables (Campillo, 2005).

Por estos motivos, se suele argumentar que la aplicación de un sistema automático para el entrenamiento de los pesos involucrados en la función de coste de selección suele aportar resultados más consistentes y robustos que dejarlo sólo en manos del usuario sin controlar el proceso (Hunt y Black, 1996; Yi y Glass, 2002). La dificultad radica en encontrar una aproximación que modele lo mejor posible el espacio de búsqueda (posibles valores de los pesos) y la respuesta perceptiva humana ante los resultados sonoros que éstos permiten obtener (Yi, 2003). La propuesta de un ajuste subjetivo de pesos eficiente presentada en este trabajo permite abordar todos estos problemas.

En el ámbito de las propuestas de ajuste automático de los pesos, cabe destacar las dos aproximaciones presentadas en el trabajo de (Hunt y Black, 1996): la búsqueda del espacio de pesos (*weight space search* o WSS), técnica introducida inicialmente en

(Black y Campbell, 1995), y el cálculo de los pesos mediante regresión multilínea (*multilinear regression* o MLR). A continuación se describen con detalle estos dos métodos así como las aportaciones que Meron y Hirose (1999) introdujeron posteriormente, ya que han sido durante años los métodos de referencia para el ajuste de los pesos de la función de coste de selección.

- **Búsqueda del espacio de pesos (WSS):** esta técnica se basa en la discretitzación del espacio de pesos $\mathcal{W}$ en un conjunto finito de valores posibles. Se trata de un espacio multidimensional de características, es decir, tiene tantas dimensiones como pesos (o subcostes) considerados. Por lo tanto, el número de posibles combinaciones a considerar puede llegar a ser muy elevado, dependiendo exponencialmente del número de pesos y el grado de discretización del espacio de búsqueda (Hunt y Black, 1996; Campillo y Rodríguez Banga, 2003; Peng, Schuurmans y Wang, 2003; Campillo, 2005). Esto provoca que este método sea computacionalmente muy costoso, y que se deba controlar el número de valores de los pesos (posibles soluciones) considerados en el barrido. Por lo tanto, se debe escoger un conjunto finito de vectores de pesos $\mathcal{W}$ (con p pesos de unidad w_j^t y q pesos de concatenación w_j^c en cada configuración del conjunto de análisis) sobre el que se determinará la configuración de pesos óptima (la figura 2.7 presenta un ejemplo de espacio 2D discretizado uniformemente).

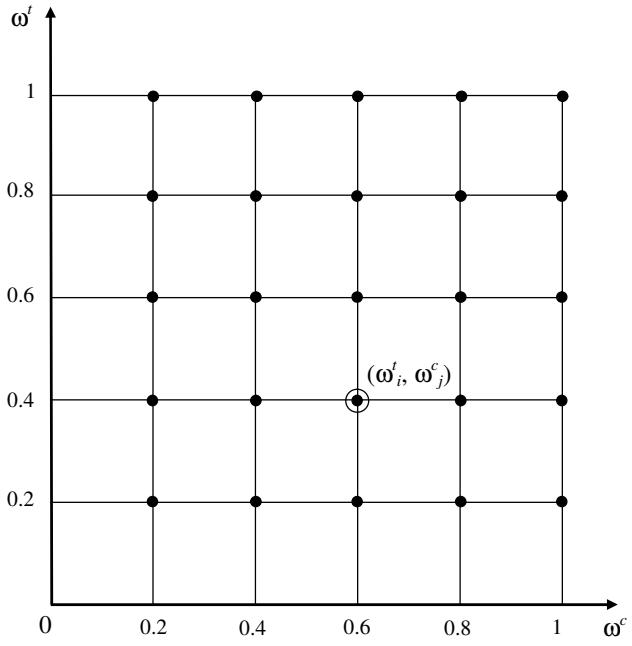


Figura 2.7: Ejemplo del espacio de búsqueda discretizado uniformemente (con incrementos de 0.2) para dos pesos: peso de unidad (w^t) y peso de concatenación (w^c).

El conjunto óptimo de pesos se obtiene mediante un proceso de análisis (selección de

unidades) y síntesis (generación de voz) que evalúa la bondad de los resultados, y por consiguiente de los pesos, de forma objetiva —sin la intervención humana. Concretamente, el proceso parte de la elección de una expresión (p.ej. una frase) presente en el corpus de voz, que es extraída como secuencia objetivo. De las configuraciones de pesos obtenidas para todo el espacio discretizado, se sintetiza la secuencia (una vez extraída esa expresión del corpus) que más se ajusta a la expresión objetivo considerada, según la función de coste (ecuación (2.3)) (Peng, Schuurmans y Wang, 2003). A continuación, el conjunto de secuencias de unidades candidatas —se obtiene una secuencia para cada posible vector de pesos considerado— se compara con la expresión original mediante una distancia objetiva, normalmente la distancia cepstral (Black y Campbell, 1995; Hunt y Black, 1996). Así, una vez alineados temporalmente los parámetros cepstrales que modelan las dos expresiones, se determina la mejor secuencia de unidades candidatas de forma objetiva. Este proceso se repite para diferentes expresiones extraídas sucesivamente del corpus, hasta escoger el mejor conjunto de pesos, $\mathcal{W}^* = (w_1^t, \ldots, w_p^t, w_1^c, \ldots, w_q^c)$, que será el que presenta un comportamiento más *consistente* a lo largo de las pruebas (Hunt y Black, 1996).

Este método se puede aplicar para el entrenamiento tanto de los pesos de unidad como de concatenación (Hunt y Black, 1996), aunque posteriormente sólo fue utilizado para los pesos de concatenación (Black y Taylor, 1997a). Este cambio de enfoque es debido a la introducción de una nueva estrategia para la organización de las unidades en el corpus de voz. En ese trabajo Black y Taylor introdujeron el concepto de agrupamiento (en inglés, *clustering*) de las unidades. La idea consiste en clasificar las unidades dentro del corpus de voz según diferentes criterios prosódicos, fonológicos, fonéticos, etc. Así, para cada tipo de unidad, se construye un árbol de regresión y clasificación (*Classification and Regression Tree* o CART) (Breiman et al., 1984) con el fin de organizar las unidades según la información que el módulo de PLN puede aportar durante el proceso de CTH. Según los autores, así se evita el problema de la estimación de los pesos de unidad (w_j^t), sustituyendo el cálculo del coste de unidad (C^t) durante el proceso de selección por la indexación de las unidades dentro del árbol. No obstante, la distancia acústica utilizada para generar los *clusters* también incluye un conjunto de pesos a optimizar (ver (Black y Taylor, 1997a)), por lo que el problema ha sido trasladado, no solucionado (Campillo, 2005). Por otro lado, el problema fundamental de esta aproximación es que separa en dos procesos independientes el entrenamiento o ajuste de los pesos de concatenación y de unidad, cuando ambos están interrelacionados e influyen decisivamente en la calidad de la señal sintética (Meron y Hirose, 1999).

Más adelante, este método fue estudiado por Meron y Hirose (1999). Éstos dividieron el proceso de entrenamiento en dos fases independientes: *análisis* + *síntesis* con el fin de acelerar su funcionamiento. A continuación se apuntan sus aportaciones fundamentales al proceso:

- *Análisis:* consiste en precalcular los costes para todas las posibles secuencias candidatas, sin ninguna ponderación, una vez definido el conjunto de expresiones

de test.

- *Síntesis:* para cada combinación de pesos estudiada se sintetiza la secuencia de unidades escogida por el proceso de búsqueda dinámico, según los costes de la fase de análisis. Sin embargo, este proceso se puede acelerar si el orden de las combinaciones de pesos estudiadas se distribuye de forma eficiente. Es decir, en cada iteración del proceso sólo se cambia uno de los pesos, aprovechando los cálculos del resto de valores del vector de ponderaciones analizado.

Por otra parte, Meron y Hirose (1999), introducen modificaciones importantes en la fase de comparación de las expresiones candidatas respecto a la expresión original. Concretamente, introdujeron la modificación prosódica de las expresiones candidatas mediante PSOLA (Moulines y Charpentier, 1990) antes de la alineación temporal de los parámetros cepstrales de las unidades a comparar mediante la distancia acústica. De esta manera, se toma en consideración el resultado del módulo de procesamiento digital de la señal (PDS) que suele acompañar a los sistemas de CTH —también para selección de unidades. Normalmente, resulta necesario introducir pequeños retoques en los puntos de unión de las unidades recuperadas del corpus con el fin de evitar sonidos artificiales en la señal sintética; ya que, aunque existen aproximaciones que no incluyen ningún postprocesamiento de la señal de voz (Black y Taylor, 1994) —concatenación directa—, la mayoría de sistemas de CTH-SU incorporan este módulo (p.ej. el sistema de AT&T (Beutnagel et al., 1999; Syrdal et al., 2000)). Pero, por otra parte, cabe destacar que, en este caso, el proceso de ajuste de pesos se hace totalmente dependiente del módulo de PDS considerado.

- **Regresión multilínea (MLR):** este método fue aplicado inicialmente sólo para el cálculo de los pesos de unidad (w_j^t) (Hunt y Black, 1996). En este caso, el conjunto de valores que definen las ponderaciones de unidad son aquéllos que mapean mejor, según una regresión multilínea, los costes de unidad (C^t) respecto a la distancia acústica que evalúa el grado de similitud entre las unidades candidatas y las deseadas (también extraídas del corpus, como en el caso de WSS). Este método pretende determinar los pesos w_j^t que permitan seleccionar las unidades candidatas más próximas acústicamente a las unidades objetivo, intentando escoger las mismas unidades que serían elegidas mediante la distancia objetiva si ésta pudiera integrarse directamente en el algoritmo de selección de unidades. En este propuesta incial de MLR no se incluyeron los pesos de concatenación, que fueron incorporados posteriormente en el trabajo presentado en (Meron y Hirose, 1999).

  En este caso, cada una de las unidades del corpus (fonemas, difonemas, etc., según sea la unidad mínima del corpus), es utilizada como unidad objetivo (t) para ser comparada con el resto de unidades del mismo tipo (u_i). Esta comparación se lleva a cabo tomando en consideración las siguientes similitudes:

  - *Similitud acústica:* se evalúa el grado de similitud entre t y cada una de las unidades candidatas u_i mediante una distancia objetivo (normalmente cepstral).

- *Similitud de unidad:* se evalúa el coste de unidad C^t mediante los subcostes correspondientes (ver ecuación (2.1)), considerando sólo las k unidades candidatas más próximas acústicamente —$k = 20$ en (Hunt y Black, 1996).

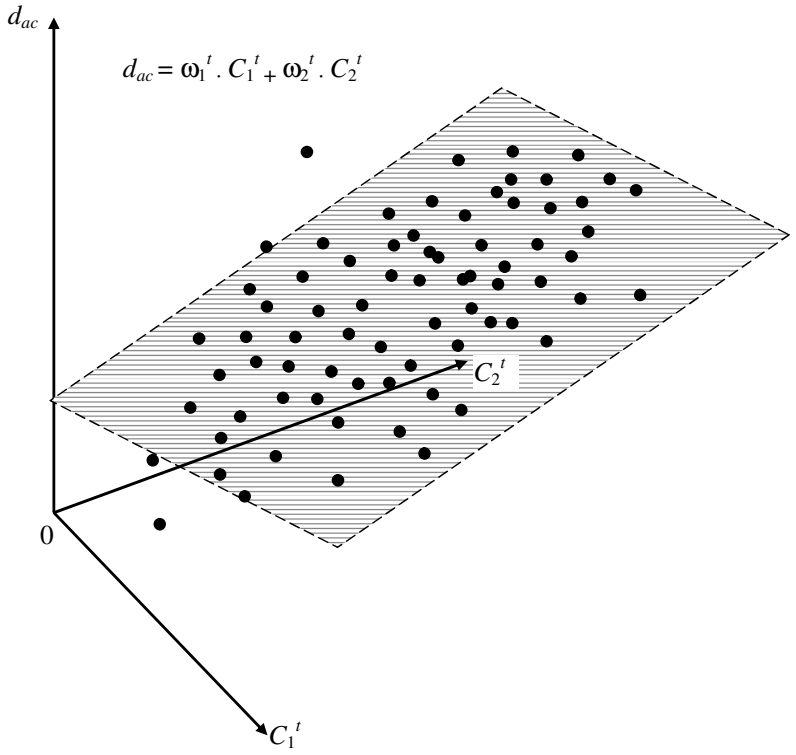


Figura 2.8: Ejemplo del resultado de una regresión lineal en tres dimensiones: distancia acústica (d_{ac}) y dos subcostes de unidad (C_1^t y C_2^t).

Este proceso se repite para *todas* las realizaciones de las que dispone la unidad estudiada dentro del corpus. Es decir, secuencialmente, cada una de las unidades candidatas es tratada como unidad objetivo y es comparada con el resto de unidades del grupo. Sobre esta nube de puntos p-dimensional, definida a partir de las parejas de valores *distancia acústica-subcoste de unidad* de las k unidades candidatas más parecidas a la unidad objetivo (ver ecuación (2.1)), se aplica una regresión multidimensional con el fin de mapear la relación existente entre los subcostes y la distancia acústica. En este caso, el mapeo es lineal y se realiza mediante un hiperplano (ecuación (2.15)). Es decir, se pretende obtener la mejor aproximación lineal de la distancia acústica a partir de los subcostes de unidad. La figura 2.8 presenta un ejemplo teórico del resultado de una regresión bilineal obtenida sobre el espacio formado por dos subcostes de unidad (C_1^t y C_2^t) y la distancia acústica entre unidades.

$$d_{ac} = \omega_1^t C_1^t + \ldots + \omega_p^t C_p^t \tag{2.15}$$

El proceso que se acaba de describir se debe repetir para todas las unidades del corpus, obteniendo todos los pesos necesarios para el proceso de selección de unidades (cada unidad tendrá asociado un vector de pesos específico). Como se comenta en (Hunt y Black, 1996), la modularidad de esta aproximación permite entrenar los pesos para todo el corpus de voz, para grupos de unidades (oclusivas sonoras, nasales, etc.) o para unidades independientes (p.ej. cada fonema o difonema del corpus). Por lo tanto, este método es una herramienta flexible, que permite al investigador o desarrollador de sistemas de CTH-SU escoger la configuración que crea más oportuna según las necesidades del sistema. Si se trabaja con un entrenamiento global para todo el corpus, los pesos obtenidos serán más genéricos (información más promediada) que si se ajustan para cada uno de los fonemas del corpus (permite ajustar los pesos de las unidades a las particularidades de cada unidad, siempre que ésta disponga de la suficiente variabilidad estadística).

Por otra parte, cabe destacar que este método teóricamente es más robusto que el WSS al utilizar en el proceso de selección y comparación (equivalente al de análisis y síntesis de WSS) *todas* las unidades del grupo de instancias estudiadas con el fin de ajustar los pesos w_j^t —análisis exhaustivo, mientras que el método WSS sólo escoge un conjunto discreto de valores del espacio de pesos, por lo que se obtiene un análisis menos exhaustivo (Meron y Hirose, 1999). Además, el coste computacional para el ajuste de los pesos se reduce considerablemente, ya que el tiempo de entrenamiento se incrementa linealmente con el número de los subcostes considerado ante el incremento exponencial que presenta el método WSS.

Posteriormente, Meron y Hirose (1999) mejoraron las prestaciones de este método mediante su aplicación al entrenamiento simultáneo de los pesos de unidad (w_j^t) y de concatenación (w_j^c). Esta variante supone el estudio de parejas de unidades (en su caso, fonemas) para el entrenamiento de los pesos. Por otra parte, como se ha descrito para WSS, el método de ajuste de pesos propuesto incluye la modificación prosódica de las unidades candidatas antes de ser concatenadas y comparadas acústicamente (d_{ac}) con las unidades objetivo. Por lo tanto, se introduce una variante del método MLR que permite el entrenamiento simultáneo de pesos (ver ecuación 2.16), con un coste computacional razonable.

$$d_{ac} = \omega_1^t C_1^t + \ldots + \omega_p^t C_p^t + \ldots + \omega_1^c C_1^c + \ldots + \omega_q^t C_q^t \tag{2.16}$$

2.2. Ajuste de pesos mediante algoritmos genéticos

Como se acaba de describir, WSS y MLR fueron los dos métodos inicialmente definidos para el ajuste automático de pesos. Sin embargo ninguno de ellos ha permitido obtener unos resultados definitivos en lo referente a la optimización del conjunto de pesos para selección de unidades. WSS, como se ha comentado, sufre dos problemas: el elevado coste computacional y la discretización del espacio de búsqueda. MLR, a pesar de presentar soluciones más robustas estadísticamente, fuerza una relación lineal entre las medidas de

los subcostes y la distancia acústica. Esta relación no tiene porqué ser cierta, ya que se puede tratar de una dependencia más compleja y que debe ser estudiada mediante la aplicación de algún método más genérico (no lineal).

A partir de esos trabajos aparecieron nuevas propuestas de métodos de entrenamiento, fundamentalmente, objetivos que buscaban mejorar las prestaciones de estas dos aproximaciones mediante la incorporación de técnicas muy diversas. Entre ellas cabe destacar, el trabajo de (Wu y Chen, 2001), que define tres categorías genéricas de pesos: w_l (*loose*), w_t (*tight*) y w_o (*overlapped*), según el tipo de unidades consideradas; el trabajo de (Park, Kim y Kim, 2003), que introduce un método no lineal basado en la técnica de entrenamiento discriminativo utilizada en problemas de clasificación, como por ejemplo el reconocimiento del habla. En este trabajo, los pesos se actualizan mediante la aproximación del gradiente descendiente tomando en consideración el error de clasificación como medida objetiva a optimizar —necesita de un ajuste heurístico de un conjunto de parámetros— en lugar de la típica distancia Euclídea cepstral utilizada por WSS o MLR. En sus experimentos Park, Kim y Kim (2003) sólo aplican esta técnica al entrenamiento de los pesos de unidad (w^t) de la función de coste, dejando el entrenamiento de los pesos de concatenación según lo indicado en (Hunt y Black, 1996).

Asimismo, a nivel nacional cabe destacar los trabajos de (Campillo y Rodríguez Banga, 2003; Campillo, 2005), donde se utiliza un método basado en MLR para ajustar los pesos de los costes debidos a parámetros fonéticos, utilizando como medida objetiva la distancia Euclídea cepstral (Mel-cepstrum), mientras que para los costes prosódicos definen unas funciones lineales limitadas por dos umbrales, que saturan el valor del coste (0 si el coste es menor que el umbral inferior y coste máximo si supera el umbral superior)— al estilo de (Coorman et al., 2000), o como simplificación de la función de coste sigmoidea descrita en (Febrer, 2001), por lo que estrictamente, no se realiza un entrenamiento automático de los pesos de los subcostes prosódicos, sino que éstos se controlan a partir de los umbrales definidos experimentalmente. No obstante, la función de coste de unidad incorpora un peso, ajustado subjetivamente, que balancea la importancia del conjunto de subcostes prosódicos respecto a los subcostes fonéticos. Posteriormente, en (Campillo, Alba y Rodríguez Banga, 2005) se extiende este enfoque sustituyendo el MLR por una Red Neuronal (perceptrón multicapa) que permite obtener mejores resultados —siempre trabajando sobre el coste de unidad. Las técnicas descritas en estos trabajos del grupo de la Universidad de Vigo, también son aplicables a grupos de unidades, es más, el agrupamiento de las unidades les permite obtener los mejores resultados (datos más homogéneos). No obstante, estos trabajos sólo se centran en el entrenamiento de los pesos de la función de coste de unidad, mientras que los pesos de concatenación son ajustados de forma *manual*.

Debido a la naturaleza del problema, éste puede ser planteado como un *problema de optimización*, en el que las variables que intervienen pertenecen al dominio real $\mathbb{R}$. En este marco y, buscando una aproximación más flexible, en este trabajo se evalúa la incorporación de los **algoritmos genéticos** al problema del entrenamiento y ajuste de los pesos de la función de coste de selección (w_j^t y w_j^c), como método de optimización de una función real de variables reales ($f : \mathbb{R}^n \rightarrow \mathbb{R}$) (Llorà, 1996). Algunos ejemplos de la aplicación de los algoritmos genéticos en el ámbito de las tecnologías del habla son: en (Boëffard

y Emerard, 1997) se utilizan en el diseño de corpus para obtener el modelo prosódico adaptado a una determinada aplicación del sistema de CTH, en (Lauri, Illina y Fohr, 2003) se aplican en la adaptación de los modelos acústicos a los cambios de locutor en el contexto del reconocimiento automático del habla, y finalmente, en (Kumar, 2004) se aplican al contexto de la CTH-SU, en este caso, definiendo un proceso iterativo para la selección de las unidades del corpus (se obtiene una solución después de unas 14 iteraciones del algoritmo) —utilizando pesos de la función de coste ajustados manualmente.

A continuación se presenta una breve introducción sobre los conceptos fundamentales en los que se basan este tipo de algoritmos de aprendizaje artificial —a partir del trabajo de (Llorà, 2001) y la bibliografía clásica que trata el tema (Goldberg, 1989; Michalewicz, 1992).

2.2.1. Los algoritmos genéticos

Los algoritmos genéticos (AG) fueron introducidos por John H. Holland a mediados de los años setenta en la Universidad de Michigan (Holland, 1975). Estos algoritmos se fundamentan en dos conceptos extraídos de la naturaleza: el principio de evolución natural de las especies de Darwin (1859) y las leyes de la herencia de Mendel (1965). Conceptualmente, los AG se basan en el hecho de tratar las posibles soluciones a un determinado problema como *individuos* de una *población* natural. Las características básicas de cada individuo están codificadas por sus cromosomas (parámetros), es decir, residen en su información genética. Estos individuos, como en la naturaleza, se cruzan entre sí según los principios básicos de la herencia genética. De este modo la población de la siguiente generación hereda las características esenciales de la población que la ha generado e intenta adaptarse mejor al *entorno* (problema). Por otra parte, como en todo proceso natural, se debe considerar la aparición de errores de cruce entre los individuos, denominados *mutaciones*. Así pues, la evolución puede incorporar pequeños cambios aleatorios en la copia de la información genética desde un individuo *padre* a un individuo *hijo* de la siguiente generación.

Gracias a los principios de cruce y mutación, la especie va adaptándose al entorno (va explorando posibles soluciones al problema), ya que éste es el que define qué individuos (soluciones) son los mejor adaptados (*selección*). Este principio de la naturaleza se puede codificar dentro de un algoritmo genético mediante la *función de evaluación* (o en inglés, *fitness function*), que modela el problema a resolver. Concretamente, esta función se encarga de determinar la bondad de los individuos (*evaluación*) para poder seleccionar aquel o aquellos individuos que permitan resolver satisfactoriamente el problema planteado. Por lo tanto, la función de evaluación guiará la evolución de la población, de manera que las mejores soluciones (producto de buenos cruces y buenas mutaciones) serán las que tendrán más posibilidades de sobrevivir en la siguiente generación. Una vez finalizado el proceso evolutivo, se obtienen unas soluciones que se consideran *bien* adaptadas al entorno. No obstante, puede ser que este conjunto de soluciones no contenga la *mejor* solución posible, aunque éstas, cuanto menos, deberían corresponder a un mínimo (o un máximo) de la función de evaluación o *fitness*.

```
t := 0
inicializar P(t)
evaluar P(t)
MIENTRAS (¬ condición) HACER
       seleccionar P(t+1) de P(t)
       recombinar P(t+1)
       evaluar P(t+1)
       t := t+1
FIN_MIENTRAS
```

Figura 2.9: Pseudocódigo de un algoritmo genético, siendo $P(t)$ la población de individuos en el instante t del ciclo evolutivo.

Cabe destacar que el proceso evolutivo de adaptación al medio de la población no es instantáneo ni propio de un único individuo, sino que precisa de un número importante de generaciones y de una población de individuos que se relacionen entre sí.

Una de las primeras y más directas aplicaciones de la teoría genética fue la *optimización* de funciones matemáticas. En este contexto, los individuos se convierten en los puntos del espacio de posibles soluciones, y el entorno está representado por la función a optimizar. No obstante la teoría genética ha sido aplicada a otros problemas dentro del mundo de la *búsqueda* iterativa de soluciones, como por ejemplo, problemas combinatorios (Michalewicz, 1992) —p.ej. el "problema del viajante de comercio" (*travelling salesman problem*)—, o la clasificación de datos (Llorà, 2001), entre otros.

Modelo de un algoritmo genético

Como se ha comentado anteriormente, un algoritmo genético trata de codificar la evolución natural de una población de individuos. Su objetivo es evolucionar un conjunto de individuos (soluciones posibles) con el fin de adaptarlos a un cierto entorno (problema a resolver). El modelo tradicional de algoritmo genético se puede resumir siguiendo el pseudocódigo de la figura 2.9 (Michalewicz, 1992).

Revisando el algoritmo propuesto, se puede apreciar que existen dos partes bien diferenciadas: la inicialización de la población de soluciones y el proceso iterativo de evolución y adaptación de estas soluciones al medio. En la inicialización, se define la población de partida (generación inicial). En esta etapa se asigna un valor inicial a cada uno de los genes que conforman los cromosomas de cada uno de los individuos de la población. Seguidamente se procede a la evaluación de los individuos que forman la población, determinando el grado de adaptación de las soluciones al problema a resolver. De esta manera se determina qué soluciones son más plausibles, por lo que deben ser tomadas en consideración para la siguiente generación (*selección*).

Posteriormente, si no se cumple la condición de finalización (indicada de forma heurística o según algún criterio de convergencia), se procede a entrar en el proceso evolutivo propiamente dicho. En cada iteración se realizan los pasos siguientes:

1. Se incrementa el contador de iteraciones.

2. Se seleccionan los individuos mejor adaptados al medio. Este proceso se consigue escogiendo los individuos de la iteración anterior que presentan una mejor función de evaluación (*fitness*). Estos individuos serán los que transmitirán su información genética a la siguiente iteración, conformando así la base de la nueva población.

3. Se recombinan, o modifican, los individuos de la población actual según los algoritmos que mapeen los procesos naturales de reproducción de las especies. Fundamentalmente, el *cruce* del material genético de los progenitores y la *mutación* del material genético de los descendientes.

4. Se evalúa el grado de adaptación al medio de la nueva población surgida de las etapas de selección y recombinación anteriores.

Convergencia de un algoritmo genético

El concepto de *convergencia* es uno de los elementos fundamentales del funcionamiento de los algoritmos genéticos. Resulta necesario demostrar que el proceso iterativo de selección y adaptación de los individuos es capaz de llegar a una solución estable. La convergencia de los AGs fue demostrada matemáticamente en (Holland, 1992) y (Goldberg, 1989; Michalewicz, 1992), inicialmente para soluciones binarias, y posteriormente extendida para problemas de variables reales (Goldberg, 2002). Los elementos fundamentales que permiten asegurar la convergencia de los AGs se basan en los siguientes conceptos:

- **Esquema:** hace referencia al hecho que las soluciones se pueden representar mediante un determinado patrón. Por ejemplo, si se trabaja con soluciones binarias representadas por tiras de 6 bits, un patrón o esquema (del inglés *scheme*) podría ser el siguiente (1∗∗001), donde '∗' indica que estos valores no influyen en el resultado de la función de evaluación. Es decir, todas las soluciones que tengan activo el primer y el último bit y desactivados el penúltimo y antepenúltimo bits presentarán un resultado equivalente para el problema planteado (idéntico o parecido, según el problema). Por lo tanto los esquemas definen los subconjuntos de soluciones del espacio de búsqueda que se comportan de forma equivalente.

- **Teorema fundamental:** de su enunciado se deduce que los patrones poco definidos (con elevado número de '∗') y compactos (tamaño reducido) que presenten una adaptación superior a la media, estarán mejor representados en la población de las iteraciones siguientes (Goldberg, 1989; Michalewicz, 1992), ayudando a la convergencia del proceso evolutivo.

- **Hipótesis de los bloques de construcción:** los esquemas cortos y compactos, con una razón de adaptación mayor que 1 , se los conoce como bloques de construcción (*building blocks* en inglés) y son los que permitirán determinar la solución del problema propuesto —serán la base de la solución final (ver (Goldberg, 1989; Michalewicz, 1992) para más detalles). Por lo tanto, los problemas en los que no se consiga obtener bloques de construcción serán difícilmente resolubles por parte de los algoritmos genéticos (p.ej. en problemas engañosos (Goldberg, 1989; Michalewicz, 1992)).

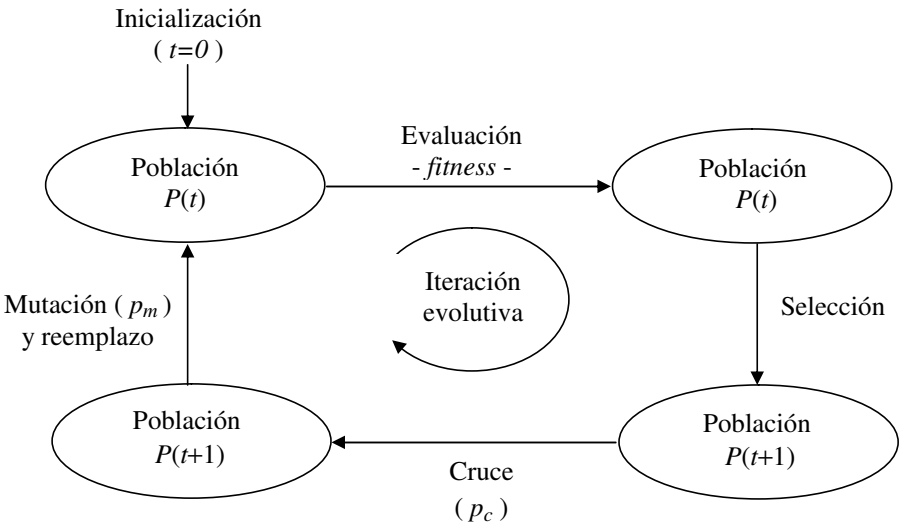


Figura 2.10: Esquema básico del funcionamiento de un algoritmo genético.

Fases y operadores de un algoritmo genético

En este apartado se describen las fases de las que consta un algoritmo genético a partir de la figura 2.10 que representa las fases de funcionamiento de un algoritmo genético (Davis, 1991; Holland, 1992; Michalewicz, 1992).

La etapa de **inicialización** es la primera fase de cualquier AG y tiene como objetivo definir la población sobre la que se llevará a cabo el proceso evolutivo. Existen dos perfiles bien diferenciados dentro del proceso de inicialización. Uno de ellos pretende diversificar al máximo las posibles soluciones y el otro pretende partir de la región o regiones de soluciones potencialmente bien adaptadas, conocidas de algún modo a priori. Por lo tanto, la primera opción busca conseguir que las soluciones que conformen la población inicial presenten la máxima cobertura del problema (un buen barrido del espacio a optimizar), al no disponer de ninguna estimación fiable del mismo, por lo que se maximiza la diversidad genética de la población. Bajo esta estrategia, se encuentran las técnicas de inicialización aleatoria de la población que se comentan a continuación. Por otro lado, la segunda estrategia de inicialización busca encontrar una población inicial que contenga soluciones que pertenezcan

a una región donde, intuitivamente, se prevé encontrar una de las soluciones al problema. Esta opción implica introducir conocimiento heurístico (experto) dentro del proceso de inicialización, con el objetivo de ayudar al algoritmo en su proceso de convergencia hacia una *buena* solución.

A continuación se describen algunas de las técnicas típicas para la inicialización de la población de soluciones (Holland, 1975; Goldberg, 1989):

- **Aleatoria:** es una de las técnicas más utilizadas para inicializar los algoritmos genéticos, por su simplicidad y por permitir un barrido amplio del espacio a optimizar. Se fundamenta en la asignación de valores aleatorios a las variables que conforman los individuos de la población siguiendo una distribución de probabilidad sobre el rango de valores sobre el que se definen las variables o genes del individuo (los pesos de la función de coste, en el presente trabajo).
- **Aleatoria repetitiva:** en este caso, la inicialización aleatoria es un paso previo a la determinación de la población inicial. Para ello, se genera una *macro*-población aleatoria (muchos más valores de los necesarios) de la que se escoge el subconjunto que esté mejor adaptado de entrada al medio.
- **Manual:** en este caso, el encargado de definir la población inicial es un experto. Esta estrategia es compatible con cualquiera de las dos estrategias de inicialización indicadas con anterioridad. Así, la población inicial a optimizar puede ser el resultado de un proceso previo, o bien, puede incorporar cierto conocimiento heurístico con el objetivo de sesgar la búsqueda hacia las regiones que, potencialmente, puedan contener buenas soluciones.

Durante la etapa de **evaluación** se asigna a cada individuo de la población una medida que indica la bondad de esa solución respecto al problema. Por ejemplo, en un problema de optimización de una función matemática, esta fase consistiría en asignar a cada individuo el resultado de calcular el valor de la función para al punto evaluado. La función de evaluación es uno de los elementos más críticos del algoritmo, ya que es la que aporta información del grado de adaptación de los individuos al medio. Por este motivo, es imprescindible un diseño cuidadoso del módulo de evaluación con el fin de obtener una estimación realista para el proceso de optimización.

Una vez evaluados los individuos que conforman la población, se procede a su **selección**. Para ello, se aplica una política de supervivencia de los individuos mejor adaptados al entorno, es decir, se escogen los individuos que permanecerán en la población en función de su grado de adaptación (evaluación) al entorno. De este modo, los individuos más *fuertes* tendrán más probabilidades de reproducirse, aumentando las posibilidades de que sus descendientes posean las mejores cualidades para adaptarse al medio. Existe un amplio abanico de posibilidades a la hora de implementar la etapa de selección de un AG. A modo de resumen, a continuación se presentan algunas de ellas:

- **Proporcional:** o también conocida como *roulette wheel selection* (RWS). Se trata

de una de las políticas de selección más utilizadas, junto con sus múltiples variantes (ver (Goldberg, 1989) para más información). En este caso, la probabilidad de que un individuo sea seleccionado es proporcional al valor de su evaluación. Es decir, cuanto mayor sea el grado de adaptación del individuo al entorno (*fitness* mayor o menor, según el problema) más probabilidad tendrá de ser escogido. Simbólicamente, esta política de selección se puede modelar mediante el uso de una ruleta donde cada individuo tiene una región asignada de área proporcional a su valor de evaluación. Seguidamente, se hace 'girar la ruleta' (mediante un proceso aleatorio) tantas veces como individuos se quieran obtener, escogiendo el individuo de la región seleccionada en cada caso (según las probabilidades acumuladas) como individuo base de la siguiente generación.

- **Rango lineal:** esta técnica también se basa en escoger los individuos de la población según su valor de evaluación mediante la representación de los individuos en una jerarquía que determina su grado de adaptación al problema (el individuo mejor adaptado se sitúa en la posición superior de la jerarquía, mientras que el que está peor adaptado ocupa la posición inferior). A continuación, a cada posición de la jerarquía se le asigna una probabilidad de selección linealmente decreciente, donde el mejor individuo es el que tiene más posibilidades de reproducirse. A partir de esta información, y también de forma aleatoria, se escogen los individuos que formarán la siguiente población siguiendo un proceso idéntico al explicado para RWS, pero, en este caso, sobre un espacio probabilístico distinto.

- **Rango uniforme:** según esta técnica, se selecciona un conjunto de K individuos (el número de individuos seleccionado se fija según las el problema y el tipo de algoritmo utilizado) que tienen la misma probabilidad (distribuida uniformemente, $\frac{1}{K}$) de reproducirse. El proceso de elección es muy similar al utilizado por las técnicas anteriores. La diferencia fundamental de esta técnica radica en el hecho que algunos individuos son descartados para la siguiente generación (técnica *extintiva*), a diferencia de las dos anteriores, donde todos los individuos podían tener descendientes (técnica *preservativa*). Por otra parte, todo método preservativo puede ser definido como extintivo, fijando un umbral de adaptación a fin de seleccionar los individuos que puedan ser progenitores.

- **Torneo:** la selección de los individuos que pasan a la siguiente generación se realiza de forma comparativa a diferencia de las técnicas de selección global anteriores (Goldberg, 1989). Por ejemplo, el torneo binario consiste en comparar parejas de individuos escogidos aleatoriamente, seleccionando aquel individuo que presente una mejor adaptación al problema. Esta será la técnica utilizada en el presente trabajo de investigación, tanto para el ajuste objetivo de pesos como para el subjetivo[8].

Una vez seleccionado el conjunto de individuos de la población, se pasa a **recombinar** su material genético para dar lugar a la nueva población mediante dos operadores: cruce y

[8] Al usuario le resulta más sencillo comparar soluciones sintéticas dos a dos que escucharlas todas y tener que ordenar su calidad.

mutación. Una vez finalizada la recombinación de los individuos, la población dispondrá de unos nuevos individuos que han heredado el material genético de sus progenitores. Se trata de un proceso probabilístico regido por la probabilidad de cruce (p_c) y la probabilidad de mutación (p_m) —de valor inferior a p_c, habitualmente— del material genético (ver figura 2.10).

Como ya se ha comentado anteriormente, el cruce de los materiales genéticos consiste en la mezcla de las soluciones preexistentes para crear nuevos individuos. Dentro del proceso de cruce se pueden distinguir dos elementos claramente diferenciados: el valor de la p_c, que deberá ser ajustado según el problema, y el operador de cruce encargado de recombinar el material genético. Existen distintos operadores de cruce, de entre los que destacan:

- **Puntual:** del inglés *1-point crossover*. Se fundamenta en un método de corte e intercambio de información genética entre dos individuos (o cromosomas). Aleatoriamente se escoge *una* posición de corte de los cromosomas y se intercambian su contenido para crear los descendientes de los dos progenitores. Por ejemplo, dados los individuos $i_1 = (0.1, 0.2, 0.3)$ e $i_2 = (0.5, 0.6, 0.7)$, y escogida la segunda posición como punto de cruce, se obtendrían los nuevos individuos $i'_1 = (0.1, \mathbf{0.6}, \mathbf{0.7})$ e $i'_2 = (0.5, \mathbf{0.2}, \mathbf{0.3})$.
- **Múltiple:** el funcionamiento es idéntico al del cruce puntual, pero en este caso se consideran n puntos de corte, escogidos aleatoriamente para el intercambio de material genético por tramos.
- **Aritmético:** mediante este operador, en lugar de simplemente intercambiar el material genético entre los cromosomas, se aplica una combinación lineal entre el valor de los mismos antes del cruce. Puede ser aplicado de forma puntual o múltiple.

En cuanto a la *mutación* del material genético de los individuos, también se determina de forma probabilística (mediante p_m) qué individuos sufrirán el *error* genético. Se trata de un operador secundario en el proceso evolutivo, encargado de mantener la diversidad genética dentro de la población. Fundamentalmente, existen dos estrategias distintas para aplicar el operador de mutación sobre la población de individuos. La primera consiste en analizar toda la población y, individuo a individuo, determinar aleatoriamente si éste debe mutar o no. La segunda se basa en determinar estadísticamente (de forma global) qué genes deben mutar, considerando la dimensión de la población y el valor de la probabilidad de mutación.

Una vez escogidos los genes a mutar, existen distintas posibilidades para definir la *acción* del operador de mutación:

- **Uniforme:** se sustituye la variable seleccionada por el nuevo valor, dentro del margen de valores al que pertenece la variable. Por ejemplo, si se trabaja con cadenas binarias, esta técnica provocaría la negación del bit de la posición escogida ('0'↔'1').
- **No uniforme:** consiste en trabajar con una elevada mutación en el inicio del ciclo evolutivo (mucha diversidad en el espacio de soluciones), reduciendo el valor de p_m

a medida que la población se acerca a la solución óptima (Michalewicz, 1992). Esta técnica surge de los problemas de optimización y permite, al mismo tiempo, explorar un número elevado de soluciones potenciales en el inicio de la búsqueda, acotando gradualmente la búsqueda de la solución a medida que el algoritmo va convergiendo.

- **Media aritmética:** consiste en sustituir el valor de la variable a mutar por la media aritmética del conjunto de variables que forman la población de soluciones. Se deberán acotar los resultados obtenidos para evitar que la mutación quede fuera del intervalo en el que se describe la variable.

Existen otros operadores más o menos complejos que se pueden adaptar a las necesidades puntuales de cada problema, pero que quedan fuera del alcance del presente trabajo de investigación (ver (Michalewicz, 1992) para más detalles).

Finalmente, dentro del esquema de algoritmo genético presentado en la figura 2.9 se incluye la fase de **reemplazo** de la población, es decir, inicialmente se trata de un modelo de AG *generacional*. En este caso, la población de descendientes reemplaza completamente a la población de progenitores que los ha generado. Aunque existen otras alternativas, como por ejemplo, los modelos no generacionales en los que conviven progenitores y descendentes (Llorà, 2001), el modelo generacional es el utilizado, por el momento, para solucionar el problema de optimización planteado en el marco de este trabajo de investigación. Los modelos más conocidos para el diseño de ésta fase del algoritmo genético son (Goldberg, 1989; Michalewicz, 1992):

- **Elitismo:** intenta mejorar la eficiencia del método generacional procurando perpetuar a los individuos mejor adaptados al medio. En el método generacional puede suceder que la mejor solución no sea seleccionada por la recombinación (sujeto a p_c), o que su descendencia presente una peor adaptación (menor *fitness*). Debido a que si pensamos en procesos biológicos, ésta situación es poco probable, el elitismo consiste en no perder la mejor solución encontrada hasta el momento, forzando que en el reemplazo ésta se mantenga siempre.

- ***Steady-state:*** este método permite la coexistencia en una misma población de los progenitores y de sus descendientes. De la misma manera que el elitismo procura no perder las mejores soluciones de una población, esta técnica consiste en sustituir las peores soluciones de la población $P(t)$ en la siguiente generación $P(t+1)$ (ver figura 2.10) por sus mejores descendientes, pero manteniendo los progenitores mejor adaptados. Se deberá controlar que la población no se sature hacia una solución subóptima, por ejemplo, aumentando la probabilidad de mutación.

Finalmente, si se quiere profundizar más en el campo de los algoritmos genéticos se recomienda consultar (Goldberg, 2002) y (Larrañaga y Lozano, 2002).

2.2.2. Adaptación de los algoritmos genéticos al ajuste objetivo de pesos

Los algoritmos genéticos proponen una metodología de trabajo genérica para la solución de problemas de optimización, una vez definida la representación de los individuos, las estrategias de cruce y mutación escogidas, el método de selección, etc. Asimismo, para cada problema concreto, además de considerar su función de evaluación específica, se deberá buscar la mejor representación y los mejores operadores posibles (Davis, 1991; Michalewicz, 1992). Los algoritmos evolutivos (AE), a menudo, se aplican sobre problemas de carácter ruidoso donde es necesario llevar a cabo una determinada optimización. La idea consiste en intentar reducir el ruido trabajando con una población de soluciones candidatas importante, tratando de cubrir el espacio de búsqueda lo mejor posible. De este modo, los AE se convierten en robustos ante los efectos del ruido, al compensar el carácter ruidoso del problema con un mayor número de individuos de lo habitual. Concretamente, en el ámbito de los algoritmos genéticos (que es un tipo de AE) existen diversos estudios que demuestran la robustez de los AG en contextos ruidosos (Miller y Goldberg, 1997). Basándose en la hipótesis de los bloques de construcción (o *building blocks*), Miller y Goldberg analizaron los efectos del ruido para diversos mecanismos de selección en el contexto de los AG. Mediante este estudio los autores llegaron a la conclusión de que los efectos del ruido pueden ser reducidos de forma considerable si se trabaja con una población suficientemente grande, donde el tamaño de la población se debe incrementar exponencialmente con la potencia del ruido (Goldberg, 2002). Estos primeros estudios aplicados inicialmente a problemas discretos, fueron contrastados posteriormente sobre espacios de búsqueda continuos y multidimensionales por Arnold y Beyer (2003), situación sobre la que se aplican los algoritmos genéticos en el presente trabajo de investigación, ya que resulta inabordable un barrido exhaustivo de todos las posibles configuraciones de pesos —el muestreo aleatorio de valores de los pesos más la capacidad de optimización del AG permitirán obtener la solución al problema.

Una vez presentados los fundamentos teóricos y las variantes fundamentales de las que disponen los algoritmos genéticos, a continuación se pasa a describir cómo se incorpora esta estrategia para el entrenamiento de los pesos de unidad y concatenación (w_j^t y w_j^c) que forman parte de la función de coste de selección (ver sección 2.1.2 y ecuación (2.3)). El esquema del funcionamiento del AG propuesto (figura 2.11) parte de una población de individuos de tamaño constante, inicializada aleatoriamente. Cada individuo es un vector $\mathcal{W}$ (configuración de pesos) que hay que ajustar. El objetivo del algoritmo es encontrar el individuo *mejor* adaptado al problema $\mathcal{W}^* = (w_1^{t*}, \ldots, w_p^{t*}, w_1^{c*}, \ldots, w_q^{c*})$. Durante la fase de evaluación se calcula el *fitness* de cada configuración de pesos (individuo) a partir de la ecuación (2.17) que promedia la función de coste para los k individuos más cercanos acústicamente al objetivo. A continuación el AG escoge los mejores individuos para actuar como progenitores de la siguiente generación. Este proceso, conocido como *selección*, se encarga de construir la nueva población a partir del muestreo de la generación anterior, considerando el grado de adaptación al problema de los individuos (estimada de forma implícita a partir de la distancia objetiva utilizada). Como ya se ha descrito en la introducción teórica sobre AG, existen diferentes opciones para implementar el operador de selección. De éstas se ha escogido trabajar con la selección por *torneo binario*, debido a su eficiencia ante la

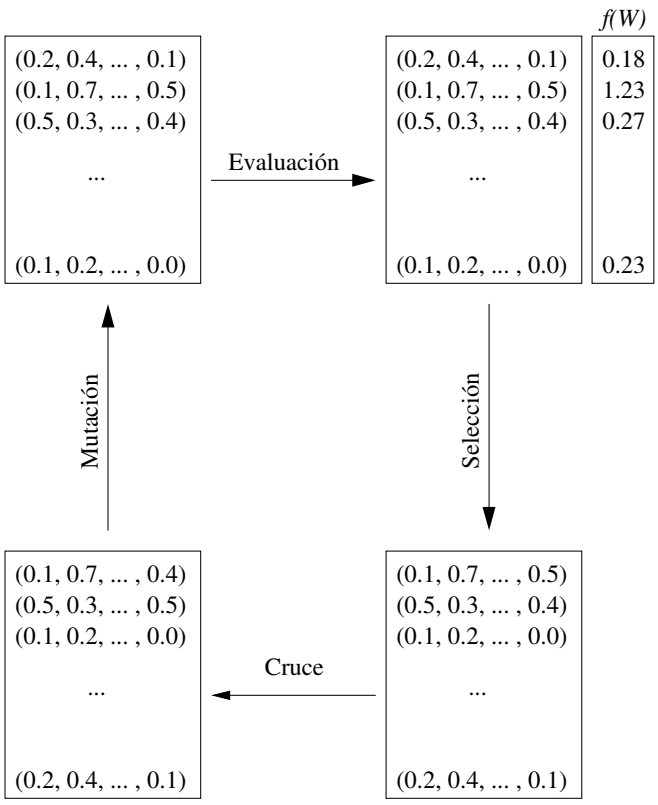


Figura 2.11: Diagrama del ciclo evolutivo para el ajuste objetivo de los pesos de la función de coste basado en un algoritmo genético clásico.

inestabilidad de las evaluaciones que se derivan de un problema ruidoso (Goldberg, 2002). Además, el muestreo aleatorio de los individuos permite acelerar el proceso de análisis del medio, tal y como se describirá más adelante en este mismo apartado.

Una vez se han escogido los progenitores de la nueva población, estos individuos se recombinan en dos fases. Primero se aplica el operador de cruce. Dados dos individuos escogidos aleatoriamente mediante la probabilidad de cruce (p_c), se combina el valor de sus pesos generando dos nuevos descendientes, reemplazando, en este caso, a sus progenitores (aproximación generacional). Este proceso se realiza a partir del operador de cruce puntual (Goldberg, 1989), descrito anteriormente. La segunda fase del proceso de recombinación es la mutación de algunos de los individuos generados. Se introduce una perturbación aleatoria sobre los valores de los individuos según una determinada probabilidad de mutación (p_m), siguiendo una mutación *uniforme* (sustitución del valor del gen actual por el nuevo valor aleatorio). Es decir, de forma probabilística se escoge de la población qué individuos sufrirán errores en su material genético (en este caso, variaciones en los valores de los pesos), para que la siguiente generación pueda analizar otras regiones del medio o problema evaluado. Una vez aplicados los operadores de cruce y mutación, se dispone de la nueva generación de

configuraciones de pesos que reemplaza a la anterior, iniciándose de nuevo el ciclo evolutivo.

Este proceso iterativo terminará cuando se alcance una determinada condición de finalización (ver figura 2.9), en este caso un determinado número de iteraciones escogido empíricamente (en el apartado donde se describen los experimentos desarrollados con el AG se presentan los valores escogidos para los parámetros de configuración utilizados).

Siguiendo el trabajo de Meron y Hirose (1999), en esta investigación también se ha escogido trabajar con parejas de unidades como elementos de entrenamiento de los pesos. De esta manera, el ajuste de los pesos de unidad (w^t) y de concatenación (w^c) se realiza de forma simultánea, evitando realizar procesos de ajuste independientes —no se consideran posibles interrelaciones—, ya que el AG es capaz de entrenarlos conjuntamente. Sin embargo, a diferencia del trabajo de Meron y Hirose (1999), donde se utilizaban fonemas, en este caso se trabaja con parejas de difonemas y trifonemas para el entrenamiento de los pesos de la función de coste. Esto es debido a que el CTH-SU utilizado se diseñó a partir de estas unidades como elementos mínimos del proceso de selección (Guaus y Iriondo, 2000a; Guaus y Iriondo, 2000b). Comentar que trabajar con difonemas y trifonemas en lugar de fonemas implica que el espacio de búsqueda se vea aumentado de forma considerable. Por lo tanto, el ajuste de los pesos de la función de coste basado en pares de difonemas y trifonemas aumentará en coste computacional respecto a los pares de fonemas; cuestión que, aunque importante, no es crítica debido a que este proceso se realiza fuera del proceso de la conversión de texto en habla a tiempo real.

Como primer paso, el cálculo de la medida de bondad de adaptación de las configuraciones de pesos —*fitness*— se llevó a cabo sobre un corpus de voz (ver apartado 2.4.1) organizado en unidades básicas. Es decir, todas las unidades, en este caso difonemas y trifonemas, han sido agrupadas por tipo de unidad (p.ej. todas las realizaciones de la unidad /b@/[9]). Por este motivo, lo que conseguirá el algoritmo genético será encontrar un vector de pesos particular para cada una de las unidades del corpus. En este punto cabe destacar que el AG diseñado también permite definir vectores de pesos para grupos de unidades (p.ej. oclusivas sonoras) o para todo el corpus en general, del mismo modo que lo permite el método de MLR (Hunt y Black, 1996; Meron y Hirose, 1999). Sólo será necesario cambiar el grupo de unidades sobre el que se quiera llevar a cabo el proceso de entrenamiento.

El proceso para evaluar a los individuos (configuraciones de pesos) en cada iteración del AG, entrenados a nivel de unidad (una configuración de pesos por unidad básica del corpus), se divide en los siguientes pasos (se trata de un similar al descrito en (Black y Campbell, 1995) y (Hunt y Black, 1996) para MLR, pero adaptado en este trabajo al uso de los AG):

- Se escoge aleatoriamente una de las realizaciones de la unidad, que realiza el papel de unidad objetivo (*target*, en inglés).
- Se calcula la distancia cepstral (evaluación objetiva de la similitud acústica) (Hunt y Black, 1996; Meron y Hirose, 1999) entre el resto de realizaciones de esa unidad (unidades candidatas) y la unidad objetivo, una vez parametrizadas y alineadas tempo-

[9]En este trabajo se utiliza la notación fonética SAMPA (Wells et al., 1992)

ralmente mediante un proceso de *Dynamic Time Warping* (DTW) (Rabiner y Juang, 1993).

- Se seleccionan las k unidades mejores, acústicamente hablando, para el cálculo de la función de coste ponderada a partir de la configuración de pesos $\mathcal{W}$ evaluada (unidades potencialmente seleccionables en tiempo de síntesis). En este trabajo, se ha escogido un valor de $k = 10$, ya que como se explicará más adelante, el número mínimo de realizaciones de las unidades estudiadas es de 25, con el objetivo de disponer de suficiente información estadística para el entrenamiento de los pesos (ver apartado 2.4.1).

Entonces, la función de *fitness* $f(\mathcal{W})$, que se utiliza para evaluar el grado de adaptación de los individuos de la población al problema planteado, se calcula como el promedio del coste de las k unidades candidatas mejores (más cercanas acústicamente a la unidad objetivo), dado el vector $\mathcal{W}$ de pesos a evaluar para una determinada iteración del proceso evolutivo la ecuación (2.17). En este caso, se opta por no incluir explícitamente la distancia acústica en el criterio de evaluación, sino que esta se utiliza de forma implícita, ya que la bondad de la configuración de pesos se evalúa a partir de los k-mejores individuos candidatos del conjunto de unidades analizado —en un futuro, se pretende evaluar funciones de *fitness* algo más complejas (p.ej. incorporando parámetros de estimación subjetiva, como PESQ (*Perceptual Evaluation of Speech Quality*) (Cernak y Rusko, 2005)) con el objetivo de mejorar los resultados obtenidos hasta el momento.

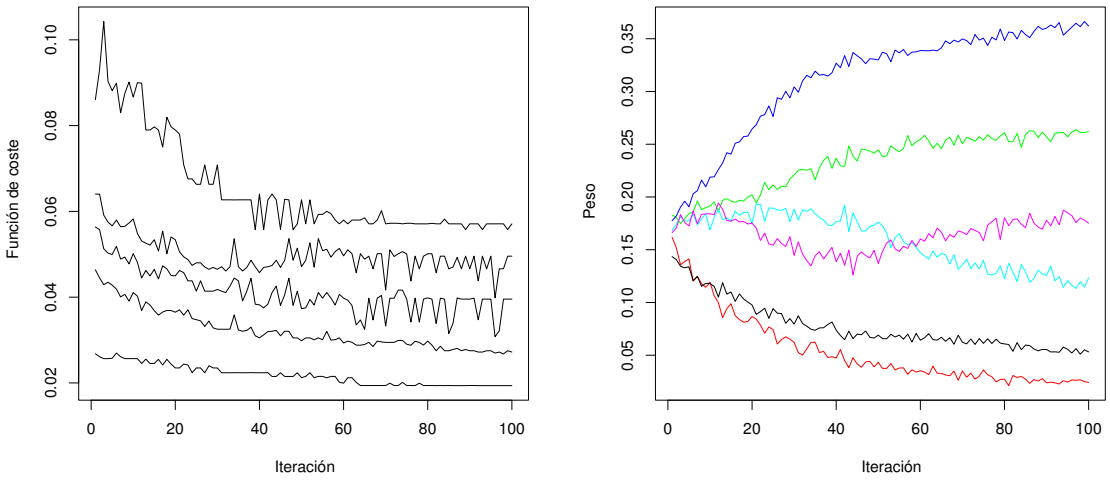
$$f(\mathcal{W}) = \frac{1}{k} \sum_{i \in k\text{-mejores}} C(t, u_i) \tag{2.17}$$

donde u_i representa la unidad i-ésima de entre las k candidatas y $C(t, u_i)$ la función de coste de la ecuación (2.3), particularizada para una unidad, en este caso —se considera el subcoste de unidad y de concatenación entre el difonema o trifonema objetivo y el candidato, según lo descrito anteriormente.

De los ajustes realizados para adaptar el AG al entrenamiento de pesos, cabe destacar la reducción del coste computacional conseguida gracias a trabajar con una estrategia de muestreo (selección aleatoria de la unidad objetivo) de los individuos analizados, a diferencia del método MLR que realiza un análisis exhaustivo (contempla todas las realizaciones de la unidad analizada). Sin embargo, el muestreo de individuos introduce ruido en las evaluaciones, cuestión que puede conducir a resultados erróneos. Este problema se resuelve gracias al buen funcionamiento de los algoritmos genéticos ante este tipo de problemas de optimización ruidosos (Goldberg, 2002; Arnold y Beyer, 2003).

A modo de ejemplo del funcionamiento del AG para el ajuste de pesos, se presenta la figura 2.12 dividida en dos subfiguras. En la subfigura 2.12(a) se muestra el proceso de adaptación de algunos individuos al problema a lo largo de las iteraciones, según su *fitness*. En concreto se puede observar el mejor (menor coste) y el peor individuo (mayor coste), junto a otros individuos intermedios escogidos al azar. En cambio, la subfigura 2.12(b)

muestra la evolución de la media de los pesos de la población a lo largo del proceso evolutivo hasta convergir en un determinado instante de tiempo (iteración).



(a) Ejemplo de la evolución del *fitness* para distintos individuos (muestreo) de la población (se incluye el mejor y el peor individuo).

(b) Ejemplo de la evolución media de los seis pesos considerados a lo largo del proceso iterativo.

Figura 2.12: Ejemplo del funcionamiento del AG sobre el conjunto de seis pesos (w_j^t y w_j^c) analizado para una unidad determinada.

2.2.3. Ajuste subjetivo de pesos mediante algoritmos genéticos interactivos

En los apartados 2.1.5 y 2.2.2 se han presentado diversas aproximaciones para el entrenamiento automático de los pesos de la función de coste a partir de una función *objetiva*. Estos métodos pretenden estimar las sensaciones que el usuario percibe subjetivamente e intentan mapearlas objetivamente. En la gran mayoría de las aproximaciones descritas, como por ejemplo WSS y MLR (Hunt y Black, 1996; Meron y Hirose, 1999), se parte de la hipótesis de que la distancia Euclídea cepstral es una *buena* herramienta para comparar la similitud acústica entre dos realizaciones sonoras —aunque existen trabajos como los de (Campillo y Rodríguez Banga, 2003; Campillo, 2005) que ponen en duda su aplicación para la evaluación de cualquier tipo de subcoste. Asimismo, dentro de la descripción de los subcostes de selección (sección 2.1.4), se han enumerado distintos trabajos que pretenden encontrar una medida objetiva que permita correlar con la respuesta subjetiva de los usuarios, cuestión que, hasta el momento, continua siendo fuente de debate —por ejemplo, recientemente, (Cernak y Rusko, 2005) propone utilizar la medida PESQ para evaluar la

calidad sintética de los sistemas de conversión de texto en habla.

En esta misma línea existen distintos trabajos en el ámbito del entrenamiento de los pesos de la función de coste que pretenden incorporar de forma implícita la subjetividad de los usuarios, yendo más allá de las propuestas que aplican directamente una función objetiva (típicamente la distancia Euclídea cepstral). Estos trabajos, parten de la evaluación subjetiva de un conjunto de expresiones sintéticas mediante un test perceptual (puntuado mediante la escala *Mean Opinion Score* o MOS, por ejemplo) para, seguidamente, ajustar los pesos de la función de coste de manera que se maximice la correlación entre la función y las respuestas de los usuarios. Entre estos trabajos se encuentra el de Lee, Lopresti y Olive (2001), que utiliza información procedente de un test perceptual para ajustar los pesos a partir de un algoritmo de optimización multidimensional denominado *Downhill simplex*, que busca la solución mediante caminos *trapeciformes*. No obstante, el ajuste de pesos se lleva a cabo sobre palabras (monosílabicas) aisladas, cuestión que, según (Campillo y Rodríguez Banga, 2003; Campillo, 2005) hace que no sea trivial extender los resultados para toda la frase. En (Peng, Zhao y Chu, 2002), se utilizan 400 expresiones sintéticas evaluadas mediante MOS: 300 para entrenar el método de optimización de los pesos y el resto es utilizado para evaluar el funcionamiento de la técnica propuesta. En este caso, se van optimizando en bloques independientes los subcostes y sus pesos (se supone que no existe interdependencia, a diferencia de lo que se indica en (Meron y Hirose, 1999)). Después de varias optimizaciones los autores obtienen un conjunto de pesos que permite optimizar la correlación entre la función de coste (los subcostes considerados) y los resultados del test perceptual. No obstante, se indica la necesidad de continuar trabajando en esta línea ya que, por un lado, el tamaño del conjunto de entrenamiento parece insuficiente para ajustar el número de parámetros (pesos) considerado, y por otro, las expresiones consideradas deben presentar una mejor cobertura (variablidad) de los subcostes considerados.

Siguiendo un enfoque similar, en (Toda et al., 2002) se introduce un estudio muy interesante a partir de otro test perceptual evaluado mediante la escala MOS (en este caso de 7 puntos —[-3,3]—), utilizando 141 expresiones recogidas de un total de 17381 expresiones sintéticas (no pertenecen al corpus). En este caso, se estudia la correlación del resultado del test (una vez normalizada la respuesta de los 8 evaluadores participantes mediante *z-score*) respecto a tres funciones de coste distintas: el coste promedio (equivalente al descrito en la ecuación (2.3)), junto a dos variantes del mismo, denominadas coste máximo y coste normalizado (ver sección 2.1.4). El mismo autor continua trabajando en la misma línea en (Toda, Kawai y Tsuzaki, 2003), donde ajusta tanto los pesos como la potencia de la función que integra los subcostes, esta vez sobre una distribución de costes dentro de un rango menor al del primer trabajo, ya que normalmente durante el proceso de selección se escogen unidades cercanas a la unidad objetivo. En este caso las correlaciones obtenidas son peores que en el caso de trabajar con un muestreo que abarque todo el rango de costes potencialmente seleccionables del corpus. Asimismo, los autores indican la necesidad de disponer de un conjunto mayor de estímulos para llevar a cabo un ajuste robusto de todos los parámetros, en la línea de lo descrito por (Peng, Zhao y Chu, 2002) —p.ej. al trabajar con el rango de costes reducido, se obtienen pesos $w = 0$. Finalmente, Toda, Kawai y Tsuzaki (2004) proponen ir más allá en la definición de la función de coste y permitir que para cada tipo de subcoste

se utilice una función de coste apropiada, consiguiendo mejorar los resultados obtenidos, a cambio de tener que adaptar tanto el tipo de función (lineal, exponencial, sigmoidea, etc.) a cada subcoste —cuestión que no se ha tratado presente proyecto de investigación.

Finalmente, existen trabajos fundamentalmente enfocados al ajuste automático de los costes, pero que incorporan una ponderación subjetiva final. Por ejemplo, en (Campillo y Rodríguez Banga, 2003; Campillo, 2005) se presenta un esquema de optimización de los pesos de unidad y de concatenación por separado, más un ajuste subjetivo final de la importancia de ambos pesos (con una única variable) al estilo de lo descrito por (Meron y Hirose, 1999) (ver ecuación (2.18)).

$$C(t_i^n, u_i^n) = \alpha \cdot C^t(t_i, u_i) + (1 - \alpha) \cdot C^c(u_{i-1}, u_i) \tag{2.18}$$

En este caso, la percepción humana se incluye explícitamente mediante el ajuste del parámetro α de la función de coste (estrategia también utilizada por los autores para ponderar los dos grupos de subcostes —prosódicos[10] y contextuales— utilizados en la función de coste de unidad).

Siguiendo un camino similar al descrito en estos trabajos, se decidió diseñar una estrategia que permitiera la incorporación *explícita* de la percepción humana en el entrenamiento de los pesos de la función de coste —tanto del coste de unidad como de concatenación, ya que éstos están interrelacionados entre sí (Meron y Hirose, 1999). Para ello, tomando en consideración la eficiencia demostrada por los algoritmos genéticos para abordar el problema (ver los resultados de los experimentos descritos en la sección 2.4.1), se propuso intentar solucionar el problema mediante el uso de un algoritmo genético interactivo (AGI) (Alías et al., 2004a), estrategia que permite la fusión de la computación evolutiva y los criterios subjetivos para resolver el problema planteado (Caldwell y Johnston, 1991; Takagi, 2001)

El funcionamiento de un AGI se basa en los mismos criterios que se han descrito para los algoritmos genéticos (ver sección 2.2), pero incorporando la subjetividad humana en la fase de selección de los individuos, es decir, se sustituye la función de *fitness* objetiva (cuantitativa) por el *fitness* subjetivo (cualitativo) del usuario. Así pues, será el evaluador el que escogerá qué individuos le parecen mejor adaptados al problema planteado, según la calidad de la síntesis conseguida para cada una de las configuraciones de pesos evaluadas.

En este contexto, resulta necesario diseñar una aplicación que permita al usuario interactuar con el proceso de entrenamiento de los pesos de la función de coste. En este caso, se ha desarrollado una plataforma *web* bajo un entorno interactivo, que permite el acceso remoto a las pruebas para agilizar la realización de las mismas (ver anexo C.2 para más detalles). Esta estrategia empieza a ser habitual en el ámbito del desarrollo de pruebas subjetivas (Jilka y Syrdal, 2002; Black y Tokuda, 2005). Asimismo, la plataforma ha sido diseñada para poder funcionar sobre varios servidores, aumentando su capacidad para dar servicio a un mayor número de usuarios y permitiendo la distribución de las pruebas a ejecutar para

[10] En este mismo trabajo se indica que "*se supone que la importancia de cada uno de los constituyentes de la parte prosódica es la misma*", una vez se hayan "*ajustando adecuadamente los respectivos umbrales*", cuestión que el autor no comparte, observando los resultados obtenidos, justamente trabajando con una función de coste fundamentalmente prosódica.

el entrenamiento de los pesos. Además, la plataforma permite automatizar el análisis de los resultados de las pruebas, agilizando el trabajo de evaluación de las mismas. Así, se evita la tediosa tarea de trabajar con pruebas escritas que posteriormente tienen que ser revisadas, transcritas y evaluadas por el experto.

Introducción

El diseño de un sistema de CTH de alta calidad basado en corpus comporta afrontar distintas problemáticas, como ya se descrito en los diferentes apartados previos. Algunas de ellas involucran conceptos puramente objetivos y cuantificables, como los que intervienen en el bloque de PLN (ver figura 2.2), p.ej. la trascripción fonética o el etiquetado del corpus de voz (frecuencia fundamental, duración, etc.). Por otra parte, hay procesos en los que intervienen parámetros que pertenecen al dominio perceptivo de las personas, siendo estrictamente subjetivos y, por lo tanto, difícilmente medibles (ver apartado de '*Subcostes de selección*' en la sección 2.1.4). Por ejemplo, la elección de los parámetros de la señal de voz considerados en la función de coste o la importancia que éstos tienen que tomar (pesos) en el proceso de selección de unidades.

Debido a que la calidad de la síntesis basada en corpus depende en gran medida del valor que tomen estos parámetros (Black, 2002), parece interesante escoger un método que permita el ajuste eficiente de estos valores (cuantificados numéricamente) a partir de un método que permita incorporar subjetividad al proceso. Aunque este objetivo pueda parecer de entrada contradictorio, la evolución artificial (Holland, 1975), y en especial los algoritmos genéticos (Goldberg, 1989; Goldberg, 2002), permiten afrontar este propósito de una forma simple y eficiente (ver los distintos apartados que describen el funcionamiento de los algoritmos genéticos en la sección 2.2). La diferencia fundamental introducida por los AGIs se encuentra en el proceso de selección, ya que se incorpora el dominio psicológico subjetivo de los usuarios. De esta forma, el algoritmo fusiona la subjetividad del usuario con el proceso evolutivo (numérico) durante el ajuste iterativo de los pesos de selección que controlan el CTH basado en corpus.

Este tipo de algoritmos han sido aplicados a distintos ámbitos de la investigación. Por ejemplo, en gráficos por ordenador, para ingeniería mecánica o para procesamiento digital de la señal (ver (Takagi, 2001) para un amplio resumen sobre el tema). También han sido aplicados a sistemas de tratamiento del habla (Watanabe y Takagi, 1995; Sato, 1996; Sato, 1997; Todoroki y Takagi, 2000). Concretamente, los AGIs han sido utilizados para ajustar los coeficientes de distorsión de los filtros FIR utilizados en un sistema de síntesis, o bien, para el ajuste de diversos parámetros que controlan la incorporación de emociones al bloque de procesamiento digital de la señal de un sistema de síntesis. Recientemente, en (Sato, 2005) se ha vuelto a aplicar los algoritmos genéticos interactivos, en este caso para controlar la prosodia en el proceso de conversión de la calidad vocal del habla. Así pues, existen varios precedentes exitosos de la integración de estos dos ámbitos del conocimiento: los algoritmos genéticos interactivos y la síntesis del habla.

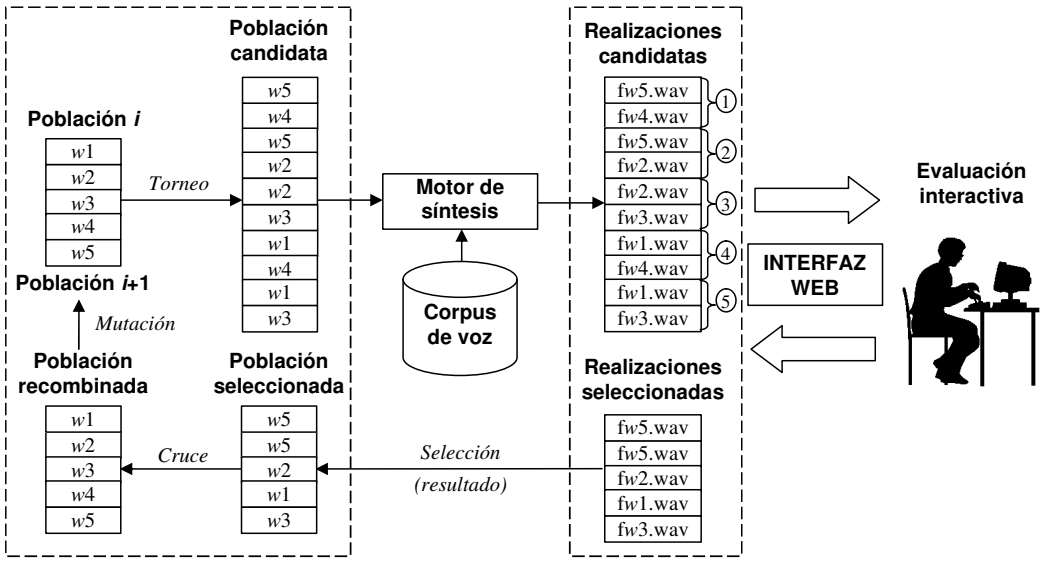


Figura 2.13: Diagrama del ajuste subjetivo de los pesos de la función de coste basado en un algoritmo genético interactivo, donde los índices marcados con círculo que acompañan a las frases sintéticas representan los distintos torneos binarios presentados al usuario.

Adaptación de los AGIs al ajuste de pesos de la función de coste

El algoritmo genético interactivo utilizado en este trabajo de investigación se ha diseñado para el entrenamiento de los pesos (w^t y w^c, en las ecuaciones (2.1) y (2.2), respectivamente) que ponderan los subcostes de la función de selección de unidades de un CTH basado en corpus. La figura 2.13 representa el funcionamiento completo del algoritmo. Como se puede observar, la estructura del algoritmo es la misma que la de un algoritmo genético convencional (ver figura 2.10). En este caso, igual que en el AG diseñado para el ajuste objetivo de pesos, el AGI evoluciona hacia un vector de individuos que corresponde a los pesos utilizados en la función de coste (ecuación (2.3)), hasta llegar a una configuración *óptima*: $\mathcal{W}^* = (w_1^{t*}, \dots w_p^{t*}, w_1^{c*}, \dots w_q^{c*})$.

La inicialización del conjunto de posibles soluciones se realiza de forma aleatoria (al igual que para el AG). A continuación, se inicializa el proceso evolutivo, que se divide en dos fases. Primero, se seleccionan los mejores individuos dentro de la población y, después, se recombinan (cruce y mutación) para generar la nueva población. Es en el proceso de selección donde los AGIs se diferencian de un AG clásico. En este caso, la función matemática de evaluación (*fitness*) queda sustituida por la percepción del evaluador. En cada una de las iteraciones del proceso de ajuste, el algoritmo dispone de un conjunto de vectores de pesos $\mathcal{W}$ (posibles soluciones) para la síntesis de la expresión considerada. El resultado de esta síntesis (realizaciones candidatas, en la figura 2.13) es evaluado interactivamente por el usuario. A éste se le presentan los resultados de la síntesis en parejas agrupadas aleatoriamente, según el

operador de *torneo binario* (en la figura 2.13 se representa dentro de la población candidata con las parejas de frases sintéticas: (fw5, fw4), (fw5, fw2), (fw2, fw3), y así sucesivamente hasta el final) —el escenario mínimo para recoger información significativa del usuario es utilizando un esquema de torneo binario ($s = 2$) (Goldberg, Korb y Deb, 1989). Así pues, el evaluador debe escoger cuál de las dos realizaciones de la misma expresión prefiere. De esta manera, escogerá, después de escuchar las veces que crea necesarias los ficheros, la solución (vector de pesos $\mathcal{W}$) que presente una mejor calidad de voz desde el punto de vista subjetivo. En esta aplicación, los pesos actuarán de *genotipo* (información genética de la solución) y los ficheros de voz actuarán de *fenotipo* (representación externa del individuo o realización acústica, en este caso), utilizando términos del argot de los algoritmos genéticos.

El proceso de recombinación del AGI (representado por los operadores de cruce y mutación) sigue los mismos conceptos que han sido descritos en el apartado de algoritmos genéticos. Siguiendo la misma filosofía, para la implementación del AGI se ha escogido como operador de cruce el clásico de punto único (*1-point*) (Goldberg, 1989). En este caso, las probabilidades de mutación y cruce (p_m y p_c, respectivamente) son mayores que las utilizadas por el algoritmo genético que trabaja con una función de *fitness* objetiva (Alías y Llorà, 2003). El objetivo de este aumento es compensar la notable disminución de individuos en la población, ya que resulta inviable que un usuario deba evaluar en cada iteración una población del tamaño típicamente utilizado en los sistemas de optimización basados en algoritmos genéticos (p.ej. $n = 200$ individuos) —provocaría la saturación perceptiva y la fatiga del usuario. En cuanto al criterio de finalización del ciclo evolutivo, será el evaluador quien decida cuándo se debe dar por concluido el proceso iterativo. Es decir, el evaluador determina cuándo le resulta imposible detectar la mejora conseguida en las realizaciones sintéticas propuestas (convergencia del algoritmo a nivel subjetivo). En el anexo C.2 se describe la implementación modular de la plataforma de pruebas que integra todos los elementos que forman parte del entrenamiento subjetivo de los pesos de la función de coste —en el trabajo de (Formiga, 2003) se puede encontrar una descripción técnica más detallada de esta plataforma.

2.3. Ajuste subjetivo de pesos mediante algoritmos genéticos interactivos *activos*

En este apartado se describe el método final desarrollado para optimizar el proceso de ajuste subjetivo de los pesos basado en algoritmos genéticos interactivos activos (aAGI) (Alías et al., 2006). Este método consigue minimizar la fatiga del usuario —y favorecer la velocidad de la convergencia del proceso—, mediante la obtención de un modelo que aproxima los criterios subjetivos que el usuario utiliza en las comparativas subjetivas. La *fiabilidad* de las decisiones del usuario, es decir, la calidad de las soluciones proporcionadas, es un elemento clave para la aplicación satisfactoria de los algoritmos genéticos interactivos (AGIs) —o cualquier otro método de computación evolutiva interactivo (Takagi, 2001). Las decisiones que el usuario toma cuando utiliza un AGI guían la búsqueda de soluciones a través del espacio de posibles hipótesis. Cualquier procedimiento interactivo, independien-

temente de su grado de eficacia, perderá eficiencia en la búsqueda de buenas soluciones si el usuario es incapaz de proporcionar evaluaciones consistentes (Llorà et al., 2005), cuestión altamente probable si las pruebas no son cortas y simples, ya que provocan un aumento de su fatiga y una reducción de la fiabilidad de sus respuestas (Campillo, 2005). Recientemente, los aAGI también han sido aplicados al ajuste interactivo de los parámetros prosódicos para conversión de texto en habla expresiva (Oversdotter y Llorà, 2006).

En este apartado, se describe cómo aplicar los aAGI al problema del ajuste de los pesos de la función de coste, introduciendo a la vez una medida que permite evaluar la consistencia de los usuarios a lo largo del proceso evolutivo. Esta información es fundamental para conocer el grado de calidad de los resultados obtenidos por cada usuario, como se describe en los experimentos desarrollados.

2.3.1. Introducción

En el apartado 2.2 se ha descrito cómo los algoritmos genéticos (AG) han sido aplicados al problema del entrenamiento automático de pesos. Esta técnica permitió superar las restricciones de las aproximaciones clásicas (Hunt y Black, 1996; Meron y Hirose, 1999) —trabajando sin restricciones lineales, p.ej. como en (Park, Kim y Kim, 2003; Campillo, 2005), y presentando un coste computacional razonable—, alcanzando mejores resultados en términos de la función de *fitness* definida (ver los experimentos de la sección 2.4.1). Sin embargo, esta aproximación, como todas las técnicas anteriores (descritas en la sección 2.1.5), necesita enfrentarse a un reto clave: la estimación fiable de la percepción subjetiva de los atributos de habla (es muy difícil de definir una función de representación de percepción sólida)(Lee, Lopresti y Olive, 2001; Peng, Zhao y Chu, 2002; Campillo, 2005; Zhao et al., 2006). Después de ese primer trabajo, se decidió incorporar de forma explícita las preferencias de los usuarios en el entrenamiento de los pesos de la función de coste. Como primer paso, se aplicó un algoritmo genético interactivo (AGI) sencillo para el entrenamiento subjetivo de los pesos, permitiendo un ajuste guiado a partir de la percepción real. No obstante, después de analizar los resultados obtenidos (ver sección 2.4.1), se observó que dejar el proceso de evaluación en las manos de un usuario crea un marco diferente comparado con las tareas de optimización clásicas (ver (Takagi, 2001) para una descripción detallada). Principalmente, se observaron dos problemas fundamentales: lo tedioso del proceso de ajuste interactivo (provoca la fatiga del usuario) y la complejidad de mantener un criterio de comparación estable a lo largo del todo el proceso de ajuste (consistencia del usuario), que son problemas relacionados típicamente con los AGIs (Takagi, 2001). A partir del análisis y discusión de los resultados obtenidos con el AGI, se observó la necesidad de continuar investigando para optimizar el proceso de ajuste subjetivo de los pesos, y combatir, así, la fatiga de los usuarios.

Un tiempo después, en el trabajo de Llorà et al. (2005) se describen las líneas maestras para abordar la consecución de un proceso de ajuste interactivo *competente*. Concretamente, en este trabajo se indica que resulta necesario abordar cinco elementos fundamentales: *(i)* una definición clara del objetivo (permite fijar la atención del usuario), *(ii)* el impacto de la visualización del problema (es esencial centrar la atención del usuario en los elementos

fundamentales de la comparativa), *(iii)* la falta de un *fitness* objetivo (naturaleza cualitativa del AGI frente a naturaleza cuantitativa del AG), *(iv)* fatiga del usuario (provoca frustración y malas soluciones), *(v)* consistencia del criterio del usuario (es fundamental incorporar estrategias que controlen la fiabilidad de sus respuestas). En mismo trabajo se abordan estos problemas mediante la definición un nuevo paradigma evolutivo: *los algoritmos genéticos interactivos activos* (aAGI), con el objetivo de combatir la fatiga de usuario a partir del modelado de su criterio de comparación mediante una función aproximante (en inglés, *surrogate function*). Concretamente, los aAGIs se basan en aprender de la interacción con el usuario para anticipar las hipótesis (posibles soluciones) en las que éste estará más interesado en la siguiente iteración (en inglés, *educated guesses*), guiando así el proceso que generación de nuevas soluciones (ver algoritmo 1 (Llorà et al., 2005)). De este modo, se consigue acelerar el proceso de convergencia del algoritmo interactivo, reduciendo la probabilidad de fatigar al usuario[11]. En este trabajo, se demuestra que, aprendiendo de la interacción con el usuario y explotando el conocimiento obtenido para guiar el proceso interactivo, se consigue reducir de forma evidente el número de evaluaciones necesarias para llegar a soluciones de gran calidad.

Algoritmo 1 Descripción algorítmica del modelo aAGI —adaptada de (Llorà et al., 2005)—, donde h es la altura del árbol de torneos ($h = log_2(n)$, siendo n el tamaño de la población $\mathcal{S}$) y $\hat{r}(v)$ es el *ranking* estimado del vértice v.

1: Crear un grafo dirigido vacío $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$, de $\mathcal{V}$ vértices y $\mathcal{E}$ conexiones
2: Crear 2^h soluciones iniciales aleatorias (conjunto $\mathcal{S}$)
3: Crear el conjunto de torneo jerárquico $\mathcal{T}$ utilizando las soluciones disponibles en $\mathcal{S}$
4: Presentar los torneos de $\mathcal{T}$ al usuario y actualizar el orden parcial de $\mathcal{E}$
5: Estimar $\hat{r}(v)$ para cada vértice $v \in \mathcal{V}$ una vez ordenados según el orden parcial de $\mathcal{E}$
6: Entrenar el *fitness* sintético mediante ε-SVM basado en $\mathcal{G}'$ y $\hat{r}(v)$
7: Optimizar el *fitness* sintético obtenido de ε-SVM utilizando el un algoritmo genético
8: Crear un conjunto $\mathcal{S}'$ con 2^{h-1} soluciones diferentes, donde $\mathcal{S} \cap \mathcal{S}' = \emptyset$ muestreando el modelo probabilístico evolucionado por el algoritmo genético utilizado —p.ej. compact GA (Harik, Lobo y Goldberg, 1999)
9: Crear un conjunto de torneo jerárquico $\mathcal{T}'$ con $2^h - 1$ torneos, utilizando las $\frac{n}{2}$ mejores soluciones de $\mathcal{S}$ y las $\frac{n}{2}$ mejores soluciones de $\mathcal{S}'$
10: $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{S}'$
11: $\mathcal{T} \leftarrow \mathcal{T} \cup \mathcal{T}'$
12: Volver al punto 4 mientras no converja

El aAGI está inspirado básicamente en el AGI propuesto en (Takagi, 2001), cuyo funcionamiento se basa en cinco elementos clave: *(i)* inicialización (generalmente aleatoria) de la población de n individuos, *(ii)* presentación al usuario de soluciones agrupadas aleatoriamente en grupos de s posibilidades (típicamente $s = 2$ o torneo binario —$\mathcal{T}$ en el algoritmo 1), *(iii)* selección de los mejores individuos según el criterio subjetivo del usuario, *(iv)* cruce de los individuos y su posterior mutación para volver al paso *(ii)* mientas el proceso

[11] Gracias a modelar la respuesta del usuario, los aAGIs pueden estimar qué soluciones serán mejor valoradas por éste, evitando repetir comparaciones sobre el espacio de optimización.

no converja[12]. Si se observa el aAGI propuesto por (Llorà et al., 2005), se pueden ver las diferencias que esta estrategia presenta respecto a los AGI clásicos. Principalmente, desaparecen las etapas de *cruce* y de *mutación* debido a que los nuevos individuos son creados mediante un proceso automático guiado por la subjetividad del usuario (puntos `5` a `8` del algoritmo 1). Este proceso automático se divide en dos fases: primero, se modela el criterio del usuario a partir de la ordenación que ha indicado de las soluciones que ha evaluado (puntos `5` y `6`) para, a continuación, evolucionar la población para encontrar nuevos individuos mejor adaptados al problema (en este caso, al criterio subjetivo del usuario) mediante un algoritmo genético automático (puntos `7` y `8`). Una vez realizada la evolución del modelo probabilístico, se añaden los $\frac{n}{2}$ mejores individuos evolucionados a los $\frac{n}{2}$ mejores individuos de la iteración anterior (punto `9`). Esta combinación es la nueva población que se agrupa jerárquicamente para que pueda ser evaluada de nuevo por el usuario (puntos `10` a `11` del algoritmo 1). Asimismo, las soluciones se presentan al usuario siguiendo también un torneo binario $\mathcal{T}$, organizadas jerárquicamente para permitir obtener una ordenación global de los resultados a partir del orden parcial generado por el usuario a lo largo de las comparativas (ver ejemplo de la figura 2.14).

Concretamente, la propuesta de (Llorà et al., 2005) se basa en dos elementos fundamentales:

1. **Ordenación de las evaluaciones:** cuando el usuario decide sobre la bondad de las soluciones vía comparaciones relativas (torneos binarios), se establece un orden parcial del conjunto de soluciones evaluado. A partir de este orden parcial, y utilizando los conceptos de *no dominancia* y *dominancia* procedentes de los algoritmos evolutivos multiobjetivo (Coello-Coello, December, 1998; Deb et al., 2000), se puede inducir el orden completo de las soluciones de la población —en este caso, los torneos se organizan jerárquicamente para disponer del número mínimo de comparaciones necesario para deducir las relaciones entre todas las soluciones (ver ejemplo de la figura 2.14).

2. **Función aproximante (*surrogate function*):** el orden global inducido puede ser utilizado para asignar puntuaciones a las soluciones mediante su ordenación (*ranking*). Esta información se utiliza para modelar (generalizar) la respuesta del usuario mediante la generación de un *fitness sintético*, capaz de evaluar las nuevas soluciones acorde con las respuestas del usuario hasta ese momento. Esta función se irá actualizando a medida que el usuario vaya evaluando nuevos individuos a lo largo del proceso evolutivo.

Gracias a disponer de una función de *fitness* que aproxima el criterio subjetivo, se disminuye el tiempo global necesario para la convergencia del proceso evolutivo —es decir, el tiempo de evaluación típico de un usuario. Esta estrategia definida por el aAGI permite presentar al usuario soluciones potenciales de alta calidad —se añaden los $\frac{n}{2}$ mejores individuos evolucionados a los $\frac{n}{2}$ mejores individuos de la iteración anterior—, según su criterio,

[12]Indicado por el usuario cuando decide que no es capaz de percibir mejora alguna o el algoritmo detecta que no existe variación de fenotipos, es decir, se vuelven a seleccionar las mismas unidades.

a partir de los torneos realizados hasta el momento[13]. La aproximación propuesta proporciona los siguientes beneficios: *(i)* permite trabajar con una población con mayor número de individuos para un mejor barrido del espacio de soluciones, a la vez que, *(ii)* acelera el proceso de convergencia del algoritmo genético interactivo, reduciendo la fatiga del usuario. Además, este método proporciona soluciones potencialmente de alta calidad para la evaluación del usuario, reduciendo el riesgo de frustración. Por lo tanto, este modelo proporciona una herramienta útil para extraer y modelar las preferencias del usuario.

Ordenación de las evaluaciones del usuario

Existen distintas técnicas para recoger las evaluaciones subjetivas del usuario en un proceso interactivo basado en AGI, por ejemplo: la puntuación de las soluciones, medida de la calidad de las soluciones en una escala de valores, p.ej. MOS (*Mean Opinion Score*) (Takagi, 2001) —ver (Llorà et al., 2005) para una revisión de las bases del proceso. La selección por torneo utilizada en el experimento de ajuste de pesos mediante AGI, es una de las técnicas más empleadas por este tipo de algoritmos interactivos. Bajo este enfoque, se presenta al usuario comparativas de dimensión s seleccionado los individuos de la población $\mathcal{S}$ de tamaño n. Este tipo de esquema de selección sustituye la necesidad de recoger evaluaciones numéricas del usuario por evaluaciones comparativas, lo cual permite aplicar los conceptos evolutivos de los algoritmos genéticos a esquemas de evaluación interactivos. Como se ha comentado, el escenario mínimo para recoger información significativa del usuario es utilizando un esquema de torneo binario ($s = 2$) (Goldberg, Korb y Deb, 1989), en el que se presentan dos opciones al usuario para que escoja la que le parezca mejora adaptada al problema (en este caso, la que presente mejor calidad sintética respecto a la secuencia objetivo). Asimismo, gracias a organizar las soluciones jerárquicamente, las evaluaciones del usuario introducen un orden parcial entre las que le han sido presentadas hasta ese momento con el mínimo número de comparaciones necesarias (se evita tener que comparar todos los individuos contra todos) (ver figura 2.14).

Grafo de ordenación parcial: En el trabajo de (Llorà et al., 2005) se propone modelar la ordenación parcial de las soluciones mediante un grafo de orden parcial $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$[14]. Cada vértice en $\mathcal{V}$ representa la solución presentada al usuario, mientras que las conexiones (*edges*, inglés) en $\mathcal{E}$ representan la ordenación parcial de las soluciones según el usuario (ver ejemplo en la figura 2.19). En este contexto, dadas dos soluciones $\{s_1, s_2\} \in \mathcal{V}$, el usuario puede dar al sistema tres respuestas distintas: *(i)* $s_1 > s_2$, *(ii)* $s_1 < s_2$, y *(iii)* $s_1 = s_2$— o *son iguales/no sabe/no importa*. La ordenación de torneos presentada en la figura 2.14 garantiza que el orden parcial introducido por las evaluaciones del usuario produzca un grafo conectado $\mathcal{G}$. Este grafo $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$ representa el orden parcial de las evaluaciones,

[13] El número de comparaciones por generación es similar al necesario utilizando AGI simple, sin embargo, el número de iteraciones necesario para encontrar la solución óptima, es mucho menor —ver sección 2.4.2.

[14] Durant et al. (2004) también intentan ensamblar globalmente las ordenaciones obtenidas a partir de comparaciones de parejas. Sin embargo, los autores del trabajo no intentan modelar las respuestas del usuario para reducir la fatiga que el proceso interactivo de ajuste puede causar.

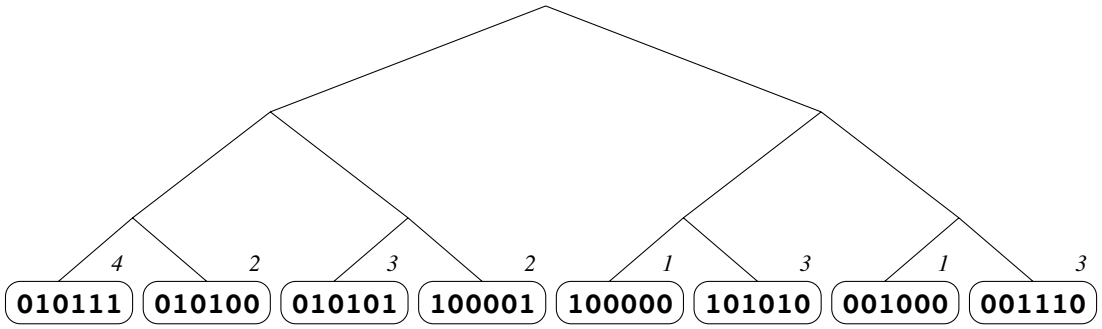
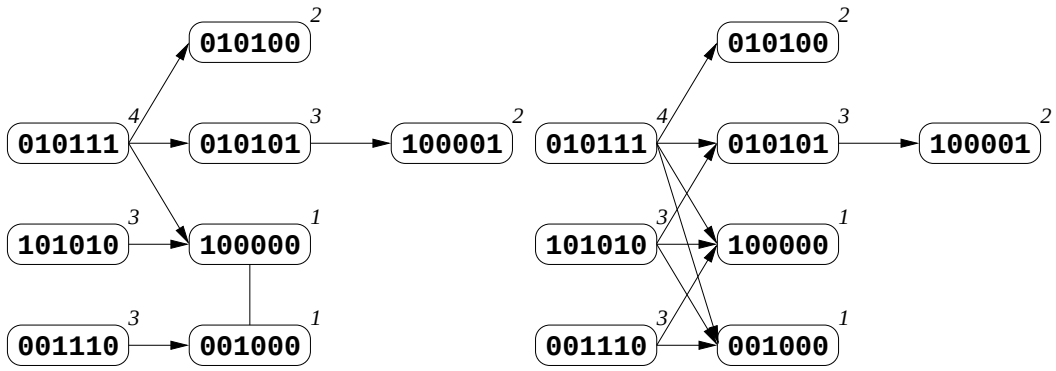


Figura 2.14: Ejemplo de torneo con ocho individuos escogidos aleatoriamente de la población de soluciones, agrupados en siete torneos distintos {(010111, 010100), (010101, 100001), (100000, 101010), (001000, 001110), (010111, 010101), (100000, 001000), (010111, 100000)}. El número que acompaña a cada individuo representa el resultado de la función de *fitness* (en este caso el número de bits activos).

representando las soluciones como vértices en $\mathcal{V}$, y las relaciones de comparación entre pares de individuos (mayor, menor, o igual) representadas como conexiones en $\mathcal{E}$. El grafo de ordenación parcial obtenido del usuario puede ser no dirigido (se permiten evaluaciones iguales), sin embargo, este grafo $\mathcal{G}$ puede ser transformado —bajo ciertas restricciones— en un grafo dirigido normalizado $\mathcal{G}'$, como muestra la figura 2.15. El grafo dirigido se obtiene al sustituir las relaciones de igualdad (conexiones no dirigidas) por las relaciones *mayor que* o *menor que* adecuadas —ver algoritmo 2. En la figura 2.14 se puede observar como existe una conexión de igualdad entre los individuos 100000 y 001000 que es reformulada a relaciones *mayor que* (indicadas por la dirección de las flechas) en el la figura 2.15(b).

Tabla 2.1: Estimación del *ranking* global de los individuos basada en los operadores de dominancia, obtenida a partir del orden parcial representado en la figura 2.15(b).

v	$f(v)$	$r(v)$	$\delta(v)$	$\phi(v)$	$\hat{f}(v)$	$\hat{r}(v)$
010111	4	1	5	0	5	1
010100	2	3	0	1	-1	4
010101	3	2	1	1	0	3
100001	2	3	0	2	-2	5
100000	1	4	0	3	-3	6
101010	3	2	2	0	2	2
001000	1	4	0	3	-3	6
001110	3	2	2	0	2	2



(a) Grafo de ordenación parcial obtenido de los torneos representados en la figura 2.14

(b) Grafo de ordenación parcial equivalente donde las relaciones de igualdad han sido sustituidas por las relaciones *mayor que* correspondientes.

Figura 2.15: Grafo de ordenación parcial proporcionado por las comparaciones realizadas por un usuario a partir de los torneos indicados en la figura 2.14. La dirección de las flechas indica las relaciones *mayor que* entre las soluciones comparadas. Las conexiones sin dirección (sin flecha) corresponden a relaciones de igualdad entre las soluciones.

Función aproximante: estimación del *fitness* subjetivo

Típicamente, los algoritmos genéticos convencionales trabajan con poblaciones de centenares de individuos, que van evolucionando a través de las iteraciones hasta encontrar una buena solución al problema. Sin embargo, en los algoritmos genéticos interactivos, resulta inviable trabajar con este volumen de individuos —y, consecuentemente, de evaluaciones. Por ejemplo, si se trabaja con una población de $n = 200$ individuos que evoluciona a lo largo de 10 generaciones (iteraciones), el usuario debería realizar 2000 evaluaciones, considerando que sólo evaluara una vez cada solución (p.ej. si se trabaja con una filosofía de 10-*fold cross validation*, este valor debería multiplicarse por 10). Por lo tanto, parece muy complicado conseguir trabajar con este volumen de datos sin causar fatiga y frustración al usuario.

Existen varias técnicas que pueden ser aplicadas para superar este problema, entre ellas, la utilizada por el AGI descrito en la sección 2.2.3, donde se reduce el tamaño de la población de individuos a cambio de aumentar la probabilidad de cruce y mutación de los individuos. Técnicas como la paralelización, la continuación en el tiempo, la relajación de la evaluación o la hibridación son aproximaciones útiles para reducir el tiempo de convergencia del algoritmo y así combatir la fatiga del usuario (Goldberg, 2002). No obstante, todas ellas trabajan sobre funciones de *fitness* totalmente objetivas, cuestión que no tiene tanto sentido en el ámbito de los algoritmos genéticos interactivos. Además, la naturaleza cualitativa de las evaluaciones en un AGI abarca otros conceptos que también se deben tener en cuenta. Por ejemplo, si un AGI involucra dos evaluadores paralelos, entonces el AGI necesita compaginar dos criterios diferentes de evaluación subjetiva. Desafortunadamente, no se puede dar por buena ninguna

Algoritmo 2 Algoritmo de normalización de $\mathcal{G}$ en $\mathcal{G}'$ para eliminar relaciones de igualdad, donde $e(\cdot,\cdot)$ representa la conexión (flecha) entre dos vértices.

proceso *normalizeGraph*($\mathcal{G} =< \mathcal{V}, \mathcal{E} >$)

1: Crear el conjunto de empates
$\mathcal{D}= \{\forall e(v_1, v_2) \in \mathcal{E} : \exists e(v_2, v_1) \in \mathcal{E} \Rightarrow e(v_1, v_2) \subseteq \mathcal{D}\}$
2: Crear el conjunto vacío $\mathcal{E}_\mathcal{N}$ de nuevas flechas
3: Copiar los caminos que llegan a los primeros ítems en los segundos ítems:
$\mathcal{E}_\mathcal{N} \leftarrow \mathcal{E}_\mathcal{N} \cup \mathcal{E}_\mathcal{I}= \{\forall e(v_1, v_2) \in \mathcal{D}\text{: } \forall v^i | \exists e(v^i, v_1) \in \mathcal{E} \Rightarrow e(v^i, v_2) \subseteq \mathcal{E}_\mathcal{I}\}$
4: Copiar los caminos que llegan a los segundos ítems en los primeros ítems:
$\mathcal{E}_\mathcal{N} \leftarrow \mathcal{E}_\mathcal{N} \cup \mathcal{E}_\mathcal{I}= \{\forall e(v_1, v_2) \in \mathcal{D}\text{: } \forall v^i | \exists e(v^i, v_2) \in \mathcal{E} \Rightarrow e(v^i, v_1) \subseteq \mathcal{E}_\mathcal{I}\}$
5: Copiar los caminos que parten de los primeros ítems en los segundos ítems:
$\mathcal{E}_\mathcal{N} \leftarrow \mathcal{E}_\mathcal{N} \cup \mathcal{E}_\mathcal{I}= \{\forall e(v_1, v_2) \in \mathcal{D}\text{: } \forall v^i | \exists e(v_1, v^i) \in \mathcal{E} \Rightarrow e(v_2, v^i) \subseteq \mathcal{E}_\mathcal{I}\}$
6: Copiar los caminos que parten de los segundos ítems en los primeros ítems
$\mathcal{E}_\mathcal{N} \leftarrow \mathcal{E}_\mathcal{N} \cup \mathcal{E}_\mathcal{I}= \{\forall e(v_1, v_2) \in \mathcal{D}\text{: } \forall v^i | \exists e(v_2, v^i) \in \mathcal{E} \Rightarrow e(v_1, v^i) \subseteq \mathcal{E}_\mathcal{I}\}$
7: $\mathcal{G}' =< \mathcal{V}', \mathcal{E}' > \leftarrow < \mathcal{V}, \emptyset >$
8: $\mathcal{E}' \leftarrow \mathcal{E} \cup \mathcal{E}_\mathcal{N}$
9: $\mathcal{E}' \leftarrow \mathcal{E}' - \mathcal{D} - \mathcal{D}_\mathcal{N}= \{\forall e(v_1, v_2) \in \mathcal{D} : e(v_2, v_1) \subseteq \mathcal{D}_\mathcal{N}\}$
10: *normalizeGraph* $\leftarrow \mathcal{G}'$

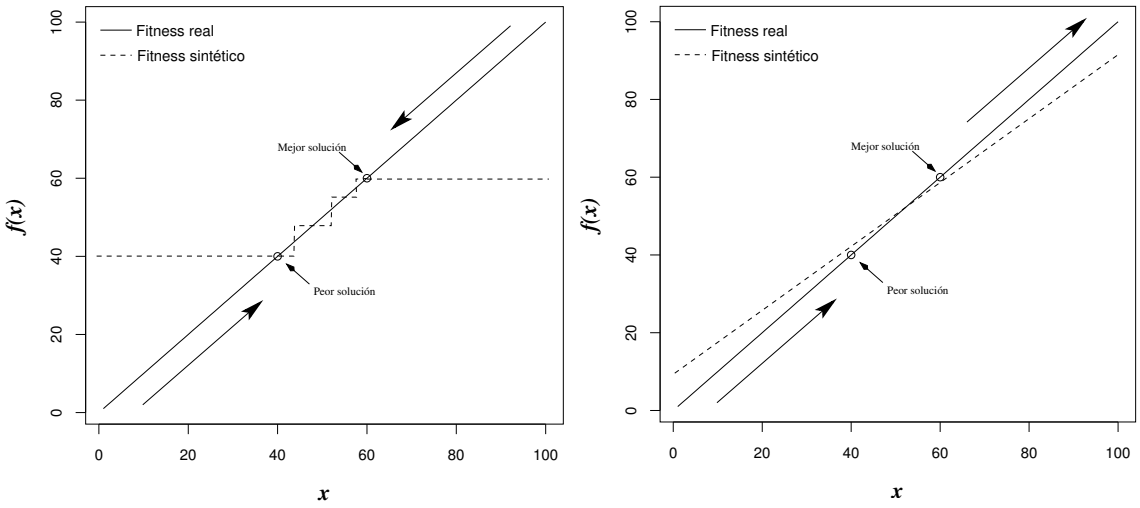
tesis sobre la coherencia de este criterio a través de los evaluadores paralelos. Para solventar estas situaciones se suelen aplicar aproximaciones multiobjetivo y/o multimodales.

Para obtener un *fitness* subjetivo, Llorà et al. (2005) proponen una heurística calculada partir del grafo dirigido normalizado $\mathcal{G}'$ obtenido de la ordenación parcial de las evaluaciones del usuario. Esta heurística está basada en el concepto de dominancia de Pareto (1896), extraído del ámbito de la optimización multiobjetivo (Coello-Coello, December, 1998; Deb et al., 2000). Concretamente se utilizan dos medidas de dominancia aplicadas sobre los vértices de grafo: $\delta(v)$ y $\phi(v)$. $\delta(v)$ se define como el número de nodos diferentes por los que pasan las flechas que salen del vértice v (p.ej. en la figura 2.15(b), $\delta(\texttt{010111}) = 5$, ya que este vértice domina a los individuos `010100`, `010101`, `100000`, `001000`, así como `100001`, a través de dominar a `010100`). Análogamente, $\phi(v)$ se define como el número de nodos diferentes de donde parten las flechas que llegan a v (p.ej. en la figura 2.15(b), $\phi(\texttt{100001}) = 2$, debido a que este vértice es dominado por los individuos `010100` y `010111`). Por lo tanto, el *fitness* de una determinada solución v puede ser calculado como $\hat{f}(v) = \delta(v) - \phi(v)$. Intuitivamente, cuanto mayor sea el número de soluciones que un vértice v domina (sea *mayor que*), mayor será su *fitness* (mayor preferencia del usuario). De lo contrario, cuantas más soluciones dominen (sean *mayores que*) una solución v, menor será su *fitness* (menor preferencia del usuario). Esta información es muy útil para inducir la ordenación global de las soluciones candidatas o *ranking* $r(v)$. El *ranking* global final estimado $\hat{r}(v)$ se obtiene después de ordenar todos los vértices $v \in \mathcal{V}$ según su $\hat{f}(v)$ (o *fitness* subjetivo). En la tabla 2.1 se presenta un ejemplo los valores $\delta(v)$, $\phi(v)$, $\hat{f}(v)$ y $\hat{r}(v)$ a partir del grafo normalizado de soluciones de la figura 2.15(b), teniendo en cuenta el problema *OneMax*, problema de referencia típico en el que se busca conseguir que todos los genes de los individuos converjan a '1' (Llorà et

al., 2005), por lo que el *fitness* objetivo $f(v)$ se computa contabilizando el número de genes activos del individuo. Como resultado, se puede observar como la estimación de ordenación $\hat{r}(v)$ sigue perfectamente el *ranking* indicado por el *fitness* $f(v)$. A continuación, resulta necesario generalizar este *fitness* subjetivo recopilado de las evaluaciones parciales del usuario a un *fitness* sintético capaz de definir una función aproximante de los criterios subjetivos del usuario, como se detalla a continuación.

Propiedades de un *fitness* sintético: Según (Llorà et al., 2005), la generación de un *fitness* sintético basado en el orden parcial de las soluciones proporcionado por el usuario debe satisfacer, al menos, dos propiedades: *(i) extrapolación de fitness* subjetivo y *(ii) mantenimiento del orden*. La primera propiedad (*extrapolación de fitness* subjetivo) requiere que el *fitness* sintético infiera la bondad de los individuos más allá de los límites del orden parcial que ha proporcionado el usuario. Como el usuario evalúa sólo una parte de la población de individuos, es necesario realizar una extrapolación de la medida deducida de las evaluaciones realizadas para poder estimar el perfil de soluciones en las que el usuario estará más interesado para guiar la creación de la nueva población de individuos. La segunda propiedad (*mantenimiento del orden*) garantiza que el *fitness* sintético mantenga el orden parcial que viene dado por las decisiones del usuario recopiladas hasta el momento.

Anteriormente al trabajo de (Llorà et al., 2005), han habido otros trabajos que han intentado generar el *fitness* sintético utilizando un modelo basado en el algoritmo *Nearest Neighbour* (Takagi, 2001). No obstante, estos modelos no son capaces de satisfacer las propiedades mencionadas anteriormente, puesto que dan por bueno el hecho que dado un conjunto de soluciones evaluadas por el usuario, el *fitness* de una nueva solución se puede estimar como el *fitness* de las soluciones más próximas en el espacio de búsqueda —utilizando algún tipo de métrica por ejemplo, el promedio de las k soluciones más próximas (*k-Nearest Neighbour*) (Duda, Hart y Stork, 2001)—, cuestión que viola el principio de mantenimiento del orden (Llorà et al., 2005) (ver ejemplo de la figura 2.16(a)). En cambio, Llorà et al. (2005) proponen utilizar un modelo basado en un regresor (como por ejemplo, una Red Neuronal (Takagi, 2001) o CART (Breiman et al., 1984)). Concretamente, utilizan un modelo de regresión generalizado, la regresión ε-insensitiva basada en máquinas de soporte vectorial (ε-SVM) que realiza un aprendizaje supervisado de los datos (en este caso de las soluciones v respecto a $\hat{f}(v)$, que corresponden a x y $f(x)$ en la figura 2.16), tanto de clasificación como de regresión —para una descripción detallada de ε-SVM, ver (Cristianini y Shawe-Taylor, 2000)). Las características de este regresor, trabajando con un *kernel* lineal, garantizan que el *fitness* sintético generado satisfaga las propiedades de extrapolación de *fitness* y de mantenimiento de orden necesarias para generalizar satisfactoriamente el criterio subjetivo del usuario obtenido de los individuos que éste ha evaluado (ver figura 2.16(b)). Además, gracias a la delimitación del espacio utilizando los vectores de soporte, el regresor ε-SVM garantiza el orden adecuado de las soluciones bajo un esquema de torneo por selección, incluso cuando se produce un elevado porcentaje de error de regresión —ver (Llorà et al., 2005) para más detalles sobre todas estas conclusiones.



(a) Generalización del *fitness* mediante *Nearest Neighbour*.

(b) Generalización del *fitness* mediante regresión.

Figura 2.16: Capacidad de generalización de los métodos de obtención de la función aproximante del *fitness* subjetivo de usuario estudiados por (Llorà et al., 2005).

Algoritmo genético compacto (cGA)

Finalmente, una vez obtenido el modelo de *fitness* sintético mediante el regresor ε-SVM (función aproximante) se procede a generar la nueva población de individuos para la siguiente generación —de la que se escogen los $\frac{n}{2}$ mejores individuos nuevos para acompañar a los $\frac{n}{2}$ individuos evaluados que más han satisfecho al usuario, reemplazando los peores individuos de la población de la generación anterior (ver algoritmo 1). Para ello, el AGI activo propuesto por (Llorà et al., 2005) hace uso de un algoritmo genético compacto (en inglés, *compact GA* o cGA) (Harik, Lobo y Goldberg, 1999), dado el marco del problema teórico al que se aplica: *OneMax*, problema de referencia del mundo de los AG, en el que se busca conseguir que todos los genes de los individuos binarios converjan a '1'. El cGA es una optimización del algoritmo denominado *Population Based Incremental Learning* (PBIL) desarrollado por (Baluja y Caruana, 1995) para trabajar sobre problemas binarios. PBIL trabaja sobre un vector de distribuciones de probabilidad que dan lugar a los nuevos individuos, en lugar de trabajar directamente sobre las soluciones (estadísticas evolutivas *vs.* poblaciones evolutivas). En un algoritmo genético clásico (AG), teóricamente, cuanto mayor sea el tamaño de la población, mayor será la probabilidad de encontrar la mejor solución al problema. En las pruebas efectuadas por (Baluja y Caruana, 1995) se observa que un AG con un tamaño de población muy grande se comporta mejor que el PBIL a la hora de encontrar la solución óptima. No obstante, trabajar con poblaciones muy grandes provoca una ralentización exagerada de la convergencia del problema —alto coste computacional—, mientras que PBIL

es capaz de obtener buenas soluciones aún trabajando con pocos individuos en la población —cuestión fundamental en el contexto de interacción con el usuario.

En el caso del cGA, cada elemento del vector representa, pues, la probabilidad de activación de su gen correspondiente (y consecuentemente de desactivación). El vector patrón de probabilidad de la población se utiliza para guiar y búsqueda de la mejor solución, controlando el proceso de generación de las nuevas soluciones candidatas a incorporar en la nueva población $\mathcal{S}'$, según lo indicado en el algoritmo 1 (ver (Llorà et al., 2005) para más detalles).

2.3.2. Adaptación de los aAGI al problema

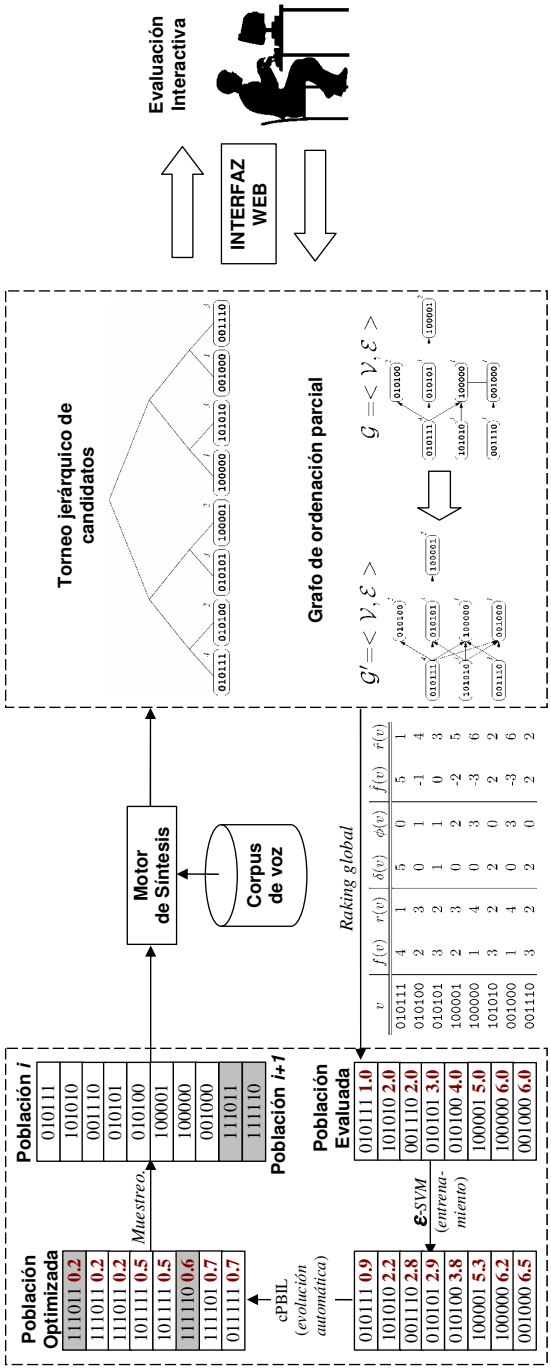
En la figura 2.17 se presenta el diagrama de bloques del proceso de entrenamiento de los pesos de la función de coste (ecuación (2.3)) basado en un algoritmo genético interactivo activo, adaptado a las características del problema. Como se puede observar, en el proceso de optimización de la población se ha sustituido el algoritmo genético compacto (cGA) utilizado en (Llorà et al., 2005) por otro algoritmo de la familia de los algoritmos PBIL adaptado para trabajar con variables de valores reales y continuas —los pesos de la función de coste—, denominado PBIL continuo (cPBIL) (Sebag y Ducoulombier, 1998). Asimismo, el regresor ε-SVM también tiene que ajustarse al problema. A continuación se describen las características principales de estos algoritmos y las adaptaciones consideradas.

Adaptación de ε-SVM a valores continuos

Como se ha comentado, a partir de la información extraída del criterio del usuario utilizando el *ranking* de las soluciones obtenida del grafo de ordenación parcial, se procede a generalizar este patrón para obtener una función de *fitness* sintético para todos los individuos de la población mediante un regresor ε-SVM, capaz de evaluar las soluciones acorde con la respuesta del usuario hasta ese momento. Este regresor debe ajustarse para funcionar adecuadamente sobre un conjunto de valores reales. Para ello, se ha utilizado la librería SVM de la National Taiwan University[15] para entrenar el regresor a partir del orden parcial de un primer conjunto de individuos ordenados, minimizando el error de regresión —ver (Formiga, 2005) para una descripción detallada del proceso.

Otra de las cuestiones, es el número de torneos que deben ser evaluados por el usuario. En las pruebas efectuadas por (Llorà et al., 2005), se demuestra que para entrenar al regresor ε-SVM con un *kernel* lineal sobre un espacio de dimensión ℓ (en su caso $\ell = 6$, ya que se trabaja con individuos binarios de seis posiciones), es necesario disponer de $\ell + 2$ ejemplos —en este caso, los individuos junto a sus *rankings* (ver tabla 2.1). No obstante, en el presente trabajo de investigación, dada la complejidad del problema de optimización, se ha aumentado el número de individuos evaluados por el usuario al doble (valor fijado empíricamente, ver (Formiga, 2005)) para disponer de suficiente información para obtener una estimación fiable del *fitness* subjetivo, ya que no existe proporcionalidad entre las variaciones de los individuos

[15] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/



v	$f(v)$	$r(v)$	$\delta(v)$	$\phi(v)$	$\hat{f}(v)$	$\hat{r}(v)$
010111	4	1	5	0	5	1
010100	2	3	0	1	-1	4
010101	3	2	1	1	0	3
100001	2	3	0	2	-2	5
100000	1	4	0	3	-3	6
101010	3	2	2	0	2	2
001000	1	4	0	3	-3	6
001110	3	2	2	0	2	2

Figura 2.17: Diagrama del ajuste subjetivo de los pesos de la función de coste basado en un algoritmo genético interactivo activo.

—genotipos— y sus representaciones finales —fenotipos—, o lo que es lo mismo, los pesos y sus correspondientes frases sintéticas.

Population Based Incremental Learning en espacios continuos (cPBIL)

Siguiendo el trabajo de (Sebag y Ducoulombier, 1998), el algoritmo cPBIL utilizado en este trabajo se estructura en los siguientes pasos (ver figura 2.18), considerando una población de n individuos:

1. *Inicialización:* Como en un algoritmo genético simple, en el que la población de individuos se inicializa, típicamente, mediante valores aleatorios, en un cPBIL se modela el valor de cada gen del individuo mediante una distribución normal $\mathcal{N}(X, \sigma)$, cuya $X_i = 0.5$ (Sebag y Ducoulombier, 1998) y $\sigma_i = 0.1$ (valor fijado experimentalmente), para $1 \leq i \leq \ell$ genes de cada individuo— en este caso el número de pesos a ajustar ($\ell = p + q$, según las ecuaciones (2.1) y (2.2)).

2. *Muestreo del modelo:* Cada nuevo individuo se genera a partir del conjunto de valores aleatorios obtenidos del muestreo de la distribución normal que representa a cada uno de los genes (pesos). Se generarán tantos individuos aleatorios como el tamaño de la población n indique.

3. *Evaluación:* Se calcula el *fitness* o medida de calidad de los individuos —obtenida, en este caso, a partir del resultado de la aplicación del regresor ε-SVM— y se ordena la población según el resultado obtenido (cuanto mayor sea el valor de *fitness* del individuo, mejor estará adaptado al criterio del usuario).

4. *Selección:* Como en los algoritmos genéticos tradicionales, cPBIL dispone de un esquema de selección encargado de elegir los individuos que actualizarán las distribuciones normales que representan a los genes de los individuos. En este caso, siguiendo lo indicado por (Sebag y Ducoulombier, 1998), se seleccionan los K mejores individuos (ejemplos positivos) junto al peor (ejemplo negativo) de la población para actualizar el modelo probabilístico.

5. *Actualización del modelo probabilístico:* por un lado, se actualiza la media de las distribuciones de los genes de los individuos, tomando en consideración los dos mejores individuos junto al peor de la población, según la ecuación (2.19) (Sebag y Ducoulombier, 1998).

$$X_i^{t+1} = (1 - \alpha) \cdot X_i^t + \alpha \cdot (X^{best_1} + X^{best_2} - X^{worst}), \tag{2.19}$$

   donde t indica la iteración acutal y α ($0 \leq \alpha \leq 1$) representa el factor de aprendizaje (*learning rate*, en inglés) que determina la velocidad de la convergencia del vector de probabilidades.

   A continuación, se actualiza la desviación de las distribuciones de los genes a partir de la varianza de los K mejores individuos, según la ecuación (2.20) —ver (Sebag y Ducoulombier, 1998) para una justificación de la estrategia utilizada.

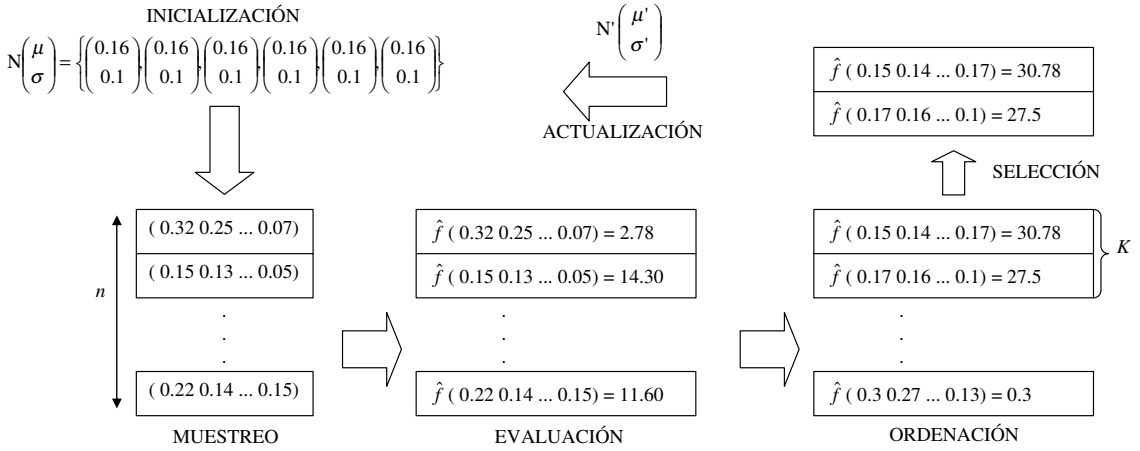


Figura 2.18: Diagrama del funcionamiento del algoritmo genético cPBIL basado en la representación de la población mediante distribuciones probabilísticas $\mathcal{N}(\mu, \sigma)$, en este caso con 6 genes por individuo.

$$\sigma_i^{t+1} = (1-\alpha)\sigma_i^t + \alpha\sqrt{\frac{\sum_{j=1}^{K}(X_i^j - \hat{X}_i^K)}{K}} \tag{2.20}$$

donde i indica el gen, j el individuo y $\hat{X}_i^K$ representa la media del gen i de los K mejores hijos $\{X^1, \ldots, X^K\}$.

6. Repetir los pasos 2 a 5 hasta que se satisfaga el criterio de convergencia.

En este trabajo se utiliza $K = n/5$, siguiendo las indicaciones de (Sebag y Ducoulombier, 1998), $\alpha = 1/n$ según (Llorà et al., 2005) y se fija experimentalmente el criterio de parada en $\sigma = 0.01$, debido a que se trabaja con pesos con resolución de dos decimales (ver sección 2.5).

2.3.3. Consistencia de las evaluaciones del usuario

Uno de los problemas detectados durante la realización de las pruebas utilizando AGI era no disponer de criterios de control de la robustez del usuario para validar la calidad de las configuraciones de pesos obtenidas después del proceso evolutivo guiado por el mismo. Típicamente, la consistencia del usuario se suele controlar mediante la inclusión de puntos de control a lo largo del ciclo evolutivo (Takagi, 2001), filosofía también utilizada en el ámbito de la evaluación subjetiva de la calidad de los sistemas de conversión de texto en habla (Vepa, King y Taylor, 2002; Breuer y Abresch, 2004). No obstante, incorporar más evaluaciones (puntos de control), contribuye a aumentar la fatiga del usuario al obligarlo a realizar la misma prueba varias veces (para evaluar su robustez de criterio).

Los aAGI permiten abordar esta problemática desde otro punto de vista. Gracias a trabajar con un grafo dirigido de soluciones ordenadas $\mathcal{G}'$, se puede determinar el grado de consistencia del usuario a la hora de evaluar las soluciones propuestas. Concretamente, si un vértice v aparece más de una vez en el cálculo de $\delta(v)$ o $\phi(v)$, entonces, existe un *ciclo* en el grafo, es decir, el usuario ha sido inconsistente en la comparativa (el ciclo representa una inconsistencia en las evaluaciones del usuario). Por lo tanto, gracias a haber relacionado las soluciones mediante conexiones *mayor que*, las contradicciones del usuario durante las evaluaciones pueden identificarse fácilmente gracias a la aparición de ciclos en el grafo normalizado $\mathcal{G}'$. Esta propiedad es la base de la medida de consistencia introducida más adelante en la ecuación (2.21). Según este enfoque, un usuario será consistente en el instante t de las evaluaciones si no aparecen ciclos en el grafo normalizado de ordenación parcial ($\mathcal{G}'^t$). Para poder determinar el grado de consistencia del usuario es necesario disponer de dos elementos:

- Un proceso automático de detección de ciclos sobre el grafo normalizado $\mathcal{G}'^t$ (ordenación de las soluciones en el instante t mediante relaciones *mayor que*),
- Una medida que permita cuantificar el grado de inconsistencia que provoca el ciclo sobre el conjunto de evaluaciones.

A continuación se presentan las soluciones implementadas para abordar estos dos elementos.

Detección automática de ciclos

Como se ha comentado, la detección de los ciclos del grafo normalizado es el proceso previo necesario para determinar la consistencia del usuario a lo largo del proceso interactivo de evaluación. La idea fundamental de estos algoritmos se basa en disponer de un criterio sólido para identificar los ciclos, y así, evitar la redundancia en la detección de ciclos (volver a detectar un ciclo o subciclo dentro de otro ciclo de soluciones) o el bloqueo del proceso (bucle infinito). Por ello, resulta fundamental disponer de unos algoritmos que controlen todas estas situaciones y den una respuesta fiable a la exploración y detección de ciclos.

En este caso, se trata de un proceso recursivo que analiza, para todos los vértices v del grafo $\mathcal{G}' =< \mathcal{V}', \mathcal{E}' >$, los caminos que parten de cada uno de ellos hacia otros vértices $\mathcal{V}_\mathcal{N} \subset \mathcal{V}'$ de forma recursiva hasta que, o bien, finalice el camino, o bien se repita un vértice en el mismo (existe un ciclo). Este proceso se lleva a cabo mediante los algoritmos 3 y 4 encargados de explorar y detectar los ciclos del grafo de soluciones. Una vez se han detectado los caminos cíclicos —el vértice que se repite no tiene porqué ser el primero del camino—, se eliminan las partes no cíclicas del camino, reteniendo sólo los nodos que corresponden al ciclo.

En este caso, para cada vértice v^i de $\mathcal{G}' =< \mathcal{V}', \mathcal{E}' >$, el proceso de detección de ciclos explora sus relaciones con otros vértices $\mathcal{V}_\mathcal{N}$ que no han sido previamente procesados —controlado mediante el conjunto acumulado de vértices visitados en el camino que parte de v ($\mathcal{V}_\mathcal{T}$ en el algoritmo 3) (puntos 2 y 3 del algoritmo). Una vez se han detectado los caminos

cíclicos mediante el algoritmo 4 (punto `4`), las partes no cíclicas de los caminos —soluciones correctamente ordenadas— son filtradas para evitar la ambigüedad de subciclos (punto `5`). Uno de los objetivos de este proceso pasa por tener en cuenta los subciclos que puedan aparecer dentro de un macro-ciclo de orden superior, considerándolos como ciclos diferentes. Por ejemplo, en la figura 2.21, se obtendrán dos subciclos independientes, $\{2 \rightarrow 1 \rightarrow 5 \rightarrow 2\}$ y $\{1 \rightarrow 5 \rightarrow 9 \rightarrow 11 \rightarrow 3 \rightarrow 1\}$, evitando entrar en un proceso de recursividad infinita al estar todos los nodos interconectados entre sí. Seguidamente, el algoritmo ordena los vértices de cada ciclo por antigüedad (índice del vértice), de tal forma que el primer y último vértice del ciclo sea el vértice más antiguo (punto `6` del algoritmo 3) Gracias a este criterio, se obtiene una ordenación consistente del contenido de los ciclos analizados. Por ejemplo, en la figura 2.20, dadas las secuencias de flechas entre vértices $\{2 \rightarrow 1 \rightarrow 5 \rightarrow 2\}$, $\{1 \rightarrow 5 \rightarrow 2 \rightarrow 1\}$ o $\{5 \rightarrow 2 \rightarrow 1 \rightarrow 5\}$, el ciclo quedaría representado por {1,5,2}, una vez ordenadas por antigüedad, ya que ambas representan un mismo ciclo de soluciones. Finalmente, todos los vértices que aparecen como mínimo en un ciclo forman el conjunto $\chi(\mathcal{G}')$, utilizado para definir la medida de consistencia que se presenta a continuación.

***Eliminación* de ciclos:** Una vez definido el conjunto de vértices que forman parte de un ciclo, resulta necesario definir algún proceso encargado de obtener un orden robusto de las soluciones evitando la aparición de inconsistencias en la ordenación. El objetivo es disponer de un *ranking* ordenado de las soluciones para entrenar al regresor lineal utilizado (sin romper físicamente los ciclos). Para ello se definen dos heurísticas. En la primera se etiqueta cada flecha (relación de preferencia entre soluciones) con el número de veces que esta relación ha sido ratificada por el usuario, rompiendo el ciclo por la conexión de menor puntuación. El hecho de emplear un torneo jerárquico a lo largo de las generaciones, permite que un mismo torneo (pareja de soluciones) sea evaluado varias veces. Se puede determinar una puntuación de solidez a la flecha como el número de veces que el usuario ha votado en ese sentido la comparativa de soluciones presentada. Por ejemplo, dado el grafo cíclico $\{3 \rightarrow 4 \rightarrow 5 \rightarrow 3\}$, con las flechas etiquetadas según las votaciones $\{3 \rightarrow 4\} : 2$, $\{4 \rightarrow 5\} : 1$ y $\{5 \rightarrow 3\} : 2$, el ciclo se rompería por la conexión $\{4 \rightarrow 5\}$, que es, en este caso, la de menor puntuación (es menos fiable). En el caso que todas las conexiones presenten la misma puntuación (p.ej. las parejas de soluciones conectadas sólo han sido evaluadas por el usuario una vez), es necesario contemplar una segunda heurística. Para ello, se considera la dominancia de cada vértice, uan vez eliminado el ciclo, suprimiendo la flecha que va desde un vértice de menor dominancia a uno de mayor dominancia (se rompe el ciclo por el punto más débil en términos de dominancia). Finalmente, y como último criterio utilizado, si ambas heurísticas no permiten romper el ciclo, se elimina el camino más antiguo de entre los que conforman el ciclo[16]. Una vez eliminada la flecha según los criterios descritos, se procede a determinar la dominancia del vértice del mismo modo que se ha descrito anteriormente (mediante las heurísticas $\delta(v)$ y $\phi(v)$).

[16]Se toma en consideración que el reducido conocimiento del usuario respecto al problema en las primeras iteraciones puede provocar la inconsistencia de la comparativa. No obstante, también existen argumentos para romper el ciclo por el camino más nuevo, si se considera que en este instante el usuario puede empezar a estar fatigado. Esta cuestión queda abierta para futuros trabajos de investigación.

No obstante, cabe puntualizar que los ciclos del grafo sólo se eliminan para poder obtener una ordenación consistente de los nodos (soluciones) y, así, poder generar el modelo de *fitness* sintético que guíe la generación de nuevas soluciones mediante el algoritmo cPBIL, por lo que los ciclos continúan representados en el grafo hasta que el usuario rompe el ciclo al repetir alguna de las comparaciones a lo largo del proceso evolutivo (se toma en consideración como mejor la última evaluación que rompe el ciclo). Gracias a trabajar con aAGI, se va consolidando de forma iterativa el modelo construido (función aproximante) a partir de las evaluaciones del usuario. De este modo, se consigue que los resultados obtenidos (vía la comparación de soluciones) sean validados varias veces debido al torneo jerárquico binario utilizado.

Algoritmo 3 Algoritmo de detección de ciclos en $\mathcal{G}'$.

proceso *cycleDetection*($\mathcal{G}'$)

1: Crear el conjunto vacío de $\mathcal{C}$ ciclos, $\mathcal{V}_{\mathcal{I}}$ de vértices visitados
2: Extraer el primer vértice $v^i \in \mathcal{V} \mid v^i \notin \mathcal{V}_{\mathcal{I}}$
3: Crear el conjunto $\mathcal{V}_{\mathcal{N}} = \{\forall\ v \in \mathcal{V}_{\mathcal{N}} : (v \neq v^i) \cap (v \in \mathcal{G}')\}$
4: Crear el conjunto de ciclos $\mathcal{C}_{\mathcal{I}} = \{\forall v \in \mathcal{V}_{\mathcal{N}}\ \forall e(v^i, v) \in \mathcal{E}' : cycleExporer(\{v^i\}, v, \mathcal{G}') \subseteq \mathcal{C}_{\mathcal{I}}\}$
5: Filtrar las partes no cíclicas y eliminar ambigüedades cíclicas de los caminos $\forall c \in \mathcal{C}_{\mathcal{I}}$
6: Ordenar los ciclos considerando el vértice más antiguo como el primer y último vértice de $\forall c \in \mathcal{C}_{\mathcal{I}}$
7: $\mathcal{V}_{\mathcal{I}} \leftarrow \mathcal{V}_{\mathcal{I}} \cup \{v^i\}$
8: $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}_{\mathcal{I}}$
9: Volver al punto 2 mientras $\forall v^i \in \mathcal{G}'$: $v^i \notin \mathcal{V}_{\mathcal{I}}$. Sino $cycleDetection \leftarrow \mathcal{C}$

Algoritmo 4 Exploración de todos los caminos que parten del vértice v en $\mathcal{V}_{\mathcal{I}}$, donde $e(\cdot,\cdot)$ representa la conexión (flecha) entre dos vértices.

proceso *cycleExplorer*($\mathcal{V}_{\mathcal{I}}, v, \mathcal{G}' =< \mathcal{V}', \mathcal{E}' >$)

1: $\mathcal{V}_{\mathcal{I}} \Leftarrow \mathcal{V}_{\mathcal{I}} \cup v$
2: Crear el conjunto $\mathcal{R} = \{\forall v^i \in \mathcal{V}_{\mathcal{I}}\ \forall e(v, v^i) \in \mathcal{E}' : e(v, v^i) \subseteq \mathcal{R}\}$
3: $(\mathcal{R} \neq \emptyset) \Rightarrow return(\mathcal{R})$
4: Crear el conjunto $\mathcal{C}_{\mathcal{I}} = \{\forall v^i \in (\mathcal{V} - \{v\}) \forall e(v, v^i) \in \mathcal{E}' : cycleExporer(\mathcal{V}_{\mathcal{I}}, v^i, \mathcal{G}') \subseteq \mathcal{C}_{\mathcal{I}}\}$
5: $return(\mathcal{C}_{\mathcal{I}})$

Medida de consistencia

Una vez detectados los ciclos del grafo, se procede a calcular el grado de consistencia (robustez) de las evaluaciones de los usuarios. Para ello, en (Alías et al., 2006) se introduce la medida de consistencia de un usuario en un instante temporal t, denominada $\kappa\left(\mathcal{G}'^t, \omega\right)$, definida según la ecuación (2.21).

$$\kappa\left(\mathcal{G}'^t,\omega\right)=1-\left(\frac{1}{\left|\mathcal{V}'^t\right|}\cdot\sum_{v\in\chi(\mathcal{G}'^t)}\omega_v\right)^{\alpha} \tag{2.21}$$

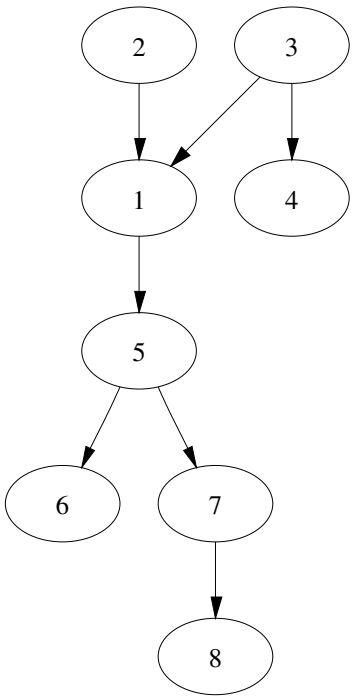
donde $\left|\mathcal{V}'^t\right|$ es el número de vértices del grafo normalizado $\mathcal{G}'$ en el instante t ($\mathcal{G}'^t$), ω_v representa el peso (importancia) del vértice v en la medida de consistencia (no confundir con los pesos de la función de coste), $\chi\left(\mathcal{G}'^t\right)$ denota el conjunto de vértices que pertenecen a los ciclos detectados en el grafo $\mathcal{G}'^t$, y α el factor de escalado global, siendo $\alpha \geq 1$. En el presente trabajo, y como paso inicial, en todos los cálculos de la medida de consistencia presentados se trabaja con $\omega_v = 1, \forall v \in \mathcal{V}'^t$ y $\alpha = 1$, por lo que κ denota la proporción de vértices que están en un ciclo respecto al total de vértices del grafo normalizado en ese instante del ciclo evolutivo. En un futuro se estudiarán otro tipo de ponderaciones, por ejemplo considerando la antigüedad del vértice v, como información para el cálculo de ω_v.

Esta medida permite conocer la consistencia del usuario en el instante t. Por lo tanto, permite visualizar la robustez de las evaluaciones del usuario a lo largo del proceso evolutivo, evitando tener que incorporar puntos de control explícitos —es decir, comparativas A-B *vs.* B-A, típicamente utilizados para validar la consistencia del criterio del usuario (por ejemplo ver (Vepa, King y Taylor, 2002; Breuer y Abresch, 2004)). Por lo tanto, es un método implícito para reducir el tiempo necesario para ajustar los pesos, ayudando de este modo a reducir la fatiga del usuario. En este trabajo de investigación, por un lado, el grado de consistencia del usuario se calcula para $t = t_f$ (tiempo final), es decir, se determina la consistencia de la solución a la que ha llegado el usuario una vez finalizado el proceso iterativo de entrenamiento de los pesos. Asimismo, en los experimentos también se analiza la evolución de la consistencia de los usuarios a lo largo del proceso evolutivo[17]. En las figuras 2.19, 2.20 y 2.21 se presentan varios ejemplos del cálculo de la medida κ para tres iteraciones de un proceso de interacción con el usuario. En ellas se representan los grafos de ordenación parcial, los valores del *fitness* subjetivo y el *ranking* calculado, junto al valor de la consistencia de usuario en el instante t, $\kappa\left(\mathcal{G}'^t,\omega\right)$. En estos ejemplos se puede observar el impacto de la inconsistencia del usuario en la toma de decisiones sobre la medida presentada. Por ejemplo, en la figura 2.20 se observa como la flecha $\{5 \rightarrow 2\}$ es la que ocasiona la aparición del ciclo —es decir, el usuario ha preferido la solución 5 ante la 2, después de haber preferido la 2 a la 1 y la 1 a la 5—, rompiendo el orden establecido en el grafo. Situación que se repite en la figura 2.21, al conectarse $\{3 \rightarrow 1\}$, después de varias nuevas comparativas.

2.4. Experimentos

El corpus de voz utilizado para los experimentos está formado por 1.520 frases en catalán (20 minutos de voz), de las cuales 1.207 corresponden a las palabras portadoras de las

[17] En un futuro se pretende incluir información de la consistencia del usuario en tiempo de ejecución del proceso evolutivo, por ejemplo a través de información visual de la interfaz interactiva —ver (Llorà et al., 2006) para un primer paso en esta dirección.



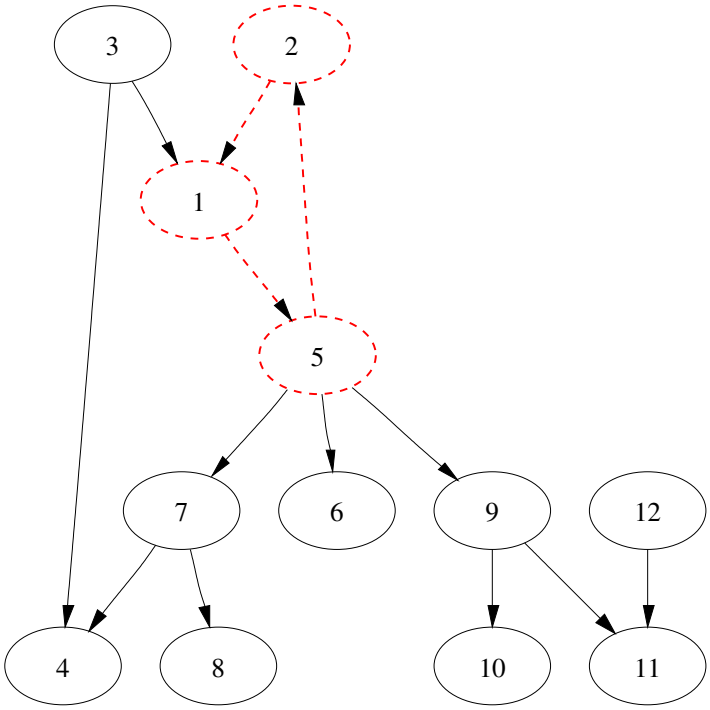
(a) Grafo de ordenación parcial.

v	$\delta(v)$	$\phi(v)$	$\hat{f}(v)$	$\hat{r}(v)$
1	4	2	2	3
2	5	0	5	2
3	6	0	6	1
4	0	1	-1	5
5	3	3	0	4
6	0	4	-4	7
7	1	4	-3	6
8	0	5	-5	8

$$\left|\mathcal{V}'^1\right| = 8,\ \left|\chi\left(\mathcal{G}'^1\right)\right| = 0,\ \kappa\left(\mathcal{G}'^1\right) = 1$$

(b) *Fitness* sintético y medida de consistencia κ.

Figura 2.19: Ejemplo de *fitness* subjetivo y medida de consistencia para aAGI en la iteración $t = 1$.



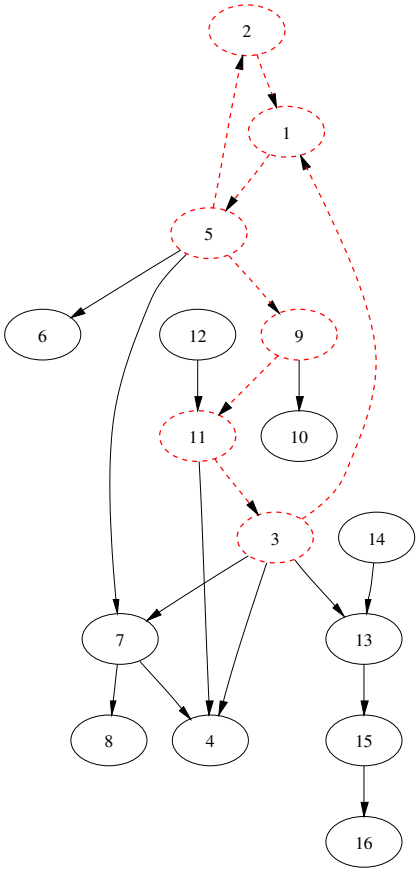
(a) Grafo de ordenación parcial.

v	$\delta(v)$	$\phi(v)$	$\hat{f}(v)$	$\hat{r}(v)$
1	9	3	6	2
2	9	3	6	2
3	10	0	10	1
4	0	5	-5	6
5	9	3	6	2
6	0	4	-4	5
7	2	4	-2	4
8	0	5	-5	6
9	2	4	-2	4
10	0	5	-5	6
11	0	6	-6	7
12	1	0	1	3

$$\left|\mathcal{V}'^2\right| = 12,\ \left|\chi\left(\mathcal{G}'^2\right)\right| = 3,\ \kappa\left(\mathcal{G}'^2\right) = 0.75$$

(b) *Fitness* sintético y medida de consistencia κ.

Figura 2.20: Ejemplo de *fitness* subjetivo y medida de consistencia para aAGI en la iteración $t = 2$.



(a) Grafo de ordenación parcial.

v	$\delta(v)$	$\phi(v)$	$\hat{f}(v)$	$\hat{r}(v)$
1	13	6	7	2
2	13	6	7	2
3	13	6	7	2
4	0	8	-8	7
5	13	6	7	2
6	0	7	-7	6
7	2	7	-5	4
8	0	8	-8	7
9	13	6	7	2
10	0	7	-7	6
11	13	6	7	2
12	14	0	14	1
13	2	8	-6	5
14	3	0	3	3
15	1	9	-8	7
16	0	10	-10	8

$\left|\mathcal{V}'^3\right| = 16,\ \left|\chi\left(\mathcal{G}'^3\right)\right| = 6,\ \kappa\left(\mathcal{G}'^3\right) = 0.625$

(b) *Fitness* sintético y medida de consistencia κ.

Figura 2.21: Ejemplo de *fitness* subjetivo y medida de consistencia para aAGI en la iteración $t = 3$.

Tabla 2.2: Análisis de la distribución de los fonemas del corpus utilizado (en %) respecto a (Esquerra, Febrer y Nadeu, 1998). Los valores significativamente diferentes están en cursiva. Los fonemas se representan según la notación SAMPA.

Unidad	**Corpus**	**Rafel**	**Diferencia**	**Esquerra**	**Diferencia**
@	15.62	20.10	*-4.48*	18.91	*-3.29*
a	4.36	4.85	-0.49	4.6	-0.24
b+B	3.18	2.96	0.22	2.72	0.46
d+D	4.45	4.48	-0.03	4.32	0.13
e	2.79	2.34	0.45	2.96	-0.17
E	2.12	1.63	0.49	1.11	1.01
f	1.26	1.18	0.08	0.99	0.27
g+G	1.76	0.95	0.81	1.07	0.69
i+j	8.00	6.16	*1.84*	7.79	0.21
J	0.77	0.28	0.49	0.25	0.52
k	3.97	4.41	-0.44	4.46	-0.49
l	5.83	6.00	-0.17	5.68	0.15
L	0.81	0.85	-0.04	0.51	0.30
m	3.42	3.83	-0.41	3.58	-0.16
n	5.41	6.38	-0.97	6.13	-0.72
N	0.81	0.36	0.45	0.5	0.31
o	2.68	1.77	0.91	2.12	0.56
O	1.58	1.43	0.15	1.1	0.48
p	2.66	2.76	-0.10	3.03	-0.37
r	3.42	3.77	-0.35	3.68	-0.26
rr	3.39	2.32	*1.07*	2.25	*1.14*
s	6.06	8.57	*-2.51*	6.63	-0.57
S	0.65	0.45	0.20	0.29	0.36
t	4.71	5.24	-0.53	5.36	-0.65
u+w	6.18	5.63	0.55	5.88	0.30
z	1.43	0.76	0.67	3.14	*-1.71*
Z	0.53	0.54	-0.01	0.95	-0.42

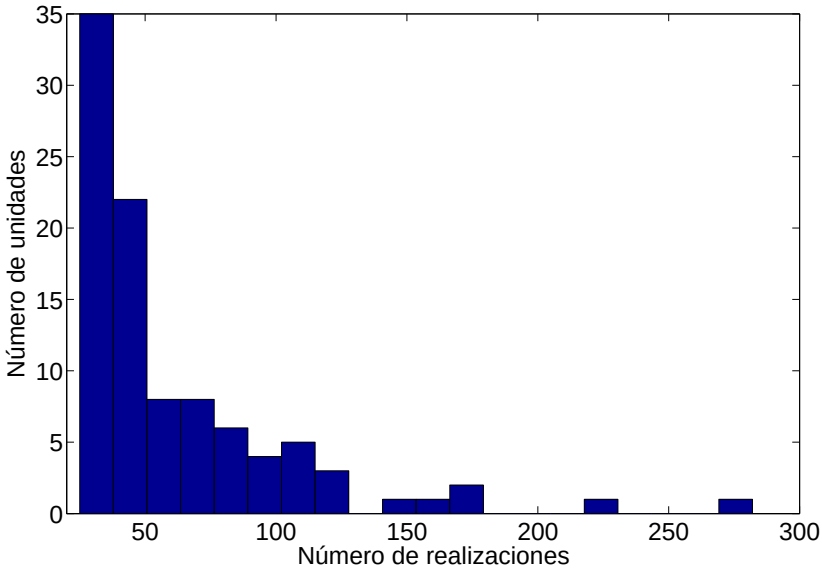


Figura 2.22: Histograma de la distribución de las unidades con más de 25 realizaciones del corpus de voz utilizado en los experimentos.

unidades básicas (difonemas y trifonemas) utilizadas para síntesis concatenativa basada en difonemas. Este corpus, grabado por un locutor profesional masculino, está fonéticamente balanceado (ver tabla 2.2) y consta de 9.863 realizaciones de las unidades básicas (entre difonemas y trifonemas). En la tabla 2.2 se presenta el análisis de la distribución de los fonemas en el corpus respecto a dos estudios previos (Esquerra, Febrer y Nadeu, 1998). Se puede observar, cómo el corpus tiene una cierta tendencia a tener menos vocales neutras (/@/) y algo más de fonemas vibrantes (/rr/) que los estudios referenciados. El resto de diferencias no son significativas, por lo que, aunque se trate de un corpus de dimensiones reducidas, está bien balanceado —con una correlación $\rho = 0.9747$ respecto al estudio de Rafel y $\rho = 0.9834$ respecto a Esquerra, en la tabla 2.2. En este corpus, el 53 % de las unidades presentan menos de 2 realizaciones, el 83 % menos de 10, y sólo un 1 % de las unidades tienen más de 25 instancias diferentes en el corpus (con 14 unidades que presentan más de 100). En los experimentos, el entrenamiento de los pesos sólo considerará el conjunto de unidades con un número mayor o igual a 25 realizaciones (97 exactamente) (ver figura 2.22). De este modo, se fija el mínimo de realizaciones necesarias para disponer de suficiente variabilidad de unidades candidatas. En un futuro se pretende agrupar las unidades según criterios fonéticos para que todas ellas dispongan de un ajuste de pesos, evitando el problema de la pobre representación de los datos (ver anexo A para una primera propuesta).

2.4.1. Experimentación y resultados preliminares

El primer experimento se centra en la evaluación del funcionamiento del algoritmo genético trabajando sobre una función de coste objetiva (como los métodos clásicos de ajuste automático de los pesos, como MLR o WSS). En este experimento se trabaja con los seis subcostes con normalización *Max-min* (ver ecuaciones (2.9) y (2.10)) —valores obtenidos para cada unidad para la que se calculan los pesos. Así, el proceso automático dispone de unos subcostes normalizados entre $[0, 1]$, evitando los posibles sesgos en el proceso de convergencia debido a diferencias en el rango de los subcostes.

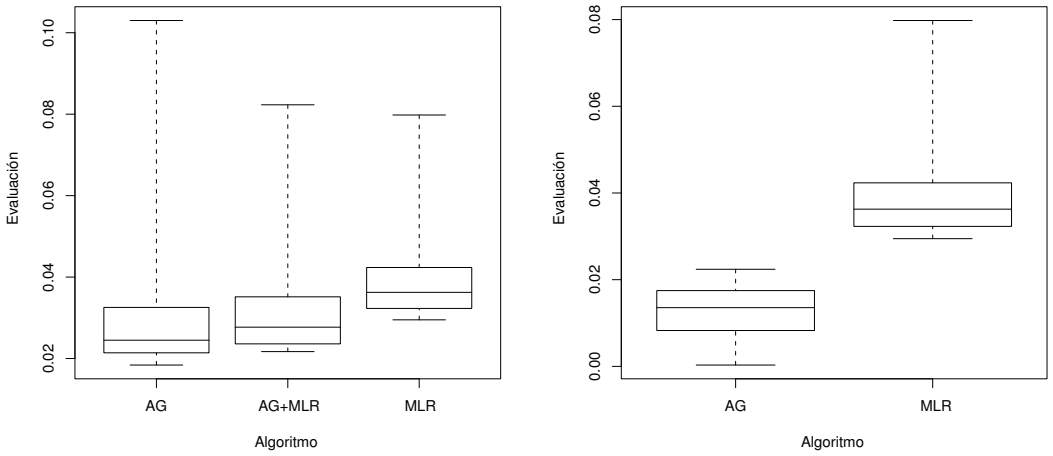
En cambio, en el segundo experimento relacionado con el AGI, se estudia el ajuste de los pesos a nivel de todo el corpus, utilizando los subcostes con normalización *mean-std* (ver ecuaciones (2.7) y (2.8)). De esta manera se simplifica su cálculo y, al mismo tiempo, se evalúa el funcionamiento en un contexto diferente: la evaluación subjetiva.

Experimento 1 - Mejoras obtenidas por el algoritmo genético sobre el ajuste objetivo de los pesos

El primer experimento realizado se centró en la evaluación del funcionamiento del algoritmo genético basado en una distancia objetiva (distancia Euclídea cepstral) respecto a los métodos clásicos de ajuste de pesos de la función de coste de selección. Este experimento se divide en dos partes. Primero, se entrenan los pesos a nivel de unidad, como primer paso para validar la viabilidad de la técnica propuesta. De las 1207 unidades del corpus, sólo se han considerado para el experimento aquellas que contenían más de 25 realizaciones (1 % del total), con el fin de disponer de suficiente información estadística que permita obtener unos resultados lo bastante robustos.

En cuanto a la configuración de los parámetros del AG, los valores escogidos son los siguientes: $n = 200$ individuos, 100 iteraciones (antes de detener el proceso evolutivo), $p_c = 0.3$ y $p_m = 0.003$. Estos parámetros, que se encuentran dentro del rango de valores habituales en la literatura específica (Goldberg, 1989; Goldberg, 2002), se han fijado de forma experimental a partir de un conjunto de pruebas inciales sobre distintas unidades utilizadas como referencia (p.ej la unidad /b@/). Después de estas pruebas preliminares, el AG ha sido adaptado a les características y necesidades planteadas por el problema de optimización sobre el que se aplica.

En cuanto a los métodos de ajuste de pesos considerados como referencia en esta primera fase del experimento, se decidió escoger MLR ante WSS por sus características, principalmente, su mayor robustez, el hecho de no discretizar el espacio de búsqueda, y su menor coste computacional (ver sección 2.1.5). Por otra parte, y durante el conjunto de pruebas, se introduce una variante del AG a partir de una inicialización heurística. En este caso, se incorpora a la población inicial la información obtenida del ajuste de pesos con MLR. De esta manera, se define un sistema híbrido, designado como AG+MLR, que prueba de aprovechar el conocimiento del medio (espacio de optimización) obtenido con el método lineal (MLR), con el fin de estudiar la posible mejora introducida respecto al sistema no lineal (AG). Se realizaron distintas pruebas, sesgando más o menos la población inicial con



(a) Unidad /b@/ (b) Todas las unidades

Figura 2.23: Distribución del valor de la función de coste (*fitness* para AG) de los pesos obtenidos por los algoritmos estudiados para de todas las unidades analizadas.

el resultado (configuración de pesos) obtenido con el método MLR (desde el 10 % hasta el 50 % de la población son copias de la solución del MLR). No obstante, los resultados obtenidos, como se pasa a describir a continuación, no fueron satisfactorios (ver figura 2.23(a)), por lo que se descartó el método híbrido para el resto de pruebas.

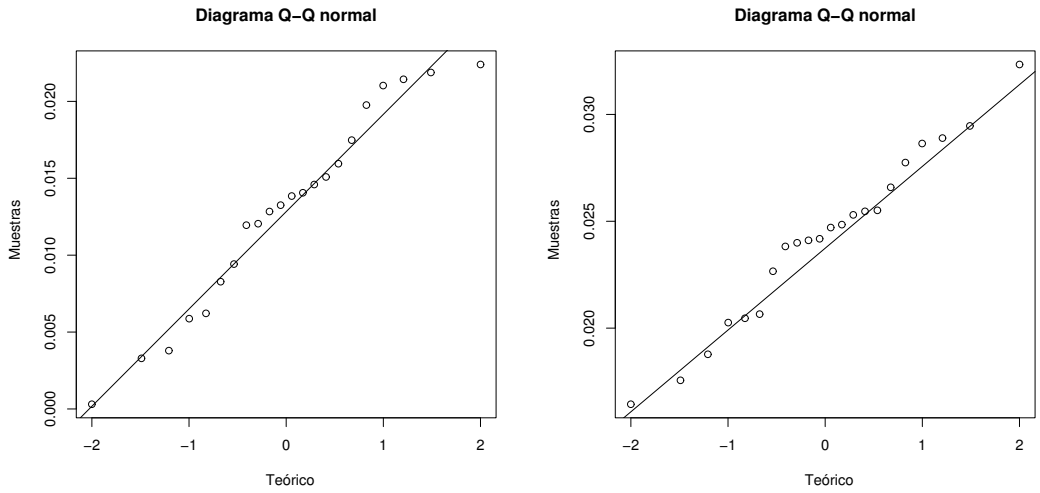
Como se ha comentado, en la primera fase del test, se escogieron varias unidades al azar (del conjunto de unidades estadísticamente significativas) para llevar a cabo los ajustes de los parámetros del algoritmo genético y validar el funcionamiento del AG en el contexto del entrenamiento de pesos. Como ejemplo, se presentan los resultados obtenidos sobre la unidad /b@/ (con 25 realizaciones) con los tres métodos estudiados: MLR, AG+MLR y AG. La figura 2.23(a) presenta las estadísticas obtenidas para todas las realizaciones de la unidad analizada (tratadas como unidad objetivo), a partir del análisis de la mejor configuración de los pesos obtenida para cada uno de los métodos aplicados. La solución obtenida con el AG presenta una mejor adaptación media al problema —en este caso, la mejor solución es la que aporta una función de coste menor— comparándola con la del MLR, aunque con una desviación estándar más elevada. En cuanto al método AG+MLR, las soluciones obtenidas sólo consiguen reducir la desviación de los resultados respecto a MLR, sin mejorar la adaptación de las soluciones al problema respecto al AG. Por lo tanto, se deduce que la heurística introducida en el problema a partir de los resultados obtenidos por el MLR no mejora las soluciones obtenidas por el AG inicializdo aleatoriamente. Después de analizar los resultados, se concluye que esto radica en que el método MLR pretende encontrar una solución que aproxime linealmente todas las posibles soluciones, mientras que el AG busca una solución *local* al problema (un mínimo). Por lo tanto, al tratarse de filosofías diferentes,

no se mejoran las prestaciones del AG al incorporar el resultado del MLR sobre los mismos datos, cuestión que provoca que el método híbrido quede descartado para el resto de pruebas del experimento.

Por otra parte, esta primera fase del experimento permite observar el efecto de la distribución de los subcostes (debido a las características del corpus de voz y el resultado de la normalización de coste utilizada), guiando las soluciones presentadas por el AG hacia soluciones con un patrón mucho más sesgado que con MLR (maximización local *vs.* global). Por ejemplo, el coste de unidad que evalúa las diferencias entre las duraciones de las unidades (`DUR T`) toma a menudo la máxima ponderación en la configuración de pesos (ver las celdas de la diagonal de los histogramas de la figura 2.25(b)). De estas pruebas iniciales, también se puede deducir que en diferentes ejecuciones del AG sobre la misma unidad se obtienen pesos con configuraciones distintas. Eso es debido al ruido introducido por el proceso de muestreo utilizado por el AG al seleccionar aleatoriamente la realización que actúa de unidad objetivo y el hecho de que la función a optimizar es bastante compleja. Por este motivo, el entorno de optimización se convierte en altamente multimodal. En este contexto, el algoritmo genético presenta un comportamiento mucho más sólido, respecto a los métodos clásicos de optimización, gracias a su robustez ante problemas ruidosos (ver el apartado correspondiente de la sección 2.2.2).

A continuación, el mismo test se lleva a cabo sobre todas las unidades que cumplen las especificaciones del experimento (con suficiente relevancia estadística), obteniendo un vector de pesos para cada una de ellas. Los resultados muestran que las soluciones (vectores de pesos) obtenidas mediante el AG mejoran a las del MLR en términos del valor medio y de la desviación de la función de coste resultante (*Evaluación* en la figura 2.23(b)). Por lo tanto, los pesos aportados por el AG están mejor *adaptados* al problema (similitud entre las unidades en términos de la función de coste o *fitness*) a solucionar.

La segunda parte del experimento se centra en la evaluación de la bondad estadística de los resultados obtenidos en las pruebas realizadas. Por este motivo, se analiza la distribución de los resultados (vector de pesos) en cuanto al valor de la función de coste ($C(t_1^n, u_1^n)$, ver ecuaciones (2.3) y (2.17)) para los dos algoritmos (MLR y AG) a lo largo de las unidades analizadas. En la figura 2.24 se observa un ejemplo de la comparación entre las muestras de la distribución de valores de $C(t_1^n, u_1^n)$ obtenida respecto de las muestras de una distribución normal teórica. Los puntos de la figura corresponden a los puntos teóricos respecto a los muestreados, mientras que la línea continua representa la distribución normal teórica respecto a la distribución de los puntos muestreados de la distribución de la función de coste obtenida (valores de la función de coste de los pesos obtenidos) a nivel de cuantilas. Como se deduce de la figura 2.24, los resultados obtenidos por los dos métodos a lo largo de las unidades estudiadas presentan una distribución *cuasi-normal*, cosa que permite llevar a cabo un test estadístico (*t-test*) para analizar la relevancia estadística de los resultados obtenidos. Si dos variables se comportan como distribuciones normales, entonces este test permite medir el grado de separabilidad de las dos distribuciones, respecto a su media y desviación. En este caso, el test muestra que la relación $C_{AG} < C_{MLR}$ se cumple con un



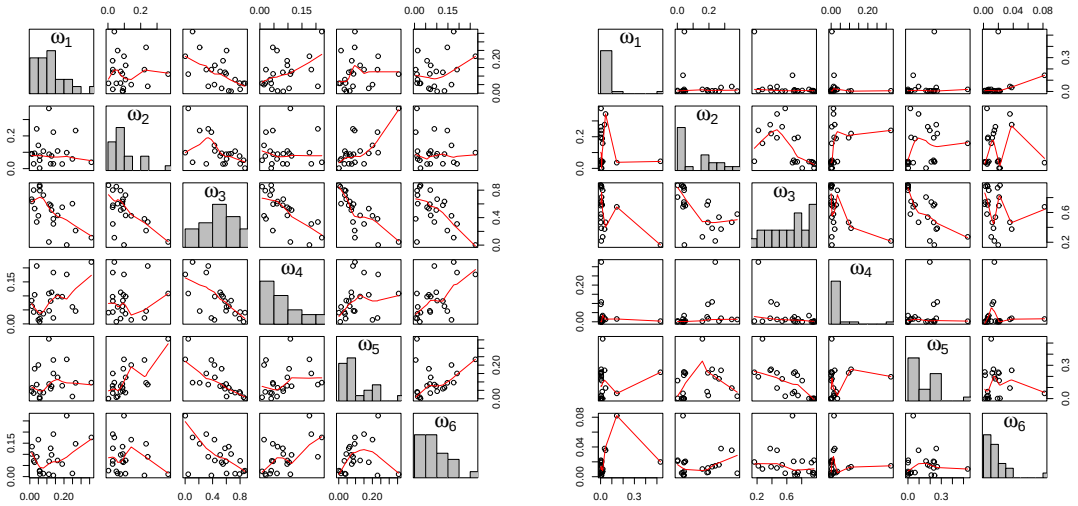
(a) AG (b) MLR

Figura 2.24: Análisis de la *gaussianidad* de los valores de la función de coste (Muestras) calculada a partir de los pesos obtenidos con los dos métodos de entrenamiento objetivo respecto a una distribución normal (Teórico).

nivel de confianza de $p = 3.756 \cdot 10^{-8}$ ([18]).

Finalmente, la figura 2.25 presenta los diagramas de correlación de los pesos obtenidos a partir de los dos algoritmos comparados (MLR y AG). En la diagonal de estas figuras (celdas ii) se muestra el histograma de cada uno de los pesos a lo largo de todas las unidades analizadas en el que se puede observar la distribución de cada peso a lo largo de las unidades. En el resto de subfiguras (celdas ij, con $i \neq j$) se representan las relaciones entre las parejas de pesos que confluyen ($w_i = f(w_j)$). Sobre la distribución de los puntos se dibuja una línea que permite determinar el carácter de la correlación entre los pesos: lineal, cuadrática, exponencial, etc. Como se puede observar en la figura 2.25, las relaciones entre los pesos obtenidos por el método el método basado en MLR son más dispersas que en los del AG, cosa que parece indicar que AG es capaz de descubrir relaciones de orden superior entre los pesos, relaciones no observadas por el MLR debido a su modelado lineal de los datos. Además, como se ha comentado también en un punto anterior de este mismo apartado, el hecho de trabajar con unos subcostes sesgados y analizar el corpus a nivel de unidades (no de grupos de unidades, p.ej. nasales, o a lo largo de todo el corpus), parece provocar que w_3 (el coste por duración de unidad) se convierta en el más importante (mayor ponderación) a la hora de escoger las unidades durante la fase de selección del CTH-SU.

[18] Por lo tanto, se puede asegurar que el algoritmo genético presenta unas soluciones claramente mejor adaptadas al problema (con una función de coste menor) que el método basado en MLR, con una seguridad del $(1 - p) \cdot 100\,\%$, en este caso, prácticamente del $100\,\%$.



(a) MLR (b) AG

Figura 2.25: Correlaciones e histogramas de los pesos obtenidos a partir de los dos algoritmos de ajuste objetivo de pesos estudiados, donde $\omega_1 \leq \omega_i \leq \omega_3$ representan los pesos de unidad (w_j^t) y $\omega_4 \leq \omega_i \leq \omega_6$, los de concatenación ($w_j^c$).

En cuanto al coste computacional de los algoritmos comparados, cabe decir que el del AG es mucho más elevado que el del MLR. Sin embargo, el coste computacional del AG crece linealmente con el número de instancias consideradas (número de realizaciones de la unidad analizada), en contra de lo que pasa con el WSS, que crece exponencialmente. Así pues, en lo que se refiere a la velocidad del método, el AG se encuentra entre el MLR y el WSS, pero aportando unos mejores resultados, como se ha visto a lo largo de las pruebas realizadas durante este primer experimento.

Por otra parte, encontrar la solución *óptima*, es decir, el mínimo global del problema, no es una tarea imposible, pero, a efectos prácticos, se convierte en una tarea difícil de conseguir. Para el método WSS, se hace imprescindible realizar una discretización intensiva del espacio (con un paso muy reducido entre los valores de los pesos) con el fin de no dejar de lado ninguna solución potencial. Este proceso implicaría varias semanas de ejecución o incluso meses, extrapolando los resultados presentados por (Hunt y Black, 1996) para este mismo método (trabajando con entre 3 y 5 valores para cada peso a ajustar y con 10 expresiones —en inglés, *utterances*— de entrenamiento, el método tarda más de 150 horas trabajando sobre un Sun SPARCStation 20). En el caso de trabajar con el AG, se tendría que incluir algún método elitista o de especiación (ver sección 2.2.1) con el fin de encontrar al individuo mejor adaptado al entorno después de varias ejecuciones del algoritmo genético.

Experimento 2 - Ajuste interactivo de los pesos mediante AGI

Una vez validada la viabilidad de los algoritmos genéticos como herramienta para abordar el problema del ajuste de los pesos de la función de coste mediante el experimento anterior, se estudia la incorporación explícita de la subjetividad de los usuarios en esta tarea. Para ello, se utiliza la plataforma *web* desarrollada para el ajuste de los pesos de la función de coste de un CTH-SU (ver anexo C.2) en la que se incorpora un algoritmo genético interactivo. Este experimento pretende, una vez contrastado el buen funcionamiento de la plataforma desarrollada y ajustados pertinentemente los distintos módulos que la conforman, evaluar la viabilidad del algoritmo genético interactivo como método para el entrenamiento de los pesos. En este caso, el entrenamiento de los pesos se realiza de forma global, es decir, se busca disponer de un patrón único de pesos para todas las unidades del corpus.

Teniendo presente los objetivos de esta prueba, se decidió trabajar con tres usuarios de perfiles distintos —*novel*, *especialista* y *experto*— encargados de desarrollar las pruebas para estudiar el impacto de los distintos criterios en el ajuste de pesos. En este caso, se seleccionaron 4 frases (expresiones) de un documental de televisión para el test[19]: *(i)*"*De la seva selva*", *(ii)* "*Fusta de Birmània*", *(iii)* "*I els han venut*" y *(iv)*"*Grans extensions*". De todas las frases —cuyas unidades se extraen del corpus para evitar su selección— se extrae su trascripción fonética y su prosodia, información utilizada como objetivo de la síntesis (estrategia *copy-prosody* o *copy-syntheis* (Toda, Kawai y Tsuzaki, 2004)). Ésta será la información de entrada del bloque de síntesis que incorpora la plataforma interactiva (módulo de selección de unidades más procesamiento digital de la señal para su concatenación, mediante TD-PSOLA en este caso). A lo largo de las iteraciones, el usuario debe escoger al mejor individuo (frase sintética) entre dos posibles candidatos —siguiendo un torneo binario, con $s = 2$ (ver secciones 2.2.3 y 2.3). En todo momento, el usuario dispone de la expresión original (objetivo) como referenica de la comparación y puede escuchar las frases sintéticas candidatas tantas veces como crea oportuno antes de tomar la decisión.

En cuanto al ajuste algoritmo genético interactivo (AGI), el experimento se desarrolló con la siguiente configuración: $p_c = 0.6$, $p_m = 0.1$ (Goldberg, 1989; Goldberg, 2002) y $n = 15$, valores fijados experimentalmente a partir de unos experimentos previos. Asimismo, después de las pruebas realizadas, se observó que entre 6 y 7 era un número de iteraciones suficiente (en promedio) para ajustar los pesos de las frases utilizadas.

En la figura 2.26 se presentan los valores obtenidos para los pesos considerados en el experimento, según cada uno de los perfiles de usuario considerados. Se puede observar una cierta correlación de los resultados obtenidos (una vez promediados para cada perfil) entre los usuarios, principalmente entre el usuario novel y el especialista en tecnologías del habla. No obstante, el usuario experto presenta un patrón con ciertas características particulares, por ejemplo, toma mayor peso el coste de *pitch* (`PIT T`) de entre los costes de unidad. Estos resultados permiten tener una primera estimación de la tendencia que presentan los pesos entrenados mediante el AGI descrito.

[19]Estas frases se escogieron aleatoriamente de entre un conjunto mayor de frases seleccionadas, considerando criterios de cobertura fonética.

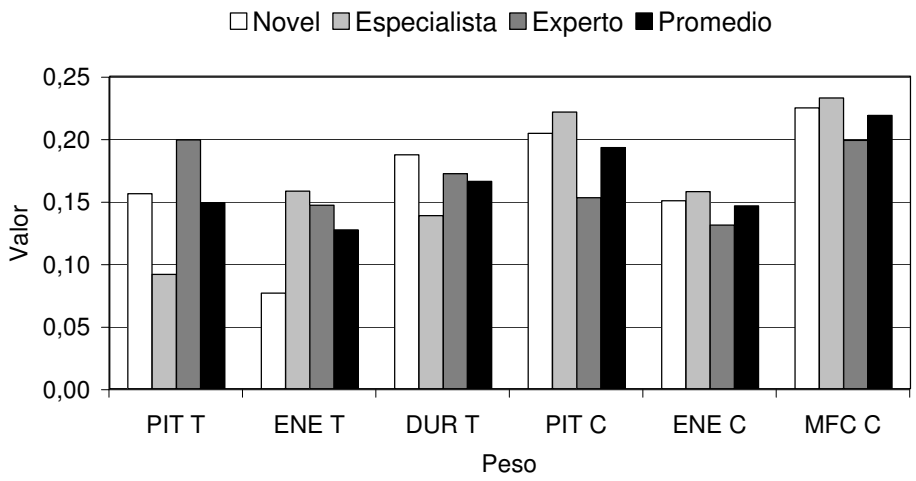


Figura 2.26: Resultado por perfil de usuario y promediado del valor de los seis pesos ajustados subjetivamente mediante el AGI desarrollado.

Del análisis de estos resultados numéricos se concluye que:

- El peso que presenta un valor más elevado es el que corresponde al cálculo del subcoste de concatenación a partir de los parámetros *Mel Frequency Cepstrum* (`MFC C`), el cual pretende codificar el grado de continuidad espectral entre las unidades concatenadas.
- A continuación, en cuanto a la importancia del parámetro, se encuentra el *Pitch* de concatenación (`PIT C`) que indica la importancia que tiene la continuidad de la frecuencia fundamental en la unión de las unidades.
- El peso de unidad (w_j^t) más importante es el que evalúa la similitud de duraciones (`DUR T`) entre la unidad objetivo y la candidata, codificando la velocidad (ritmo) del habla.
- El peso de *Pitch* de unidad (`PIT T`) toma menos importancia según los resultados de que se dispone, aunque como se comenta a continuación, la comparación de entonaciones no es sencilla.
- Los pesos que presentan una menor relevancia en el proceso de selección, según los usuarios, son los asociados a la energía de las unidades, tanto para el coste de unidad (`ENE T`) como para el coste de concatenación (`ENE C`). Eso es debido a que el bloque de síntesis aplica un proceso de ajuste de energía durante la fase de concatenación de las unidades escogidas. Por lo tanto, a los usuarios se les hace difícil descartar las expresiones sintéticas por su variabilidad energética, cosa que provoca que estos pesos tiendan a tomar menor relevancia. De algún modo, estos pesos hacen las veces de elemento de control de la prueba.

Una vez finalizado el experimento, los resultados obtenidos permiten llegar a una serie de conclusiones respecto a las pruebas en sí y al comportamiento de los pesos obtenidos. Asimismo, el experimento permite fijar algunos problemas y funcionalidades de la plataforma desarrollada. Después de distintas deliberaciones entre los tres evaluadores que realizaron el test, se llegaron a las siguientes conclusiones:

- Se hace difícil mantener el mismo criterio de comparación de individuos a lo largo de toda la prueba (consistencia del usuario). Asimismo, resulta complicado que los distintos usuarios utilicen el mismo criterio para seleccionar las frases sintéticas candidatas. Esto se debe, por un lado, a que cada perfil de usuario fija su propio criterio, y por otro, al hecho que, en este experimento, todas las unidades de la frase eran susceptibles de ser cambiadas en cada iteración del AGI (ajuste global de pesos).
- La aparición de un error en una palabra de la expresión (p.ej. un pequeño ruido o un fonema erróneo) implica que aquella realización quede descartada ante la otra frase que forma la pareja de candidatas, favoreciendo una de las dos configuraciones de pesos, cuando quizás el error no es debido al valor de los pesos sino a un problema en la etiquetado del corpus. Por lo tanto, resulta fundamental disponer de un corpus de voz etiquetado de forma robusta para evitar este tipo de situaciones —este fue uno de los motivos por los que se desarrolló el método de ajuste robusto de las marcas de *pitch* que se describe en el capitulo 4 del presente trabajo de investigación.
- En determinadas expresiones, y después de un cierto número de iteraciones, las diferencias entre las frases sintéticas resultaban prácticamente imperceptible, por lo que la prueba se convierte en tediosa al no aportar nueva información, provocando la fatiga del usuario. Esta situación está motivada, por un lado, por la necesidad de realizar un número importante de comparaciones antes de llegar a una solución buena (convergencia del AGI) y por otro, debido al tamaño del corpus de voz utilizado —como ya se ha comentado, se trata de un corpus de tamaño moderado, con un porcentaje elevado de unidades que disponen de un número reducido de realizaciones (por muchas variaciones que sufra el patrón de pesos considerado, si el número de opciones es reducido, fácilmente se repetirán las unidades seleccionadas a lo largo de las iteraciones). Este problema puede ser abordado mediante el agrupamiento de unidades fonéticamente similares —ver anexo A para una primera propuesta.
- Finalmente, comentar que las expresiones originales del documental pertenecen a un locutor distinto al del corpus de voz utilizado para la síntesis, presentando diferencias de prosodia importantes —como también sucede en (Lee, Lopresti y Olive, 2001), cuestión que se debe controlar en futuras pruebas.

Comparativa de los resultados con los métodos objetivos: A partir de los resultados de este experimento, se obtiene un patrón de pesos (columna 'Promedio' en la figura 2.26) que puede ser comparado con el patrón de pesos que se obtiene de promediar los resultados obtenidos a nivel de unidad con los métodos objetivos estudiados con anterioridad (MLR y AG). La figura 2.27 muestra los valores promedio de los pesos obtenidos mediante

el algoritmo AGI, basado en la percepción subjetiva, respecto a los resultados obtenidos mediante MLR y AG, basados en una distancia objetiva. Ambos métodos objetivos hacen hincapié en la importancia del peso `DUR T` respecto al resto de pesos, contrastando con los resultados del AGI donde todos los pesos presentan resultados similares (el peso `MFC C` es ligeramente más importante). Este resultado muestra que los métodos objetivos presentan un comportamiento (tendencia) claramente distinta respecto al proceso propuesto de ajuste subjetivo mediante AGI. Concretamente, AG acentúa al máximo el peso que parece ayudar a minimizar la función de coste (*fitness* del AG), mientras que MLR, que muestra una tendencia similar, presenta un patrón menos sesgado, debido a las restricciones lineales sobre las que se basa. Parece pues, visto el patrón de resultados obtenido mediante el AGI trabajando sobre los criterios subjetivos de los usuarios, que optimizar la distancia objetiva, no hace más que alejar la configuración de pesos de deseada por los usuarios, según el experimento desarrollado, dado el patrón de pesos claramente distinto obtenido con AGI respecto a MLR y AG.

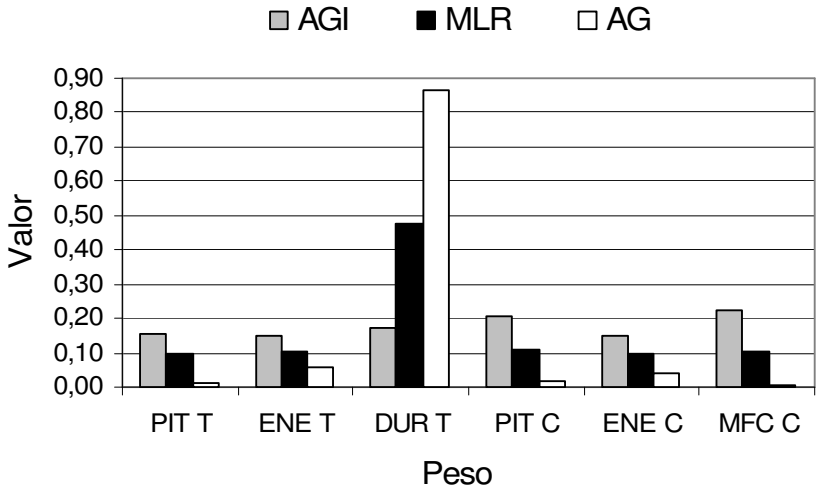


Figura 2.27: Patrón global de los pesos obtenidos mediante los tres métodos de ajuste de pesos comparados, promediando los resultados de unidad de los métodos objetivos y los resultados obtenidos para todos los usuarios mediante el método subjetivo.

Una vez finalizado el experimento, se puede concluir que resulta necesario diseñar una técnica de ajuste de pesos que, tomando en consideración los criterios subjetivos del usuario —los métodos objetivos presentan un patrón de pesos claramente distinto—, sea capaz de abordar los dos problemas fundamentales detectados: la dificultad de los usuarios para mantener un criterio consistente a lo largo del proceso evolutivo (*consistencia* del usuario) y la reducción del número de evaluaciones necesarias para conseguir que converja el algoritmo antes de causar una saturación perceptiva al usuario (*fatiga* del usuario). Por ello, y después que se presentaran los algoritmos genéticos interactivos activos (Llorà et al., 2005), descritos en este capítulo, se procedió a adaptar esta nueva filosofía de optimización interactiva para

el problema del ajuste subjetivo de pesos, repitiendo las pruebas realizadas utilizando AGI, tal y como se describe a continuación.

2.4.2. Resultados obtenidos mediante el método basado en aAGI

Una vez conocida la nueva técnica de optimización interactiva eficiente presentada por (Llorà et al., 2005), se procedió a incorporarla a la plataforma interactiva de ajuste de pesos, después de adaptar la propuesta de algoritmo genético interactivo *activo* (aAGI) al problema, como se ha descrito en el apartado 2.3. A continuación, se repitieron las pruebas descritas en el "*Experimento 2*", a partir de las mismas frases, los mismos usuarios —novel, especialista y experto— y el mismo corpus de voz, pero utilizando, en este caso, una población de $n = 200$ individuos para cPBIL, manteniendo los $n = 15$ torneos usados también en el AGI para la interacción con el usuario (torneo binario jerárquico de altura $h = 4, 2^4-1$). Asimismo, a partir de las conclusiones obtenidas con los experimentos previos, se decidió utilizar los seis subcostes con normalización *sigmoidea* (ver ecuaciones (2.11) y (2.12)) para las pruebas, ya que consiguen una distribución más homogénea de los valores.

Los objetivos fundamentales del experimento desarrollado pasan por analizar el impacto de la inclusión del aAGI para resolver el problema del entrenamiento eficiente de los pesos en términos de *(i)* la mejora de la consistencia de las evaluaciones de los usuarios, *(ii)* la reducción de su fatiga, y *(iii)* el aumento de la calidad sintética de la CTH-SU, objetivo final de este trabajo. Para ello, en el experimento se realizan un conjunto de evaluaciones tanto objetivas como subjetivas —comparando las preferencias de nuevos usuarios ante los resultados sintéticos conseguidos por los distintos métodos de ajuste de pesos estudiados en el presente trabajo de investigación—, como se pasa a describir seguidamente.

Análisis 1 - Aumento de la consistencia en el entrenamiento de pesos

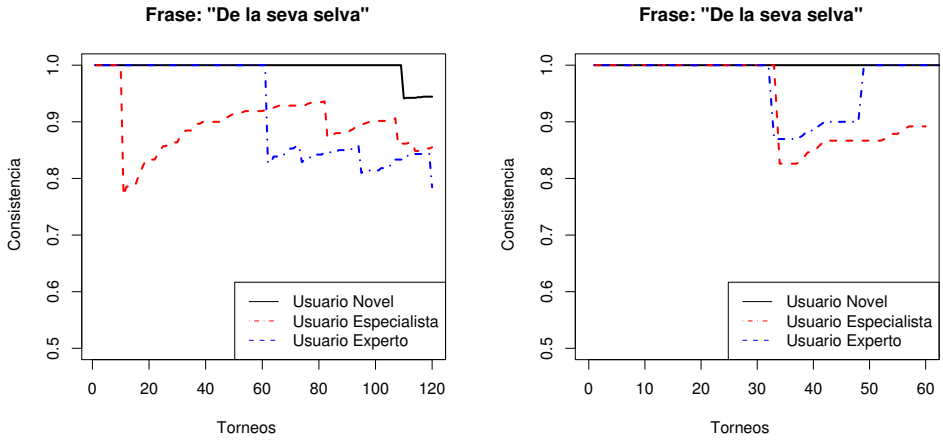
Primero, se analiza la consistencia de los usuarios utilizando la medida $\kappa\left(\mathcal{G}'^{t_f}, \omega\right)$ (t_f indica tiempo final) propuesta en la ecuación (2.21) a lo largo del proceso evolutivo obtenida a partir del experimento desarrollado en (Alías et al., 2004a) —"*Experimento 2*", que se acaba de describir—, en el cual se utilizaba el algoritmo AGI simple para llevar a cabo el ajuste subjetivo de los pesos de la función de coste. Como se puede observar en la parte superior de la tabla 2.3, sólo el usuario *experto* fue consistente en un determinado experimento. La evolución de $\kappa\left(\mathcal{G}'^{t}, \omega\right)$ para las cuatro frases analizadas a lo largo del proceso evolutivo utilizando el AGI simple se presenta en las subfiguras (a) de las figuras 2.28, 2.29, 2.30 y 2.31. En ellas se puede observar como todos los usuarios, independientemente de su perfil —*novel*, *especialista* o *experto*— tuvieron problemas para mantener un criterio consistente a lo largo del experimento utilizando el algoritmo genético interactivo simple. Esto es debido, fundamentalmente, por un lado, al gran número de evaluaciones necesarias antes de conseguir la convergencia del AGI y, por otro, a las pequeñas diferencias perceptuales entre las soluciones candidatas (frases sintéticas obtenidas para distintas configuraciones de pesos) que complican la toma de decisiones. Otro descubrimiento importante de este análisis fue observar la presencia de inconsistencias ya desde el inicio del proceso iterati-

Tabla 2.3: Consistencia final $\kappa\left(\mathcal{G}'^{t_f}, \omega\right)$ (ecuación (2.21), según el perfil de usuario, para las cuatro frases del experimento.

AGI simple **Frase**	**Usuario** **Novel**	**Usuario** **Especialista**	**Usuario** **Experto**
"De la seva selva"	0.944	0.855	0.784
"Fusta de Birmània"	0.857	0.769	0.911
"I els han venut"	0.894	0.867	0.731
"Grans extensions"	0.942	0.800	1.000
AGI activo **Frase**	**Usuario** **Novel**	**Usuario** **Especialista**	**Usuario** **Experto**
"De la seva selva"	1.000	0.892	1.000
"Fusta de Birmània"	1.000	1.000	1.000
"I els han venut"	1.000	1.000	0.948
"Grans extensions"	1.000	1.000	1.000

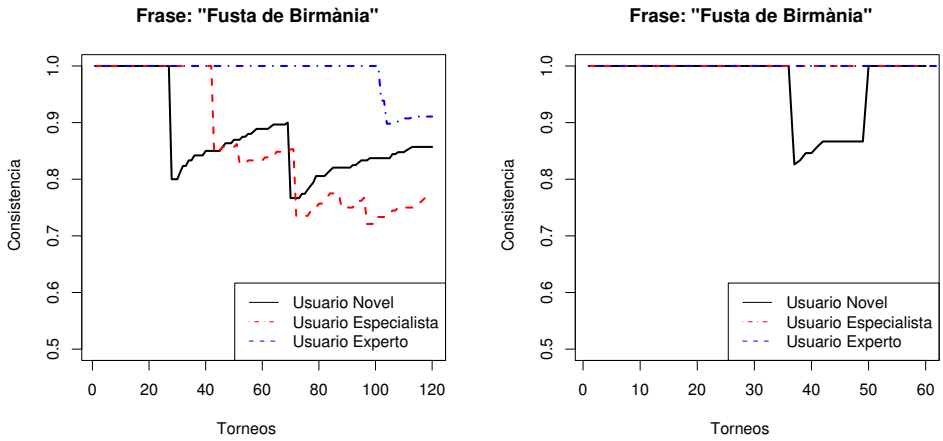
vo del AGI. En promedio, los usuarios tendieron a contradecirse (p.ej. $A > B$, $B > C$ y $C > A$) alrededor del torneo 14, con un promedio de 2.83 contradicciones por ejecución. Estas inconsistencias pueden ser interpretadas como una función de *fitness* ruidosa para el AG, provocando un incremento del número de evaluaciones necesario para que el usuario consiga obtener soluciones de elevada calidad (Miller y Goldberg, 1995; Goldberg, 2002; Sastry y Goldberg, 2002). Esto provoca el aumento de la fatiga del usuario y hace que el proceso de aprendizaje interactivo pierda efectividad.

Seguidamente, se repitió el experimento realizado con anterioridad, pero esta vez reemplazando el AGI simple de la interfaz de ajuste de pesos (ver anexo C.2) por el AGI activo (aAGI) introducido por (Llorà et al., 2005) y resumido en este trabajo. La parte inferior de la tabla 2.3 muestra los resultados del cálculo de la medida de consistencia $\kappa\left(\mathcal{G}'^{t_f}, \omega\right)$ para los tres perfiles de usuario. Asimismo, las sufiguras (b) de las figuras 2.28, 2.29, 2.30 y 2.31 muestran la evolución temporal de $\kappa\left(\mathcal{G}'^{t}, \omega\right)$ a lo largo de los torneos. A simple vista, se puede observar como reemplazar el algoritmo genético interactivo simple por el activo permite aumentar decididamente la consistencia de las evaluaciones de los usuarios (ayuda a mantener su criterio de evaluación), ya que tan sólo dos de los doce experimentos finalizó inconsistentemente —$\kappa(\mathcal{G}'^{t_f}, \omega) < 1$ (ver la porción inferior de la tabla 2.3). No obstante, la consistencia de las ejecuciones utilizando aAGI está muy por encima de la conseguida mediante el AGI simple —sólo un usuario (el experto) fue capaz de ser consistente a lo largo de toda la prueba. Otra de las observaciones interesantes que se extraen del análisis de los resultados utilizando aAGI es que, aún cuando el usuario comete una contradicción en las evaluaciones, la selección *activa* de los torneos de candidatos basado en el grafo de ordenación parcial $\mathcal{G}'$ ayuda al usuario a volver al camino de la consistencia (ver subfiguras (b) de las figuras 2.28, 2.29, 2.30 y 2.31, donde por ejemplo en la figura 2.31(b) el usuario novel recupera la consistencia entorno del torneo 45).



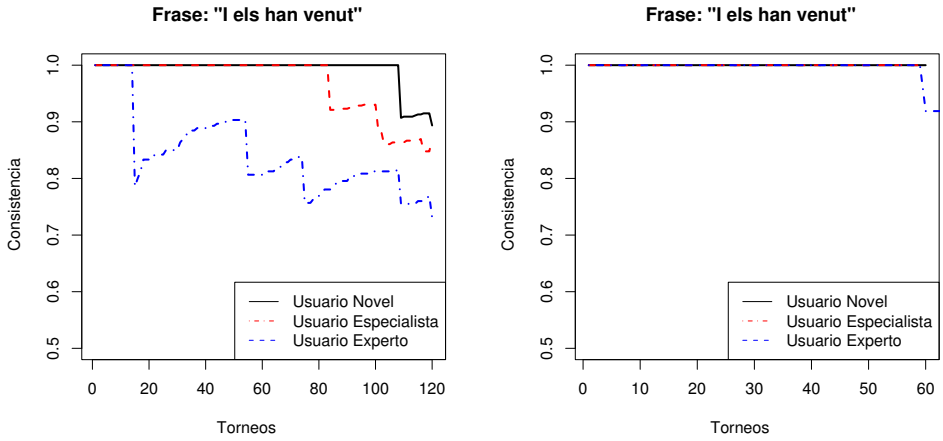
(a) AGI simple (b) AGI *activo*

Figura 2.28: Evolución de la consistencia de usuario evaluada mediante la medida $\kappa\left(\mathcal{G}'^{t}, \omega\right)$ para la frase *"De la seva selva"*. Las figuras comparan la evolución de la consistencia para distintos perfiles de usuario utilizando el algoritmo interactivo simple o el algoritmo interactivo *activo*.



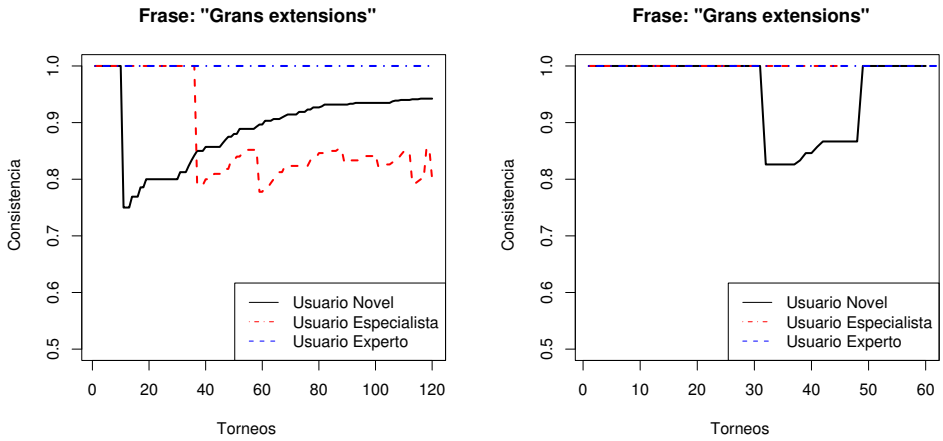
(a) AGI simple (b) AGI *activo*

Figura 2.29: Evolución de la consistencia de usuario evaluada mediante la medida $\kappa\left(\mathcal{G}'^{t}, \omega\right)$ para la frase *"Fusta de Birmània"*. Las figuras comparan la evolución de la consistencia para distintos perfiles de usuario utilizando el algoritmo interactivo simple o el algoritmo interactivo *activo*.



(a) AGI simple (b) AGI *activo*

Figura 2.30: Evolución de la consistencia de usuario evaluada mediante la medida $\kappa\left(\mathcal{G}'^{t}, \omega\right)$ para la frase *"I els han venut"*. Las figuras comparan la evolución de la consistencia para distintos perfiles de usuario utilizando el algoritmo interactivo simple o el algoritmo interactivo *activo*.



(a) AGI simple (b) AGI *activo*

Figura 2.31: Evolución de la consistencia de usuario evaluada mediante la medida $\kappa\left(\mathcal{G}'^{t}, \omega\right)$ para la frase *"Grans extensions"*. Las figuras comparan la evolución de la consistencia para distintos perfiles de usuario utilizando el algoritmo interactivo simple o el algoritmo interactivo *activo*.

Tabla 2.4: Aumento de la consistencia conseguida al reemplazar el AGI simple por el AGI *activo*, calculado como la diferencia absoluta entre las consistencias de cada método presentadas en la tabla 2.3.

AGI simple Frase	**Usuario Novel**	**Usuario Especialista**	**Usuario Experto**	**Usuario Promedio**
"De la seva selva"	5.89 %	4.30 %	27.50 %	12.56 %
"Fusta de Birmània"	16.67 %	30.01 %	9.81 %	18.83 %
"I els han venut"	11.91 %	15.00 %	29.76 %	18.89 %
"Grans extensions"	6.12 %	25.00 %	0.00 %	10.37 %
Promedio	10.15 %	18.58 %	16.77 %	15.16 %

A partir de estos resultados, se puede calcular la mejora en la consistencia conseguida al reemplazar el AGI simple por el AGI activo en el proceso de ajuste de los pesos de la función de coste mediante la interfaz interactiva (ver tabla 2.4). Como conclusión, se puede observar que gracias a utilizar aAGI la consistencia a lo largo del proceso evolutivo mejora de forma evidente, ayudando al usuario a evaluar las soluciones propuestas de forma consistente e inequívoca.

Mejora de la eficiencia del entrenamiento: Una de las razones fundamentales del diseño de los aAGIs se centra en conseguir una reducción importante del número de evaluaciones que debe realizar el usuario, de forma que se consiga reducir su fatiga (Llorà et al., 2005). Aunque el objetivo fundamental de la aplicación del aAGI al problema del ajuste subjetivo de los pesos de la función de coste fue aumentar la consistencia de las evaluaciones del usuario, sustituir el AGI simple por el AGI activo permite, a la vez, reducir el número de iteraciones del proceso evolutivo. Aunque no se tomó ninguna medida en especial para conseguir mejorar la eficiencia del proceso, los resultados obtenidos son realmente buenos. Concretamente, la tabla 2.5 muestra una reducción media del 50 % en el número de evaluaciones que debe realizar el usuario antes de converger, por lo que las pruebas utilizando aAGI duraron, en promedio, la mitad que las realizadas anteriormente mediante el AGI simple (se pasó de unas 6 generaciones —120 torneos— a unas 3 generaciones —60 torneos— en las figuras 2.28, 2.29, 2.30 y 2.31). Así pues, incorporar el aAGI como algoritmo base del proceso de entrenamiento de los pesos permite abordar satisfactoriamente dos de los problemas fundamentales en cualquier proceso de ajuste interactivo (subjetivo): la consistencia y la fatiga del usuario, cuestión que, como se verá en el siguiente análisis, permite, a la vez, conseguir unas soluciones de mayor calidad, en este caso.

Análisis 2 - Evaluación subjetiva de los resultados

Una vez analizado el buen comportamiento objetivo (mejora en la consistencia y la eficiencia, más reducción de la fatiga del proceso), resulta necesario contrastar los resultados

Tabla 2.5: Mejora de la eficiencia conseguida al reemplazar el AGI simple por el AGI *activo*, calculada como el cociente entre el número de torneos necesarios antes de converger.

Frase	**Usuario Novel**	**Usuario Especialista**	**Usuario Experto**	**Usuario Promedio**
"De la seva selva"	2.00	2.00	2.00	2.00
"Fusta de Birmània"	2.00	2.00	2.00	2.00
"I els han venut"	2.00	2.00	2.00	2.00
"Grans extensions"	2.00	2.67	2.00	2.23
Promedio	2.00	2.17	2.00	2.06

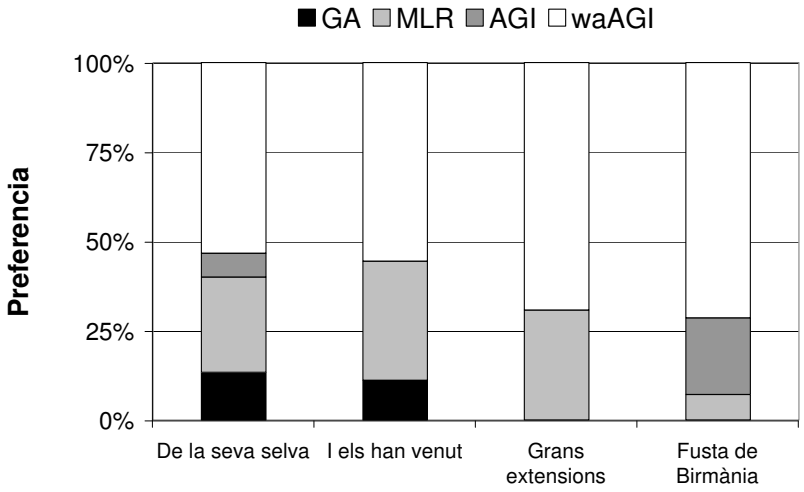


Figura 2.32: Preferencias de los usuarios según el método de ajuste de pesos de la función de coste utilizado para las cuatro frases consideradas.

obtenidos subjetivamente, es decir, la calidad sintética obtenida por las configuraciones de pesos ajustadas mediante el uso del algoritmo genético interactivo activo descrito en este capítulo. Para ello, se desarrolló un test perceptual que incorpora las frases sintéticas generadas a partir de los cuatro esquemas de ajuste de pesos estudiados a lo largo del presente capítulo: aAGI y AGI simple —basado en criterios subjetivos— o MLR y AG — basados en una medida objetiva (distancia cepstral). La función de coste, vuelve a considerar los subcostes prosódicos descritos en la sección 2.1.4 (coste de *pitch*, energía media y duración de la unidad, más coste *pitch*, energía y MFCC en el punto de concatenación). En este caso, 10 usuarios independientes (que no participaron en la fase de ajuste de pesos) del Departamento de Comunicaciones y Teoría del Señal de Enginyeria i Arquitectura La Salle fueron los encargados de realizar las pruebas subjetivas. Cada test subjetivo duró entre unos 15 y 20 minutos por usuario.

Después de analizar los resultados del experimento previo, se pudo observar que el resultado del entrenamiento de los pesos, aún presentando un grado de consistencia elevado, presentaba patrones de pesos distintos según el perfil de usuario (novel, especialista o experto). Por esta razón, y después de presentar los resultados a los usuarios encargados de ajustar los pesos y contrastar que continuaban prefiriendo sus patrones de pesos ante los obtenidos por el resto, se decidió llevar a cabo un test perceptual para validar los resultados de la fase de entrenamiento de los pesos. Para ello, se presentó a los evaluadores los resultados sintéticos obtenidos a partir de los distintos patrones de pesos entrenados mediante aAGI por los tres perfiles de usuario considerados. Las frases se presentaron en parejas de dos para poder llevar a cabo un test de preferencia A/B/indiferente. Éstas se presentaron de forma aleatoria y se realizaron todas las comparativas posibles para determinar la preferencia de cada usuario. Una vez finalizada la prueba, por un lado, se pudo observar de las preferencias de los evaluadores una tendencia clara a escoger las configuraciones de pesos propuestas por el usuario *experto* respecto a los otros dos perfiles (44 % de las elecciones), pero por otro lado, también se concluyó que no hay ningún perfil que quede totalmente descartado (27 % de las elecciones para el usuario especialista y 29 % para el *novel*). De algún modo, estos resultados demuestran la dificultad de la tarea de ajustar los pesos de la función de coste, ya que, aunque los resultados obtenidos presentan una cierta tendencia hacia el criterio del usuario experto, esto no se cumple para todos los evaluadores, seguramente, porqué toman en consideración otros criterios subjetivos distintos a la hora de valorar los resultados sintéticos.

Seguidamente, se utilizó cada una de las configuraciones de pesos ganadoras (*winning* aAGI o waAGI) de cada frase para incorporarla a un nuevo test de preferencia en el que se incluyen los resultados sintéticos de las frases obtenidos a partir de los patrones de pesos entrenados mediante AGI, MLR y AG. La figura 2.32 presenta los resultados obtenidos de esta prueba. Como se puede observar, en más del 50 % de las ocasiones los usuarios seleccionaron la frase sintética generada a partir de las combinaciones de pesos waAGI obtenidas de la primera fase del test. Este experimento demuestra la mejora subjetiva conseguida gracias a la incorporación del método basado en aAGI al proceso del entrenamiento eficiente de los pesos de la función de coste a partir de los criterios subjetivos de los usuarios.

Estos resultados representan una clara evidencia de la importancia de conseguir disponer

de un proceso eficiente de ajuste subjetivo de pesos para conversión de texto en habla basada en selección de unidades. No obstante, una vez analizados con más detalle los resultados, se observa que a mayor dificultad de escoger entre los patrones de pesos candidatos del aAGI en el primer test (es decir, mayor número de comparativas antes de tomar una decisión) —primera prueba perceptual—, menor es el grado de aceptación de la configuración ganadora waAGI entre los participantes del test —segunda prueba perceptual. Este es un problema bastante conocido en el ámbito de los algoritmos genéticos, donde cuanto menor sea la diferencia entre los candidatos, mayor dificultad presentará el problema para ser solucionado mediante este tipo de métodos de optimización (Goldberg, 2002). En este contexto, parece que puede ser interesante agrupar las unidades en *clusters* para ayudar a los evaluadores a focalizar las comparaciones en determinadas diferencias de la señal, en lugar de llevar a cabo un ajuste global de los pesos —ver anexo A para una primera propuesta de algoritmo de *clustering* basado en criterios de distribución de las unidades en el corpus y similitud fonética.

2.5. Discusión

En este capítulo se ha abordado uno de los problemas más complejos en el contexto del diseño de los sistemas de conversión de texto en habla basados en selección de unidades: el ajuste subjetivo de pesos eficiente. Durante la investigación que se ha llevado a cabo, se ha podido contrastar la gran dificultad que esta tarea conlleva y lo complicado que resulta dejar el proceso de ajuste en manos de las personas. No obstante, la técnica que se ha presentado basada en un sistema de ajuste subjetivo controlado por un algoritmo automático permite conseguir unos resultados satisfactorios en términos de la calidad de la síntesis, junto a la mejora de la consistencia y la reducción de la fatiga que conlleva este tipo de procesos interactivos. El trabajo se ha centrado en el diseño y validación de esta nueva estrategia que permite el entrenamiento conjunto de los pesos de unidad y de concatenación (sin restricciones lineales), basada en la aplicación de los algoritmos genéticos al problema, primero, validados como método automático de ajuste de pesos (Alías y Llorà, 2003), a continuación adaptados a un entrenamiento subjetivo mediante los algoritmos genéticos interactivos (Alías et al., 2003; Alías et al., 2004a), terminando con la propuesta de algoritmos genéticos interactivos activos (Alías et al., 2006), adaptando el trabajo de Llorà et al. (2005) al problema.

A continuación, se discuten algunas cuestiones relacionadas con el trabajo desarrollado hasta el momento.

Normalización de los datos y precisión de los individuos

Proceso evolutivo: Dos de los elementos importantes en lo que se refiere a la adaptación del método de optimización propuesto (aAGI) al escenario de la optimización de los pesos de la función de coste de selección son: *(i)* decidir dónde se efectúa la normalización de los pesos y *(ii)* definir la precisión de los individuos. Experimentalmente, se ha decidido

trabajar con una precisión de dos decimales, ya que, en las pruebas realizadas ha demostrado aportar buena sensibilidad a la selección de unidades. Tal y como se muestra en la tabla 2.6, sobre el conjunto de frases utilizado en las pruebas subjetivas, esta resolución consigue un ratio de 0.677 entre el número de pesos (genotipos) y el número de frases distintas sintetizadas a partir de ellos (fenotipos) —se computan como distintas todas aquellas frases que continenen, como mínimo una unidad distinta al resto de frases. Si se trabaja con mayor precisión, por un lado, las diferencias entre los resultados sintéticos serán menos perceptibles, y por el otro, se aumenta el ruido del proceso evolutivo, ya que el aAGI no es capaz de reaccionar de forma distinta (escoger soluciones distintas) ante cambios tan pequeños en los valores de los pesos. Por otro lado, trabajar con menor precisión (1 decimal) provocaría un ratio —entre el número de vectores de pesos y el número de frases sintéticas distintas— de 1, por lo que el barrido de unidades del corpus sería demasiado pobre. No obstante, será necesario realizar un estudio más exhaustivo en lo referente a la resolución de los pesos en futuras investigaciones.

En cuanto a la normalización de los individuos (siguiendo las restricciones de las ecuaciones (2.13) y (2.14)), en el esquema de ajuste de pesos basado en aAGI, resulta necesario mantener las distribuciones sin normalizar los valores de los pesos a lo largo del proceso iterativo para permitir la convergencia cPBIL, debido a que éste representa a la población como distribuciones normales $\mathcal{N}(\mu, \sigma)$. Así pues, la normalización se realiza una vez ha finalizado el ajuste interactivo de los pesos. Si esto no fuera así, experimentalmente, se ha podido observar que la normalización intermedia de los datos provoca inestabilidades en el proceso evolutivo ya que los valores de μ y σ del vector patrón de pesos van oscilando provocando que el aAGI (el método de optimización cPBIL y de generación del *fitness* sintético mediante ε-SVM) deje de funcionar satisfactoriamente.

Normalización de los subcostes: Aunque el objetivo final de este trabajo de investigación se ha centrado en el entrenamiento de los pesos, del pequeño análisis que se ha realizado sobre los distintos métodos de normalización de los subcostes, se puede deducir que la mejor estrategia pasa por utilizar la función sigmoidea, ya que permite saturar los valores extremos (valor 1 o 0), ensanchando el rango de valores *críticos* de los subcostes centrados en la media de sus valores para cada una de las unidades del corpus. Sin embargo, la normalización *Max-min* sufre la presencia de valores espurios —generalmente, debidos a errores de etiquetado— (máximo y mínimo) que pueden provocar una compactación exagerada del rango de valores *críticos*. Por otro lado, la normalización *mean-std* (o *z-score*) aunque es menos sensible a los valores espurios, no consigue evitar su influencia debido a que la media los considera. Asimismo, no asegura que el rango de los subcostes esté acotado entre $[0, 1]$ debido a las distintas unidades de trabajo de cada subcoste, por lo que influyen en el cálculo de la función de coste pueden enmascarar la importancia que deben tomar los pesos en la definición de esta función. Otro de los elementos a destacar es que la normalización sigmoidea consigue una distribución muy similar entre los distintos subcostes, cuestión fundamental para el enfoque seguido en este trabajo de investigación donde la correlación entre la función de coste y los criterios perceptuales de los usuarios se mapea a través del entrenamiento subjetivo de los pesos. No obstante, parece interesante continuar

trabajando en el diseño de nuevas funciones de coste más genéricas —como las descritas en (Toda, Kawai y Tsuzaki, 2004), por ejemplo—, que permitan disponer tanto de pesos como de subcostes sin restricciones de linealidad.

Tabla 2.6: Relación entre el número de frases sintéticas distintas obtenido a partir de los pesos utilizados en la selección de las frases analizadas durante el entrenamiento subjetivo de pesos. La columna ratio indica la relación entre el número de frases candidatas y el número de vectores de pesos utilizados para obtenerlas.

Frase	**Candidatas**	**Pesos**	**Ratio**
"De la seva selva"	82	132	62.12 %
"Fusta de Birmània"	82	118	69.49 %
"Grans extensions"	81	109	74.31 %
"I els han venut"	75	125	60.00 %
Total	406	602	67.44 %

Métodos objetivos

En este trabajo de investigación, la distancia Euclídea cepstral ha sido utilizada como medida objetiva para determinar la similitud entre las unidades comparadas en el contexto de los métodos de ajuste de pesos totalmente automáticos (MLR y AG), utilizada típicamente para este tipo de aproximaciones. No obstante, como se ha comentado en diversos trabajos, p.ej. (Campillo y Rodríguez Banga, 2003; Campillo, 2005), este tipo de distancias se centra fundamentalmente en contabilizar la diferencia espectral entre las unidades, dejando de lado otras informaciones que no quedan claramente reflejadas en el espectro, como por ejemplo, la información relacionada con la frecuencia fundamental (aunque existe una cierta relación entre el la forma de la envolvente espectral y la frecuencia fundamental de la señal), o la duración de las unidades. Este puede ser el motivo que la propuesta de método objetivo basada en algoritmos genéticos clásicos, aunque consigue unos pesos mejor ajustados a la distancia objetiva que los métodos WSS y MLR, los resultados obtenidos perceptualmente demuestran que subjetivamente no es la mejor solución —tanto en el patrón de pesos global obtenido en la figura 2.27, como los resultados del experimento perceptual de la figura 2.32, donde se comparan los cuatro métodos analizados (MLR, AG, AGI y aAGI). Por ejemplo, en la figura 2.32, los resultados obtenidos por el algoritmo genético son los menos votados de los comparados. Sin embargo, el objetivo de esta tesis no es el de desarrollar un método automático de ajuste de pesos modelando de forma objetiva la percepción humana, sino, disponer de un método eficiente de ajuste de los pesos a partir de los criterios subjetivos de los usuarios.

Mejora de la eficiencia del algoritmo genético

Los algoritmos genéticos proponen una metodología de trabajo genérica para la solución de problemas de optimización, una vez definida la representación de los individuos, las estrategias de cruce y mutación escogidas, el método de selección, etc. Asimismo, para cada problema concreto, además de considerar su función de evaluación específica, se deberá buscar la mejor representación y los mejores operadores posibles (Davis, 1991; Michalewicz, 1992).

Reducción de la fatiga del usuario: Con el objetivo de minimizar el número de comparaciones que debe llevar a cabo el usuario durante el proceso evolutivo, el algoritmo genético interactivo activo (aAGI) incorpora un proceso automático que guía las comparaciones presentadas al usuario. Concretamente, el proceso automático controla que una pareja de soluciones que haya sido anteriormente comparada satisfactoriamente por el usuario —es decir, no se encuentre dentro de un ciclo—, no tenga que volver a ser evaluada[20]. En este caso, el aAGI evalúa por el usuario la pareja de individuos comparada, evitando que el usuario deba volver a escucharlos otra vez. En este caso, se está aumentando la velocidad del proceso, gracias a considerar las evaluaciones ya realizadas anteriormente en el proceso evolutivo. Por el momento, dada la buena consistencia de los resultados obtenidos parece que esta mejora de la eficiencia del aAGI es positiva —si esto no fuera así, se deberían presentar todas las parejas de comparación (equivaldría, de algún modo, a incorporar puntos de control —A-B vs. BA— dentro del proceso interactivo). No obstante, en un futuro, se pretende estudiar con más detalle el impacto de este proceso automático sobre un conjunto de pruebas más amplio.

Especiación: A partir de los resultados obtenidos en este trabajo de investigación, se intuye que el problema del ajuste subjetivo de pesos de la función de coste de selección presenta ciertas particularidades especiales. Después de realizar las pruebas tanto de entrenamiento como de evaluación, se observa que existen distintos perfiles de preferencia (criterios subjetivos distintos), aunque demuestren una cierta tendencia común. Parece pues intuirse que se trata de un problema multimodal. En este contexto, donde pueden existir diferentes soluciones más o menos bien adaptadas al problema, puede resultar interesante considerar el concepto de *especiación* de las soluciones. Para ello, se puede incorporar el concepto de *especie* dentro de la arquitectura clásica de un algoritmo genético, con el objetivo de encontrar las múltiples soluciones óptimas al problema. En este caso, se incorporan distintas especies de individuos dentro del mismo medio para que evolucionen conjuntamente tratando de encontrar la *mejor* solución al problema. Normalmente, se trabaja con especies distantes entre sí (situadas en zonas distintas del medio), por lo que la probabilidad de que individuos de distintas especies se aparejen es baja. En este caso, la competencia por el medio

[20]Se puede volver a presentar a un usuario una misma pareja de soluciones debido a que se repita la configuración de pesos (el mismo genotipo) o porqué se haya obtenido la misma frase sintética (el mismo fenotipo) —es decir, la nueva configuración de pesos ha recuperado la misma secuencia de unidades que otra configuración anteriormente evaluada.

continúa existiendo, pero no es una competencia global sino que sólo se desarrolla dentro de la subpoblación escogida. De este modo, se consigue aumentar el grado de diversificación de la población, cuestión que puede ayudar al proceso de búsqueda para encontrar una mejor solución o conjunto de soluciones en un entorno multimodal. Como contrapartida, esto provocará una ralentización del proceso evolutivo, ya que se dispone de más grupos de individuos (especies) a evolucionar (Goldberg, 1989; Goldberg, 2002). Se pretende estudiar en un futuro la incorporación de las técnicas de especiación al problema del ajuste subjetivo de pesos para analizar de forma más detallada el carácter multimodal del mismo (que se intuye de los resultados obtenidos hasta el momento).

Asimismo, en este contexto, resulta necesario definir algún criterio para obtener las mejores configuraciones de pesos sin necesidad de realizar pruebas subjetivas formales entre las candidatas (proceso extremadamente costoso). Para ello, se pretenden estudiar técnicas de integración de los distintos criterios de usuario, por ejemplo, mediante procesos de entrenamiento conjuntos —distintos usuarios entrenando a la vez los pesos para un grupo de unidades de forma cooperativa— o bien, definiendo técnicas robustas de integración de los grafos de ordenación parcial obtenidos por los distintos usuarios. De este modo, se puede conseguir que afloren las configuraciones óptimas, situadas en la zona superior de la jerarquía definida por el grafo normalizado (es decir, son las más votadas/preferidas globalmente por los usuarios). Esta es una de las líneas prioritarias del trabajo de investigación que se inicia a partir de esta tesis en la línea del ajuste de pesos subjetivo eficiente presentado.

Agrupación de unidades

Como se ha comentado en este capítulo, el entrenamiento de pesos se puede realizar a tres niveles: para cada unidad, por grupo de unidades o para todo el corpus (Hunt y Black, 1996; Black y Taylor, 1997a). Hasta el momento, las pruebas realizadas se han llevado a cabo a nivel de unidad (objetivas) o a nivel global (subjetivas). No obstante, actualmente, se está trabajando en el ajuste de pesos para grupos de unidades, como nivel intermedio a las estrategias de entrenamiento desarrolladas hasta el momento —como se ha llevado a cabo en otros trabajos, como p.ej. (Campillo, 2005). En el anexo A se describe una primera aproximación para particionar las unidades del corpus de voz, considerando sus particularidades fonéticas y su distribución a lo largo del corpus. El objetivo es disponer de un nivel de ajuste de pesos que sea abordable mediante el paradigma de entrenamiento desarrollado —basado en la interacción con el usuario guiada mediante un algoritmo genético interactivo activo— sobre grupos homogéneos de unidades. Por un lado, el entrenamiento subjetivo de los pesos a nivel de unidad es inabordable subjetivamente y puede ser insuficiente estadísticamente para determinados tipos de unidad, si no se controla el número mínimo de realizaciones por unidad debido al propio balanceo fonético del idioma. Por otro lado, el ajuste subjetivo global de los pesos puede generar un patrón de pesos demasiado promediado, incapaz de disponer de los matices particulares de las unidades, p.ej. se puede intuir que el patrón de pesos de las unidades sonoras debe ser distinto al de las sordas, tanto por la información de *pitch* como en los puntos de concatenación. De este modo, se pretende concluir el proceso iniciado en este trabajo de investigación definiendo un marco de entrenamiento más allá de

las pruebas de resíntesis desarrolladas, que permita obtener los pesos ajustados subjetivamente para grupos de unidades con perfiles acústicos similares, cuestión que está siendo abordada en la actualidad.

Medida de consistencia

A lo largo de este trabajo, se ha colaborado estrechamente con el grupo de investigación del Illinois Genetic Algorithms Lab, de la Universidad de Illinois en Urbana-Champaign. Gracias a esta colaboración, se ha podido desarrollar la propuesta de ajuste subjetivo de pesos eficiente. Asimismo, esta colaboración ha permitido proponer alguna cuestión, que desde la resolución del problema planteado, es útil para otros problemas de interacción. Concretamente, se ha presentado una medida que permite validar la calidad de las soluciones propuestas por un usuario a lo largo de un proceso interactivo basado en la comparativa de parejas de soluciones. A partir de un grafo de ordenación parcial de las soluciones (Llorà et al., 2005), en este trabajo se ha definido un método que permite determinar la consistencia del usuario mediante la detección de los ciclos presentes en el grafo (un ciclo $A > B$, $B > C$ y $C > A$ representa una inconsistencia), junto a una medida que permite computar el grado de consistencia de ese proceso interactivo, la medida $\kappa\left(\mathcal{G}'^{t}, \omega\right)$ (ver ecuación (2.21)). Este proceso puede ser aplicado a cualquier otro problema evolutivo que necesite de la interacción del usuario para incorporar sus criterios subjetivos en la resolución del mismo.

Actualmente se sigue trabajando en esta línea de investigación para mejorar el control de la consistencia del usuario por parte del administrador de las pruebas. Para ello, se está trabajando en la incorporación de información visual en tiempo de ejecución para ayudar a optimizar el proceso evolutivo de convergencia, que por el momento está en manos del usuario y de la capacidad del algoritmo genético interactivo activo y el torneo jerárquico de soluciones para volver a presentar una pareja de comparativas que permita romper un ciclo de las evaluaciones del usuario. En (Llorà et al., 2006) se presentan los primeros pasos realizados en este sentido sobre un problema de optimización simple.

Otros elementos de mejora

Las pruebas desarrolladas, además de haber sido útiles para validar la viabilidad de la propuesta, han permitido observar diversas cuestiones relacionadas con el ajuste subjetivo de los pesos. Concretamente, se ha podido constatar que incorporar el módulo de modificación prosódica —en este caso, TD-PSOLA— en el proceso de ajuste de los pesos —como se indica en (Meron y Hirose, 1999)—, permite, por un lado, considerar el impacto de la modificación de la señal para adaptarse a la prosodia deseada (objetivo) durante el proceso de selección de unidades, pero por otro lado, provoca que las diferencias entre las frases sintéticas sean menores, según la percepción de los usuarios a lo largo de las pruebas (tanto las realizadas utilizando AGI como aAGI). Por otro lado, no incorporar el módulo de modificación prosódica en el proceso de ajuste de los pesos permite independizar el resultado del entrenamiento del método de modificación de la señal utilizado. Para ello, será necesario desactivar el bloque de PDS, sustituyéndolo por un proceso de concatenación directa de

las unidades seleccionadas. Por ejemplo, recientemente, en el trabajo de Chen, Chen y Kao (2006) se propone un método para estimar la degradación de la señal de voz debida a las modificaciones introducidas por TD-PSOLA. Mediante este modelo se puede estimar la calidad final de la señal, por lo que las estrategias que no incorporan la modificación prosódica de la señal tienen un método para incorporar esta información al proceso. En un futuro se pretende estudiar los resultados obtenidos según esta segunda estrategia para compararlos con el método utilizado hasta el momento.

En otro orden de cosas, los usuarios han destacado la necesidad de disponer de la opción de indicar que cuando se selecciona la opción *indiferente* de la plataforma, debido a la aparición de dos soluciones perceptualmente muy buenas o muy malas, perdiéndose el carácter de la similitud para el proceso interactivo. Esta situación provoca que la mayoría de los usuarios marcaran ese torneo (comparativa) como empate, provocando que se modele peor el criterio del usuario —cuanto mayor sea el número de empates, menos claro queda el criterio del usuario, por lo que se dificulta la tarea de extraer la función aproximante y generar el *fitness* sintético para presentar al usuario soluciones en las que potencialmente estará más interesado. La teoría de los aAGI propone, ante este problema, incorporar en la plataforma de test la opción de indicar que dos soluciones *no son comparables*. Esta puede ser una primera opción a considerar en un futuro, aunque será necesario controlar que esto no provoque la creación de subgrafos disjuntos (no conectados) dentro del grafo de ordenación parcial que recoge las evaluaciones del usuario. También se pretende ponderar de algún modo las soluciones de alta calidad que aparecen esporádicamente en las primeras iteraciones, para ayudar a la convergencia del algoritmo, para así, reducir la fatiga del usuario.

Capítulo 3

Conversión de texto en habla multidominio

En este capítulo se describe otra de las contribuciones del presente trabajo de investigación: la conversión de texto en habla multidominio (CTH-MD). Esta estrategia constituye un paso más hacia la consecución de un CTH genérico de alta calidad, siguiendo la línea de investigación que prima la calidad de la síntesis ante la generalidad de la conversión de texto en habla (Yi y Glass, 1998; Taylor, 2000). Así pues, la CTH-MD 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 contexto es necesario que el sistema de CTH-MD conozca, durante el proceso de conversión de texto en habla, qué dominio o dominios son los más adecuados para poder sintetizar de la forma más apropiada posible el mensaje de entrada, con el objetivo de mejorar la calidad de la señal sintética obtenida (Black, 2002).

Por lo tanto, 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 (Black, 2003) (p.ej. en (Iida et al., 2000; Iida et al., 2003) 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 (Yamagishi et al., 2003; Yamagishi et al., 2005) (p.ej. órdenes militares *vs.* tristeza o miedo (Johnson et al., 2002), mensajes conceptualmente complejos *vs.* voz de niño (Black, 2003)), y por otro lado, existen mensajes que presentan una clara correlación con el modo de locución a utilizar (Iida et al., 2003) (p.ej. mensajes positivos o negativos *vs.* a unos patrones prosódicos determinados (Hamza et al., 2004; Sagisaka, Yamashita y Kokenawa, 2005), o frases más típicas de un niño o de una niña (Black, 2003)). 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 (Campbell, 2002) (p.ej. "*Veo que hay mucha comida en la nevera*", en tono alegre o sarcástico). En este caso, esco-

ger 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, ... del hablante), que también pueden afectar a la comunicación (Campbell, 2005), cuestiones que quedan fuera del alcance del presente trabajo de investigación.

Algunas de las aplicaciones de la CTH-MD abarcan, entre otras: la síntesis multimodal, por ejemplo, la síntesis audiovisual basada en cabezas parlantes (ver apéndice D.2) o la creación de personajes virtuales (ver apéndice D.1), donde a partir de un texto de entrada se busca generar el mensaje audiovisual más adecuado (gestos, expresividad, etc.); los sistemas de diálogo multidominio, donde se puede utilizar la información del dominio de la consulta para adaptar la respuesta del sistema, yendo más allá de la función habitual de los CTH en estos sistemas como meros transmisores de información (p.ej. ver diagrama de bloques en (Pérez-Piñar y García, 2005)), y en general, la CTH-MD, al igual que la CTH clásica, puede formar parte de cualquier sistema de interacción persona-máquina que requiera de un canal de salida oral. Como primer paso para articular la estrategia de CTH-MD expuesta, ésta ha sido desarrollada bajo el marco de los sistemas de conversión de texto en habla basados en selección de unidades (CTH-SU), trabajando con un corpus que contiene distintos dominios independientes. No obstante, la filosofía de CTH-MD es adaptable a otras técnicas de síntesis y tipologías de corpus, como se discute al final del presente capítulo.

En el presente trabajo de investigación, además de presentar una propuesta para la CTH multidominio, también se aborda el problema de determinar de forma automática el dominio más adecuado para llevar a cabo la síntesis del mensaje (texto) de entrada. Esta cuestión se ha abordado a partir de un análisis del texto desde un punto de vista distinto al habitual en CTH, es decir, yendo más allá de las funcionalidades típicas del módulo de procesamiento del lenguaje natural de un CTH. Concretamente, la CTH-MD incorpora un módulo de clasificación de textos a la arquitectura clásica de los sistemas de CTH, que será el encargado de escoger, en tiempo de ejecución, el domino o dominios más adecuados para llevar a cabo la síntesis. El clasificador de textos diseñado en este trabajo incorpora tanto parámetros temáticos como estructurales del texto gracias a la representación del texto utilizada, cuestión que permite abordar satisfactoriamente la clasificación de textos en el contexto de la CTH multidominio presentada. Finalmente, comentar que para poder validar el sistema de clasificación de textos propuesto, se ha escogido trabajar con un corpus multidominio, con distintos subcorpus —uno por dominio—, cada uno de ellos grabado con un estilo de locución particular —se dejan relaciones más complejas para futuros trabajos de investigación. El capítulo finaliza con la descripción detallada de los experimentos desarrollados para estudiar el rendimiento de la propuesta en términos de eficiencia de clasificación y coste computacional, así como la calidad subjetiva obtenida por el sistema de CTH multidominio propuesto.

3.1. Introducción

Como se ha comentado en la sección 1.1 del presente trabajo de investigación, 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 3.1) (Yi y Glass, 1998; Taylor, 2000). 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 (Taylor, 2000). Así pues, el diseño de los sistemas de síntesis procuró, primero, obtener una buena inteligibilidad sintética dejando la mejora de su naturalidad para investigaciones futuras. 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 (o flexibilidad) de la síntesis[1], al utilizar textos pertenecientes al dominio (ya que, si no es así, la naturalidad no es tan elevada). 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 3.1, introducida por Yi y Glass (1998) y posteriormente recogida en (Taylor, 2000). Bajo este punto de vista, el camino para llegar al objetivo marcado se articula a través de la consecución de productos de calidad notable (aunque restringidos a aplicación) a medida que se avanza en la investigación —a diferencia de la *Aproximación A*, donde, para obtener mejoras de calidad significativas, resulta necesario realizar grandes esfuerzos (según el gráfico), al trabajar siempre con síntesis genérica (muy flexible) (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 conversores de texto en habla de dominio restringido (CTH-DR), así como la presente propuesta de conversión de texto en habla multidominio (CTH-MD) (ver figura 3.1), que busca aumentar la flexibilidad de la síntesis a partir de la generalización de los resultados obtenidos mediante la síntesis de carácter restringido.

A continuación se presenta una visión general de las disciplinas implicadas en la nueva propuesta de CTH-MD. Primero, se resumen las características principales de otros sistemas orales multidominio, haciendo hincapié en el bloque de reconocimiento automático del habla multidominio que la mayoría de ellos contiene. Seguidamente, se describe la nueva estrategia de CTH-MD propuesta, en relación a las dos filosofías típicamente utilizadas en el diseño de sistemas de CTH: la síntesis de propósito general y la restringida a dominio, junto a las características de sus corpus de voz y las distintas tipologías de corpus multidominio descritas hasta el momento. A continuación, para abordar el problema de la selección automática de dominio dentro del corpus multidominio utilizado, se describe el módulo de clasificación

[1]Incluso se puede trabajar con voz pregrabada, siempre que la aplicación lo permita (p.ej. información en trenes, metro,...).

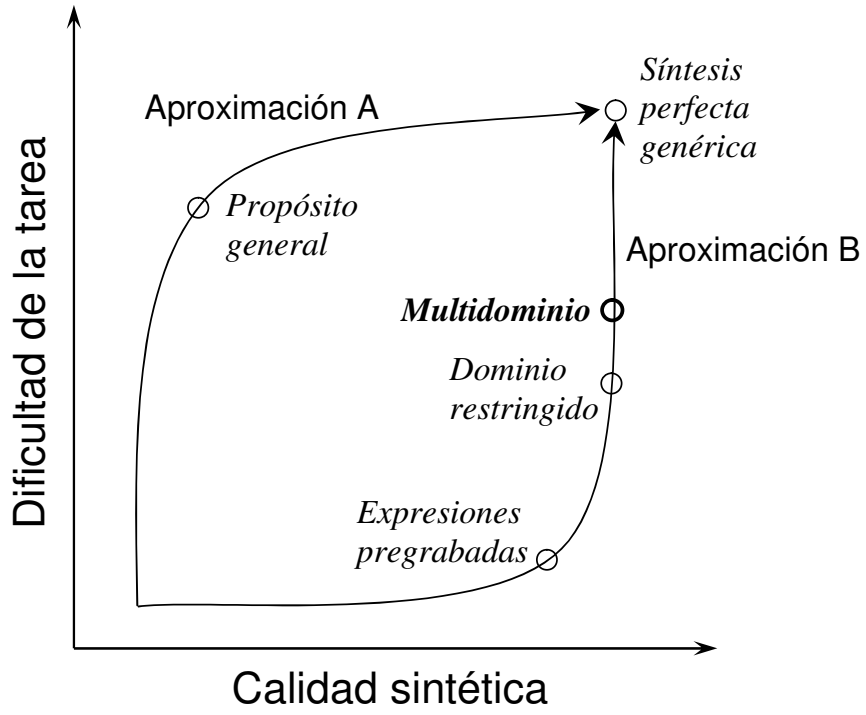


Figura 3.1: Aproximaciones seguidas en la investigación hacia la consecución de una CTH genérica perfecta, representadas según el binomio *calidad sintética - dificultad de la tarea* (o flexibilidad) — figura adaptada de (Taylor, 2000).

de dominios desarrollado. Para ello, primero, se presenta una breve introducción al campo de la clasificación de textos, que permite contextualizar la propuesta respecto a las técnicas de clasificación utilizadas generalmente en ese ámbito de la investigación. Acto seguido se presenta el algoritmo de clasificación diseñado, haciendo énfasis en que su objetivo no es tanto el de ser un excelente categorizador de textos, sino el de ser una herramienta más que permita aumentar la flexibilidad del proceso de síntesis, para así poder disponer de un sistema de conversión de texto en habla multidominio de calidad equivalente a la de los sistemas de dominio restringido actuales. Posteriormente, se presentan los experimentos realizados que permiten validar la viabilidad de la propuesta de CTH-MD presentada, analizando detalladamente las prestaciones del módulo de clasificación automático de dominio desarrollado, elemento clave de la arquitectura multidominio propuesta, así como la calidad sintética obtenida del sistema de CTH-MD basado en selección de unidades implementado. Finalmente, se presentan diversas reflexiones sobre los resultados y las decisiones tomadas a lo largo de la implementación de la propuesta de CTH-MD, discutiendo también sobre posibles líneas de trabajo que quedan abiertas a tenor de los resultados obtenidos del presente trabajo de investigación.

3.1.1. Sistemas orales multidominio

El desarrollo de sistemas multidominio —introducida en (Alías, Iriondo y Barnola, 2003) para CTH—, es una de las nuevas líneas de investigación en el ámbito de los sistemas de lenguaje hablado (SLH) (o *spoken language systems*, en inglés) (Rüggenmann y Gurevych, 2004), entre los que destacan los sistemas de diálogo, los de traducción del habla y los de enrutamiento de llamadas (Lane et al., 2005). 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 (Lane et al., 2004), 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 (Lane et al., 2004). 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) (Rüggenmann y Gurevych, 2004).

Sistemas de diálogo

Los sistemas de diálogo permiten proporcionar o gestionar información al usuario de forma automática mediante voz (p.ej. información meteorológica, compra de productos, reservas, etc.). La idea de incorporar distintos dominios en una misma aplicación propuesta por la CTH-MD (Alías, Iriondo y Barnola, 2003), también ha sido abordada por los sistemas de diálogo en su viaje hacia los futuros sistemas de propósito general (como p.ej. el descrito en (Galibert, Illouz y Rosset, 2005)). Estos sistemas han pasado de trabajar sobre un único dominio —p.ej. Wheels (Yi y Glass, 1998), Jupiter e ILEX (Taylor y Black, 1999; Taylor, 2000)—, a disponer de múltiples dominios —p.ej. GALAXY (Seneff et al., 1998), DARPA Communicator (Rudnicky et al., 2000), NoVo (López-Cózar et al., 2000), SmartKom (Wahlster, Reithinger y Blocher, 2001; Portele et al., 2003), Butler (Pakucs, 2004), o el proyecto EDECÁN[2], entre otros. De este modo, se ha conseguido aumentar la flexibilidad del sistema de diálogo, ya que el usuario puede hacer varias gestiones sobre tareas distintas dentro de una misma consulta. En los primeros sistemas de diálogo multidominio, la elección del dominio se dejaba a iniciativa del usuario (indicación explícita de cambio de dominio) (p.ej. la elección del tipo de noticias en el sistema NoVo (López-Cózar et al., 2000)). Sin embargo, recientemente han aparecido trabajos que pretenden, a partir de la interacción persona-máquina, descubrir el dominio de la consulta de forma automática, mejorando así la usabilidad del sistema (Pakucs, 2003; Lane et al., 2005) y, por lo tanto, la satisfacción del usuario. De este modo, por un lado, se puede adaptar el esquema del diálogo a la solicitud del usuario en tiempo de ejecución (Asami, Takezawa y Kikui, 2002), al determinar de forma implícita el dominio de la consulta, y por otro, se pueden controlar las consultas a

[2] *http://www.edecan.es/* (TIN2005-08660-C04).

las que el sistema no es capaz de dar respuesta, ya que pertenecen a peticiones que quedan fuera de los dominios para los que se ha diseñado el sistema de diálogo (Lane et al., 2004; Pérez-Piñar y García, 2005).

Enrutamiento de llamadas

El enrutamiento de llamadas (o *call-routing* en inglés) es otra de las aplicaciones que pueden ser catalogadas como multidominio (p.ej. el sistema *How may I help you?* de AT&T (Gorin, Riccardi y Wright, 1997)). En este contexto, resulta necesario clasificar las llamadas de los usuarios para dirigirlos a los departamentos o las personas que puedan dar respuesta a su consulta (Haffner, Tur y Wright, 2003), controlando las consultas que no tengan una destinación clara (Lane et al., 2004) o adaptando la derivación de la llamada según la importancia (peso) del tipo de llamada (Tur, 2004). En los trabajos de Myers et al. (2000) y Schapire y Singer (2000) se ataca la clasificación automática de los mensajes orales en el ámbito del enrutamiento de llamadas mediante la combinación de técnicas de clasificación de textos y de reconocimiento del habla (Sebastiani, 2002; Sebastiani, 2005); así como en el trabajo de Tur (2004), donde se indica que la longitud media del texto es de unas 11 palabras por consulta, clasificándolas mediante el modelado del texto con *n-gramas*[3], por poner algunos ejemplos.

Sistemas de traducción oral

Los sistemas de traducción automática del lenguaje oral (*speech-to-speech* o *speech translation systems*, en inglés) buscan convertir el mensaje verbal en un idioma de partida en su correspondiente señal sonora en el idioma de destino. Esta línea de investigación constituye una de las áreas de trabajo más complejas en el ámbito de las tecnologías del habla, debido a la convivencia en un mismo sistema de distintas tareas (reconocimiento, síntesis y traducción automática) que en sí mismas son grandes áreas de investigación y que, a día de hoy, todavía no están completamente resueltas. En este contexto, conocer la temática del mensaje ayuda a mejorar la precisión de la traducción (Asami, Takezawa y Kikui, 2002), tanto por el aumento en la eficiencia del sistema de reconocimiento como por la mejor precisión del bloque de comprensión (p.ej. ayuda a desambiguar la polisemia) (Nakata et al., 2002), así como al controlar las traducciones sobre temas para los que el sistema no tiene buena cobertura (Lane et al., 2004).

Sistemas de recuperación de información

Existen aplicaciones en el ámbito de la recuperación de información (RI) que incorporan la gestión de información oral. Estos sistemas suelen incorporar un módulo de reconoci-

[3]Un *n-grama* es un modelo de caracterización estadística de la probabilidad de aparición de una secuencia de n ítems (caracteres, palabras,...) consecutivos. En el ámbito de las tecnologías del habla, los *n-gramas* se suelen utilizar para modelado del lenguaje y el procesamiento del lenguaje natural, con aplicación al reconocimiento automático del habla, la detección de idioma, etc.

miento automático del habla (RAH) como medio para acceder a datos orales (locuciones, entrevistas, etc.), o bien, para realizar consultas mediante voz (RI con interfaz oral). En este contexto, la interacción RI-RAH da lugar a dos categorías de investigación diferentes (Itou, Fujii y Ishikawa, 2001), que conviene mencionar: *i)* la recuperación de documentos orales, donde, mediante consultas escritas, se pueden consultar documentos orales (p.ej. noticias de televisión o radio) o recuperar algún pasaje relevante (p.ej. la intervención de un determinado ministro en un debate parlamentario), y *ii)* la recuperación de información textual (documentos escritos) mediante consultas orales. La categoría correspondiente a la recuperación de documentos orales ha sido estudiada en las competiciones TREC-SDR (*Text Retrieval Conference - Spoken Document Retrieval*), así como en las competiciones de detección y seguimiento de temáticas (*Topic Detection and Tracking*, TDT) (ver (Allan, 2001) para más detalles). Después de la competición TREC-9, se concluyó que la SDR es una tarea *prácticamente* solventada, dados los resultados obtenidos (Allan, 2001). Esto es debido a que los errores de reconocimiento de los RAH actuales prácticamente no afectan a la tarea de clasificación *temática*, ya que, por un lado, el volumen de palabras tratado es enorme, y por otro, el objetivo es sólo determinar el tema del texto, por lo que no es necesario conocer con exactitud todas las palabras del documento oral (utilizadas como datos de acceso a su contenido).

Por otro lado, la detección y seguimiento de temáticas es una iniciativa de la asociación DARPA (*Defense Advanced Research Projects Agency*, *http://www.darpa.mil*) que investiga sobre la segmentación, el seguimiento y la detección de noticias (en radio, televisión, etc.), agrupando aquellas que discuten un mismo tema, identificando la aparición de nuevas noticias y determinando su duración en el tiempo (Allan, 2001). Aunque el análisis de documentos orales como fuente de información ha sido ampliamente tratado, existe un número relativamente reducido de trabajos dentro del ámbito de la investigación en RI dedicado a la recuperación de datos a partir de consultas orales (Itou, Fujii y Ishikawa, 2001), cuestión que parece más cercana al enfoque seguido por la comunidad científica de las tecnologías del habla. Mientras la comunidad de RI pone el acento en la gestión y recuperación de información, preocupándose, fundamentalmente de analizar el impacto de los errores del RAH en el proceso de recuperación de información (Allan, 2001), la investigación en el ámbito de las tecnologías del habla está más preocupada por mejorar los RAH como medio de acceso a la información.

3.1.2. Reconocimiento del habla multidominio

En cualquiera de los citados SLH, la interacción del usuario con el sistema se realiza mediante voz, por lo que el módulo de reconocimiento automático del habla (RAH) que incorporan los SLH juega un papel fundamental. Los sistemas de RAH han evolucionado desde el reconocimiento de palabras aisladas pertenecientes a un vocabulario reducido y adaptadas a un único locutor (p.ej. dígitos), hasta los sistemas de reconocimiento de habla continua sobre grandes vocabularios e independientes de locutor (Taylor, 2000). Los RAH se sustentan, fundamentalmente, sobre dos elementos clave: el modelo acústico y el modelo lingüístico. En el camino de evolución seguido por los sistemas de RAH, ambos

modelos se han ido transformando, pasando de utilizar estrategias muy controladas (p.ej. reglas gramáticas o modelos acústicos definidos explícitamente para la tarea y el locutor) a aplicar técnicas más flexibles (p.ej. modelos acústicos multilocutor o modelos de lenguaje estocásticos) para poder dar respuesta a las nuevas necesidades planteadas. Generalizar el funcionamiento de los RAH implica aumentar la dificultad de la tarea al incrementar el número de usuarios y el tamaño del vocabulario. Uno de los problemas más críticos que esta cuestión provoca es el desajuste que se puede producir entre las características de los datos —acústicos y lingüísticos— usados para el entrenamiento del sistema respecto a las propiedades de los datos introducidos durante su explotación. Estas discordancias pueden deberse a la evolución del dominio (p.ej. cambios en las palabras más habituales), diferencias entre el estilo del hablante (p.ej. según su personalidad o estado emocional), etc. (ver (Bellegarda, 2004) para más información). Esta cuestión puede atacarse mediante el desarrollo y aplicación de técnicas de adaptación de los modelos a la tarea y/o al usuario —tanto lingüísticos (p.ej. (Hori, Willett y Minami, 2003; Diéguez, García y Cardenal, 2005)) como acústicos (p.ej. (Aiikita y Kawahara, 2004) (ver también cap. [9.6, 9.7, 11.5] de (Huang, Acero y Hon, 2001) y (Bellegarda, 2004) para una visión general de las técnicas de adaptación comentadas).

Arquitectura del RAH multidominio

Según Hazen, Hetherington y Park (2001), a mediados de los 90, las primeras versiones de los sistemas orales multidominio utilizaban un único RAH para todos los dominios, realizando cambios explícitos entre dominios. Sin embargo, estos RAH multidominio (RAH-MD) no eran lo suficientemente robustos como para ser integrados en aplicaciones reales. Posteriormente, se utilizaron RAH construidos para cada dominio siguiendo una arquitectura de reconocedores en paralelo, tomando en consideración la hipótesis de mayor confianza de entre los distintos RAH de dominio. Según los mismos autores existen dos aproximaciones fundamentales para la construcción de un RAH-MD a partir de la combinación de distintos RAH dependientes de dominio, en lo que se refiere al proceso de búsqueda de la mejor hipótesis de reconocimiento: *i*) se puede trabajar en paralelo con distintos RAH adaptados a dominio, seleccionando como resultado del reconocimiento aquel que aporte un mayor grado de confianza (como p.ej. en (Pérez-Piñar y García, 2005)), o *ii*) se puede 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[4]. 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

[4]En la sección 3.2.2 se describen las dos estrategias utilizadas para el diseño de corpus de voz multidominio, estrategias que son equivalentes a las descritas para el diseño de los RAH multidominio.

de la arquitectura paralela, destacando el compromiso existente entre la reducción del coste computacional de la búsqueda y el aumento de la perplejidad[5]en la resolución de la tarea del reconocimiento (ver (Hazen, Hetherington y Park, 2001) para más detalles).

Además de lo expuesto por Hazen, Hetherington y Park (2001), existen otros trabajos que presentan nuevas estrategias orientadas a la definición de la arquitectura de un RAH-MD. Entre ellas, destaca la que divide 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 (Chung, Seneff y Hetherington, 1999; Chung, 2001), donde esta estrategia se emplea para sistemas de comprensión multidominio). En este contexto parece necesario incorporar un módulo de detección de dominio —más concretamente, un módulo detector de temática— para poder escoger, a partir del resultado del RAH entrenado sobre todos los dominios considerados, el RAH dependiente de dominio más adecuado a la expresión (consulta) del usuario (Lane et al., 2005).

Designación de dominio a partir de voz

En el contexto de los sistemas multidominio guiados por voz, el módulo de detección y asignación automática de dominio a partir de las locuciones del usuario toma un papel claramente relevante (Rüggenmann y Gurevych, 2004). Este módulo, por un lado, permite al usuario cambiar de dominio —dentro de los dominios contemplados por la aplicación multidominio— de forma eficiente (Pakucs, 2003), y por otro, permite mejorar el resultado de la recuperación de información al adaptarlo al dominio de la consulta (Allan, 2001). Típicamente, en el ámbito de los SLH el concepto de *dominio* suele ser sinónimo de *temática*, cuestión relacionada con cada una de las tareas que el sistema puede resolver (p.ej. cine, restaurantes, viajes, transacciones, etc.). 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 (*Word Error Rate* -WER- en inglés) (Aiikita y Kawahara, 2004; Lane et al., 2005; Diéguez, García y Cardenal, 2005; Sethy, Georgiou y Narayanan, 2005).

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 (Allan, 2001)), en el contexto de interacción persona-máquina, normalmente se trabaja con un volumen mucho más reducido de datos (Lane et al., 2005) —p.ej. entre 10 y 20 palabras para reservas de viajes (Asami, Takezawa y Kikui, 2002). Bajo este punto de vista, la detección de temática a partir de las expresiones del usuario se convierte en una tarea más complicada que la clasificación de artículos o noticias (Asami, Takezawa y Kikui, 2002).

[5]La perplejidad es una medida relacionada con la entropía de los datos, en este caso, del modelo del lenguaje utilizado, dando información sobre la incertidumbre de un *ítem* dentro del modelo. Así, la perplejidad de una cierta palabra está relacionada con el número potencial de palabras que pueden acompañar a esa palabra, dada la historia de palabras que la precede.

La detección automática del dominio del mensaje en el ámbito de los SLH, se ha abordado desde dos estrategias distintas, basadas en una selección de dominio explícita o implícita:

- **Explícita:** mediante un conjunto predefinido de comandos (palabras clave) el usuario puede navegar mediante voz a través de los servicios (dominios) del sistema de información. Por lo tanto, cuando el usuario quiere cambiar de dominio debe indicarlo al sistema de forma explícita, es decir, diciendo la palabra clave indicada. En este contexto, *sólo* será necesario que el sistema de reconocimiento detecte estas palabras dentro del mensaje del usuario para cambiar de dominio, cuestión que simplifica la tarea del módulo de clasificación de dominio. No obstante, la usabilidad de estos sistemas es reducida, ya que es necesario que el usuario conozca de antemano qué palabras debe utilizar así como el esquema de navegación del sistema (Lane et al., 2005).

- **Implícita:** la detección de dominio se realiza a partir de las hipótesis de reconocimiento recogidas durante la interacción del usuario con el sistema (Hazen, Hetherington y Park, 2001; Asami, Takezawa y Kikui, 2002; Rüggenmann y Gurevych, 2004; Lane et al., 2005). Éste no está obligado a utilizar unas palabras u otras, sino que el sistema procura detectar la temática a partir de pequeños fragmentos de sus expresiones. Esta estrategia es muy útil, por ejemplo, para los sistemas de diálogo, donde la detección implícita de la temática permite reducir el número de turnos del diálogo respecto a los sistemas en los que el usuario tiene que indicar explícitamente el dominio de consulta, mejorando su eficacia y usabilidad (Asami, Takezawa y Kikui, 2002; Lane et al., 2005), o bien, para los sistemas de RI mediante voz (p.ej. RI telefónica, sistemas de navegación en vehículos o interfaces persona-máquina), donde es complicado definir un conjunto finito de palabras clave para recuperar una información u otra (Itou, Fujii y Ishikawa, 2001).

3.2. Arquitectura del sistema de CTH multidominio

La investigación en el ámbito de la CTH no ha incorporado, hasta el momento, la filosofía multidominio seguida por otros sistemas orales de interacción persona-máquina, en su camino hacia mejorar la naturalidad y la usabilidad de los mismos. Como se ha comentado, 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.

En el presente trabajo de investigación se aborda la síntesis multidominio mediante una propuesta que permite la implantación de esta filosofía en el ámbito de la CTH. Para ello,

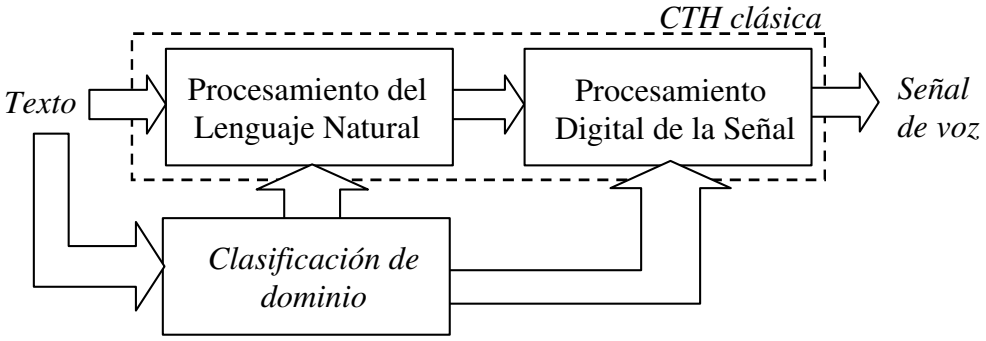


Figura 3.2: Diagrama de bloques de la arquitectura de un conversor de texto en habla multidominio con clasificación automática de dominio.

se define una nueva arquitectura que incorpora un módulo de clasificación de dominio a la arquitectura clásica de los sistemas de CTH (ver figura 3.2). El hecho de poder determinar el dominio del texto de entrada permite mejorar la manera de sintetizarlo: *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)* escogiendo el modelo prosódico más adecuado de los disponibles (si el módulo de predicción prosódica dispone de distintos modelos de locución, uno por dominio) o considerando varios perfiles prosódicos en la búsqueda (p.ej. como en (Campillo y Rodríguez Banga, 2002)); *iii)* eligiendo el subcorpus más adecuado para la síntesis (si existe más de uno en el corpus); *iv)* 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. (Hamza et al., 2004)); *v)* 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 (Turk et al., 2005)); *vi)* 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 u otro en el funcionamiento del sistema de síntesis. Por otro lado, cabe añadir que la propuesta de CTH-MD definida permite la coexistencia de distintas tipologías de corpus de voz y estrategias de síntesis, gracias al diseño de una arquitectura flexible y adaptable a las necesidades de cada aplicación.

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-LD).

A continuación, se procede a ubicar la CTH-MD respecto a las estrategias de CTH de propósito general y de dominio restringido, describiendo los elementos más destacables del diseño y el dominio de sus corpus de voz, así como las aproximaciones que, hasta el momento, se han abordado en la construcción de corpus multidominio. Para terminar esta sección, se presenta el primer paso seguido en este trabajo de investigación para llevar a la práctica la filosofía de CTH multidominio introducida, en este caso, según la estrategia de síntesis basada en corpus o *selección de unidades*.

3.2.1. Estrategias de conversión de texto en habla

Seguidamente, se presenta un pequeño resumen de los elementos clave de las estrategias de conversión de texto en habla, basadas en síntesis concatenativa, con las que se ha abordado hasta el momento la generación de la señal de voz a partir de un texto de entrada —CTH de propósito general y de dominio restringido—, desde el punto de vista de su relación con la CTH-MD, ya que esta busca alcanzar una calidad sintética próxima a la de los sistemas de dominio restringido, pero intentando aproximarse a la flexibilidad de la síntesis de propósito general.

Conversión de texto en habla de propósito general (CTH-PG)

Los sistemas de propósito general —también denominados sistemas de dominio ilimitado o abierto (van Santen y Buchsbaum, 1997; Zhu et al., 2002; Schweitzer et al., 2003)— permiten convertir *cualquier* texto de entrada, en el idioma de trabajo, en su correspondiente mensaje oral, primando la generalidad ante la calidad de la síntesis (inteligibilidad *vs.* naturalidad). Como ya se ha comentado en la sección 1.1, existen distintas técnicas para la consecución de este objetivo. En el contexto de los sistemas de síntesis concatenativa (estrategia ampliamente utilizada en la actualidad), el corpus de voz juega un papel muy importante. En el caso concreto de los sistemas basados en selección de unidades, el corpus se diseña con el objetivo de alcanzar la máxima cobertura fonética, prosódica y lingüística del idioma considerado (Black, 2002). En este contexto, la calidad de la síntesis de voz está fuertemente ligada al *buen* diseño del corpus (Black, 2002; Zhu et al., 2002); cuestión, por otro lado, nada sencilla y que todavía es fuente de investigación (ver sección 2.1.2).

Conversión de texto en habla en dominios restringidos (CTH-DR)

Los sistemas de dominio restringido anteponen la calidad a la generalidad de la síntesis (Yi y Glass, 1998; Taylor, 2000), es decir, persiguen maximizar la naturalidad sintética

de los mensajes generados dentro del marco (dominio) de la aplicación desarrollada. En el ámbito de la síntesis basada en selección de unidades, los CTH-PG, a pesar de obtener a menudo una naturalidad sintética bastante buena, todavía presentan expresiones sintéticas de baja calidad (Breuer y Abresch, 2004; Toda y Tokuda, 2005). Por lo tanto, el usuario percibirá saltos en la calidad sintética del habla, a pesar de presentar una elevada calidad en la mayor parte del mensaje generado. Esta problemática se debe, fundamentalmente, a que no siempre se dispone de las unidades *deseadas* en el corpus, por lo que se seleccionan unidades *próximas*[6] (ver capítulo 2). Una vez recuperadas estas unidades, o bien son concatenadas directamente (p.ej. ver (Black y Taylor, 1994)), o bien son modificadas mediante técnicas de procesado de la señal (p.ej. ver (Breen y Jackson, 1998)). El primer proceso puede provocar discontinuidades en la señal sintética y, el segundo, la aparición de *artefactos* sonoros (Toda y Tokuda, 2005). Una de las estrategias para la minimización de este tipo de problemas consiste en diseñar el corpus según el ámbito (dominio) sobre el que está previsto que el sistema de CTH se aplique, minimizando el impacto de la discordancia entre los contenidos del corpus y los del texto a sintetizar. Bajo este enfoque, se desarrollan los sistemas de conversión de texto en habla para dominios restringidos (CTH-DR). Su objetivo será, pues, resolver óptimamente la síntesis en el ámbito de aplicación considerado. El concepto de ámbito restringido, puede ser muy variado: desde dominios muy cerrados (p.ej. un reloj parlante (Black y Lenzo, 2000)) hasta dominios bastante flexibles (p.ej. sistemas de diálogo (Wahlster, Reithinger y Blocher, 2001; Raux et al., 2005)), pero sin llegar nunca a ser de propósito general. También es importante indicar que existen aplicaciones donde el concepto de dominio *restringido* no va ligado al de dominio *estático* (Yi y Glass, 1998; Taylor y Black, 1999; Black y Lenzo, 2000), es decir, se trata de aplicaciones que deben adaptarse a un entorno de información cambiante.

Según las características del ámbito de aplicación del sistema, existen diferentes aproximaciones al problema de la CTH para dominios restringidos, que permiten abordarlo mediante distintas estrategias de síntesis:

- **Concatenación de voz pregrabada:** se utilizan expresiones pregrabadas (*prompts*, en inglés) en lugar de incorporar un CTH propiamente dicho (Yi y Glass, 1998; Chu et al., 2002) (ver figura 3.1), ya que su calidad no es suficiente para las necesidades de la aplicación considerada (p.ej. información en los trenes, metro, etc.). Es una aproximación de *síntesis* muy básica (con mínima flexibilidad), donde se utilizan plantillas que son rellenadas con expresiones pregrabadas, como por ejemplo, nombres, precios, números, etc. encajadas dentro de frases portadoras (Black y Lenzo, 2000; Montero et al., 2000; Taylor, 2000) (p.ej. "*Próxima estación*"+"*Tordera*").

- **Concatenación de palabras:** como su nombre indica, la síntesis de la señal de voz se basa en la concatenación de las palabras grabadas previamente en el corpus. Existen distintas aproximaciones que aprovechan el vocabulario limitado del dominio restringido para trabajar con unidades de síntesis de tamaño superior (grupos de

[6]La situación ideal correspondería a la recuperación completa del mensaje correspondiente al texto de entrada, sin tener que modificar la señal, en la línea de lo descrito por Balestri et al. (1999).

sílabas, palabras, etc.) a las utilizadas para la síntesis de propósito general (fonemas, difonemas, etc.) (Möbius, 2000). Uno de los ejemplos más representativos de este enfoque es el proyecto Verbmobil para la planificación de viajes (Stöber et al., 1999). Los principales problemas de estos sistemas radican en los saltos de naturalidad de la señal sintética obtenida (sobretodo en los puntos de concatenación de las unidades (Yi y Glass, 1998)), a pesar de haberlas grabado con una gran variedad de contextos prosódicos y lingüísticos; así como cuando resulta necesario sintetizar palabras fuera del vocabulario (Möbius, 2001). Este problema, a menudo, se aborda mediante la síntesis basada en unidades de tamaño menor a la palabra (sílabas, fonemas, difonemas, etc.) para poder disponer de la señal sintética correspondiente. No obstante, esto provoca un empeoramiento sustancial de la calidad global obtenida respecto a los resultados generados a nivel de concatenación de palabras.

- **Concatenación de locuciones o *phrase-splicing*:** consiste en utilizar locuciones pregrabadas en un corpus de voz que son empalmadas durante el proceso de síntesis para generar el mensaje oral correspondiente. Esta unión se puede dar a nivel de locuciones completas o a partir de combinaciones de fragmentos de las mismas, rellenando los huecos que puedan aparecer mediante unidades de menor tamaño (Yi y Glass, 1998; Donovan et al., 1999; Donovan et al., 2001; Cosi et al., 2001; Hamza y Pitrelli, 2005). Evidentemente, la naturalidad obtenida es muy elevada, siempre y cuando la frase a sintetizar se ajuste al grupo de locuciones pregrabadas, decreciendo ésta a medida que el texto a sintetizar se aleje del dominio del corpus (calidad *in-script* > calidad *in-domain* > calidad *out-of-domain* en (Hamza y Pitrelli, 2005)). El sistema de CTH suele seleccionar los segmentos más largos posibles como base de la generación de la señal sintética, añadiendo las unidades ausentes mediante el proceso clásico de CTH que funciona como núcleo del sistema. En este contexto, nuestro grupo de investigación ha desarrollado, recientemente, un sistema de CTH basado en corpus para una aplicación meteorológica que se basa en esta estrategia de síntesis (Alías et al., 2005) (ver anexo D.1).

- **Selección de unidades:** consiste en aprovechar la filosofía de los sistemas de CTH basados en grandes corpus de voz ajustándola a una determinada tarea donde el vocabulario es restringido aunque no limitado del todo (Schweitzer et al., 2003). Estas tareas pueden ir desde aplicaciones muy sencillas, como p.ej. un reloj parlante (Black y Lenzo, 2000), pasando por aplicaciones audiovisuales con vocabularios controlados (Johnson et al., 2002), hasta sistemas de diálogo mucho más complejos, como: el *CMU DARPA Communicator* (Rudnicky et al., 2000), que permite planificar viajes, vuelos, hacer reservas de hoteles y alquilar coches telefónicamente; el sistema SmartKom, que puede manejar consultas para varios dominios (cine, TV, información turística, etc.) mediante una interfaz multimodal (Wahlster, Reithinger y Blocher, 2001); o más recientemente, el sistema *Lets Go Public!* (Raux et al., 2005), para consultas sobre el servicio de autobuses de una ciudad. En estos sistemas de diálogo, el bloque de síntesis sigue la filosofía de la selección de unidades, recuperando del corpus los segmentos más adecuados a las características de la locución a sintetizar (dentro de un contexto controlado). A pesar de la mejora de calidad, esta estrategia todavía sufre

el efecto de las palabras ajenas al dominio (*out-of-vocabulary* -OOV- en inglés). Este problema suele resolverse, o bien, mediante la síntesis basada en fonemas o difonemas del mensaje que corresponde a los vocablos de fuera del vocabulario (Black y Lenzo, 2000; Rudnicky et al., 2000) —manteniendo una calidad uniforme, aunque muy inferior a la media—, o bien, mediante una aproximación híbrida, donde conviven unidades genéricas y adaptadas a dominio, que se seleccionan mediante estrategias de búsqueda ajustadas al problema (Schweitzer et al., 2003).

3.2.2. Corpus de voz multidominio

En el ámbito de los sistemas de síntesis concatenativa, el corpus de voz es uno de los elementos clave, ya que influye decisivamente en la calidad obtenida del habla sintética generada, sobretodo en el caso de los CTH basados en selección de unidades. Es por ello que, normalmente, se encarga a un locutor profesional la grabación del contenido del corpus de voz para garantizar la calidad de la señal almacenada (sonoridad, vocalización, etc.), así como para conseguir que el estilo de locución sea lo más *consistente* posible a lo largo de las distintas sesiones de grabación realizadas, sobretodo en el caso de corpus de tamaño considerable (Kawai et al., 2004) —gracias a la capacidad que tienen estos profesionales de repetir con gran exactitud el estilo utilizado en las sesiones previas. Por otro lado, el diseño del corpus de voz estará estrechamente ligado al tipo de aplicación en la que trabaje el sistema de conversión de texto en habla (Zhu et al., 2002), diferenciándose claramente los corpus genéricos de los corpus (más o menos) restringidos, como se pasa a describir a continuación.

Diseño del corpus

El corpus de voz para síntesis de propósito general se suele construir a partir de textos de contenido genérico que permitan la síntesis de *cualquier* texto de entrada. Una de las fuentes de datos más utilizadas en el diseño de los corpus de voz genéricos son los textos periodísticos. Existen numerosos ejemplos de sistemas de conversión de texto en habla para selección de unidades (CTH-SU) basados en este tipo de textos. Entre ellos se encuentran: una de las primeras versiones del CTH-SU de IBM (Donovan et al., 2001), el sistema de la Universitat Politècnica de Catalunya (Febrer, 2001) —construido a partir de artículos de *El Periódico de Cataluña* y el diario *Avui*—, el utilizado en el CTH de Microsoft Asia (Chu et al., 2001) —basado en la recopilación de cinco años del periódico *People's Daily*—, o la porción genérica del corpus de voz para el sistema de diálogo SmartKom (Schweitzer et al., 2003), entre otros. Por otra parte, en la mayoría de diseños de corpus para síntesis genérica, estos textos se completan con otras fuentes de información. En este contexto, se encuentra el sistema de CTH de AT&T (Beutnagel, Conkie y Syrdal, 1998; Conkie, 1999; Beutnagel y Conkie, 1999), formado por textos del *Wall Street Journal* más una recopilación de conversaciones de un servicio automático de atención telefónica; la nueva versión del corpus de IBM, que contiene, además de artículos periodísticos, correos electrónicos, previsiones del tiempo junto a textos relacionados con aplicaciones guiadas por voz (Fischer, Botella y

Kunzmann, 2004); el sistema XIMERA de ATR, construido sobre un corpus formado por textos extraídos de noticias, novelas y conversaciones de viaje (Kawai et al., 2004); o el primer corpus para un CTH-SU en portugués, formado por artículos periodísticos enriquecidos con textos procedentes de una entrevista, junto a un grupo de frases interrogativas (Teixeira et al., 2001), por citar algunos. Asimismo, cabe añadir que en todos estos corpus se asegura la cobertura de las unidades mínimas que describen el idioma mediante la incorporación de frases fonéticamente balanceadas, palabras aisladas, o logatomos (pseudopalabras, generalmente sin significado). Finalmente, añadir que, recientemente, se ha desarrollado un corpus de libre distribución denominado CMU ARCTIC a partir de textos de novelas pertenecientes al proyecto Gutenberg (Kominek y Black, 2003; Kominek y Black, 2004), con el objetivo de disponer de textos sin restricciones de derechos de autor. Concretamente, se han seleccionado frases fonéticamente balanceadas para surtir de contenidos al corpus de propósito general desarrollado (Kominek y Black, 2003; Kominek y Black, 2004),

Por otro lado, en (Montero et al., 2000; Black, 2003) se comenta que el diseño de un corpus de dominio restringido se puede llevar a cabo siguiendo los mismos criterios de cobertura utilizados para los grandes corpus de voz de los sistemas de CTH-PG pero restringiendo su cobertura al dominio objetivo. Según este enfoque, si se dispone de un corpus con una buena cobertura para el dominio objetivo, éste podrá ser sintetizado con una calidad sintética muy buena (Black, 2003).

Estilo del corpus

El estilo de locución escogido durante el proceso de grabación tiene que ser consistente con la aplicación o tarea para la que se desarrolla el CTH al que pertenece el corpus (Zhu et al., 2002), ya que el habla sintética reflejará el estilo y la cobertura del corpus grabado, principalmente, en el contexto de los sistemas de CTH-SU (Breen y Jackson, 1998; Guaus y Iriondo, 2000a; Chu et al., 2001; Black, 2002; Kominek y Black, 2003; Toda, 2003). Para el caso de un sistema de dominio restringido, el ámbito de la aplicación fijará el estilo más adecuado de grabación (p.ej. (Black y Lenzo, 2000; Möbius, 2001; Batůšek, 2002; Alías et al., 2005)). Sin embargo, para un corpus de voz de propósito general suele escogerse un estilo de locución *neutro*[7], es decir, se busca que el corpus no tenga *ningún* estilo en particular, para ser coherente con la filosofía de sistema de *transmisión de información* bajo la que se enmarca la CTH-PG (Campbell, 2002). Asimismo, este estilo es el que parece más adecuado al contenido del corpus, ya que, como se acaba de comentar, éste suele estar formado fundamentalmente por textos de contenido genérico. Grabar el corpus con un estilo neutro permite, por un lado, conseguir una buena cobertura prosódica y espectral de las unidades de voz, gracias a su reducida variabilidad (Chu et al., 2001; Zhu et al., 2002), y por otro, minimizar la necesidad de modificar prosódicamente las unidades seleccionadas durante la fase de síntesis para llegar a la secuencia objetivo (Breen y Jackson, 1998; Black, 2003) —incluso, se puede llegar a concatenar directamente las unidades de voz, como ya se ha comentado en el presente trabajo de investigación (ver capítulo 1). Como

[7]En la literatura: *relax reading style* (Chu et al., 2001), *impersonal reporting style* (Campbell, 2002), *news reader style* (Black, 2003) o *normal reading speech style* (Kawai et al., 2004), por ejemplo.

consecuencia, la señal de voz generada sonará con un estilo de locución *neutro*, similar al utilizado para la lectura de noticias (Black, 2002; Black, 2003) —los corpus suelen estar construidos a partir de textos periodísticos, como se acaba de comentar. De todos modos, también existe la posibilidad que, durante la síntesis, se adapte la prosodia de las unidades seleccionadas del corpus a la prosodia (estilo de locución) de la frase a sintetizar mediante un postprocesamiento de la señal grabada (Breen y Jackson, 1998), aunque este proceso puede provocar que se degrade la calidad de la señal sintética obtenida (Black, 2003; Toda y Tokuda, 2005).

Aproximaciones al corpus multidominio

La calidad sintética obtenida por los CTH-SU decrece de forma notable cuando el dominio del que proviene el texto a sintetizar no se ajusta al del corpus grabado (Black, 2003), cuestión que perjudica la calidad tanto de los sistemas de CTH-PG (Chu et al., 2002; Fischer, Botella y Kunzmann, 2004) como la de los CTH-DR (Black y Lenzo, 2000; Hamza y Pitrelli, 2005; Alías et al., 2005). Por ejemplo, si el corpus de voz se ha grabado con un estilo neutro —síntesis de propósito general—, y se introduce un texto de estilo *literario* —poético quizás—, el resultado no será el deseado por el usuario (p.ej. la variación prosódica no será la más adecuada). 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 (Chu et al., 2002; Fischer, Botella y Kunzmann, 2004). 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** (Alías, Iriondo y Barnola, 2003; Alías et al., 2003). 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 (Taylor y Black, 1999; Chu et al., 2002; Meng et al., 2002; Schweitzer et al., 2003; Hamza et al., 2004; Fischer, Botella y Kunzmann, 2004), o *ii)* disponer de tantos subcorpus como dominios (Iida et al., 2000; Johnson et al., 2002; Campbell, 2002; Iida et al., 2003). Estos dos enfoques —o tipologías de corpus multidominio— han sido denominados *blending* y *tiering*, respectivamente, en la literatura de síntesis basada en corpus[8] (Lenzo y Black, 2002; Black, 2002; Black, 2003; Hofer, Richmond y Clark, 2005). Las principales características que los definen son:

- ***Tiering*:** se construyen subcorpus independientes para diferentes tareas o dominios, realizando cambios explícitos de subcorpus en tiempo de síntesis (Black, 2002; Black,

[8]En el ámbito de la síntesis basada en los modelos ocultos de Markov (HMM) estos dos enfoques también han sido utilizados para el modelado de distintos estilos de voz y emociones presentes en un mismo corpus (Yamagishi et al., 2003; Yamagishi et al., 2005). En este caso la denominación ha sido *style dependent modeling* y *style mixed modeling*, conceptos sinónimos a los utilizados en la literatura de CTH-SU.

2003). Esta aproximación toma sentido cuando los dominios están claramente definidos y/o acotados (p.ej. temáticamente: información meteorológica, reservas, deportes, etc. o por calidad vocal: estilo de locución, emoción, etc.). La principal ventaja de esta aproximación radica en el ajuste de cada uno de los subcorpus al dominio considerado, consiguiendo una elevada calidad (equivalente a la de los CTH-DR) para la síntesis de textos pertenecientes a esos dominios. No obstante, su mayor dificultad reside en la necesidad de crear un nuevo subcorpus cada vez que se desee ampliar el número de dominios del corpus, con todo el trabajo que ello implica.

- ***Blending*:** se combinan los distintos dominios dentro de un mismo corpus, típicamente con un mayor contenido de datos para cubrir la síntesis genérica (Black, 2002), con una definición de los límites o fronteras entre los subcorpus menos estricta comparándola con el enfoque *tiering*. Esta estrategia permite la variación gradual entre los distintos estilos de voz contenidos en el corpus (todos ellos, generalmente, con una calidad vocal similar), tal y como se detalla en (Black, 2003; Hofer, Richmond y Clark, 2005). En este contexto, la elección de dominio se realiza implícitamente a partir de las unidades a sintetizar, sin que esto implique un cambio explícito de subcorpus. Por ejemplo, si un texto contiene órdenes, estas frases deberían ser generadas a partir del subcorpus que contiene mensajes de órdenes, mientras que la información general del mismo texto debería ser generada a partir de las unidades neutras del corpus (Black, 2003). La mayor dificultad de esta aproximación reside en el buen diseño del corpus mixto (cobertura, balanceo, etc.) y en la elección adecuada de las unidades del corpus durante el proceso de selección (evitando la unión de unidades heterogéneas).

3.2.3. Implementación de la propuesta

Como se ha comentado en la introducción del presente capítulo, la filosofía de conversión de texto en habla multidominio permite utilizar distintas técnicas y formatos de corpus para conseguir que el CTH pueda generar mensajes sintéticos lo más ajustados posible a las características del texto (estilos de locución, emociones, énfasis, etc.) (ver sección 3.5 para más detalles). No obstante, para poder validar la viabilidad de la propuesta y como primer paso para implementar un sistema de CTH-MD, se ha optado por abordar el problema mediante la estrategia de síntesis bajo la que se enmarca el presente trabajo de investigación: la CTH basada en corpus o selección de unidades. En lo que se refiere al corpus de voz multidominio, éste se ha construido siguiendo la técnica de *tiering* (un subcorpus por dominio) debido a las características de los dominios escogidos (ver sección 3.4.2), por lo que el vocabulario que contiene cada uno de los dominios que constituirán el corpus de voz debe presentar una cobertura fonética y prosódica suficiente para ese dominio (Montero et al., 2000; Black, 2003), minimizando el problema de las palabras fuera de vocabulario (Möbius, 2001), al incorporar las unidades mínimas de síntesis —en este caso, difonemas y trifonemas— del idioma tratado.

La figura 3.3 (Alías, Iriondo y Barnola, 2003; Alías et al., 2003) muestra un ejemplo de la arquitectura del CTH-MD que se propone para CTH-SU basada en *tiering*, en la que se

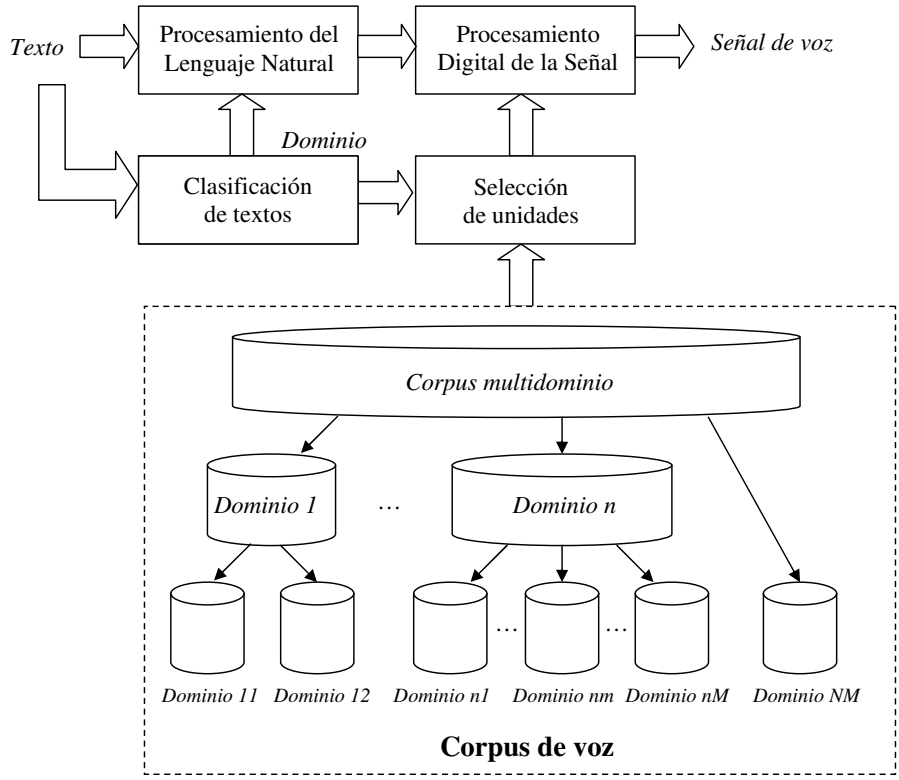


Figura 3.3: Diagrama de bloques de la arquitectura de un conversor de texto en habla multidominio basado en selección de unidades con subcorpus independientes organizados en una jerarquía de tres niveles.

puede observar que el corpus multidominio está dividido en distintas secciones, denominadas en la figura como *dominios*. Asimismo, la figura muestra un ejemplo de posible estructura jerárquica del corpus, donde, a medida que se bajan niveles, el contenido de las secciones se especializa cada vez más. Según la profundidad de la jerarquía en la que se encuentre cada nivel, se tratará de un dominio ($Dominio_n$) o de un subdominio ($Dominio_{nm}$, subdominio *m-ésimo* del dominio *n-ésimo*), y así sucesivamente. Según las particularidades del corpus y del CTH-MD desarrollado, esta estructura presentará configuraciones distintas (mayor o menor profundidad, más o menos dominios, algunos dominios agrupados, otros no, etc.). Por ejemplo, se puede pensar en una arquitectura para un corpus de voz que permita diferentes *emociones* (neutra, triste, alegre, etc.), diferentes *estilos de locución* (periodístico o de lectura de noticias, contador de cuentos, etc.), distintas *temáticas* (política, sociedad, teatro, deportes, ...), etc., estructurando y agrupado los dominios según sus contenidos y las características de la señal de voz que los definen. En este contexto, la dificultad radicará, fundamentalmente, en el diseño del contenido de cada uno de estos niveles de la jerarquía, así como del proceso encargado de escoger la sección del corpus más adecuada para el texto

de entrada con el objetivo de sintetizarlo con la mayor calidad sintética posible.

Aunque, inicialmente, esta arquitectura jerárquica del corpus multidominio se definió sólo pensando en el ámbito de la CTH-MD (Alías, Iriondo y Barnola, 2003; Alías et al., 2003), es decir, para disponer de distintos dominios de síntesis (p.ej. diversos niveles de énfasis de la emotividad del habla: *elevado, medio, tenue, ...*), posteriormente se observó que constituye un marco genérico para el desarrollo de cualquier tipo de sistema de CTH. 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), que pueden ser incorporados explícitamente como subcorpus de dominio (p.ej. (Iida et al., 2003; Black, 2003)), o como pequeños subcorpus de dominio acompañando a un corpus genérico (p.ej. (Chu et al., 2002; Fischer, Botella y Kunzmann, 2004; Hamza et al., 2004; Hofer, Richmond y Clark, 2005)), o dividir un mismo corpus (con la misma calidad vocal) en distintas temáticas (p.ej. distintas temáticas 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.

Como se puede observar también de la figura 3.3, otro de los elementos clave de la propuesta de CTH-MD es el módulo que se encarga de indicar al sistema de CTH el dominio (de la estructura jerárquica) más apropiado a las características del texto de entrada, por lo que se puede seleccionar el subcorpus de voz y el modelo prosódico más adecuados para llevar a cabo la síntesis de ese texto. En el siguiente apartado de este capítulo, se detalla el diseño e implementación de la propuesta desarrollada para la clasificación automática de dominios a partir del texto de entrada en el ámbito de la CTH-MD. Concretamente, se incorpora un módulo de clasificación de textos entrenado a partir de los textos correspondientes a los distintos dominios que constituyen el corpus multidominio.

3.3. Clasificación automática de dominios para CTH-MD

El hecho de incorporar distintos dominios en un mismo corpus de voz no es en sí un problema, sino que la principal dificultad radica en conseguir que el sistema de CTH sea *capaz* de manejarlos adecuadamente; es decir, según la cobertura del dominio, su etiquetado, las características acústicas de la señal de voz, etc. los datos almacenados en cada dominio deberán ser tratados de forma distinta. Si esta variabilidad no es controlada adecuadamente por el CTH, puede provocar una pérdida de la calidad de la señal sintética cuando se concatenen segmentos de voz que provengan de partes del corpus con características distintas (efecto *popurrí* o *patchwork effect* (Breen y Jackson, 1998)) —por ejemplo, si se unen señales

con calidades vocales diferentes (Kawai et al., 2004) o condiciones de grabación (amplitud, estilos de locución,...) no uniformes (Black, 2003). No obstante, algunas de estas diferencias son minimizables mediante el procesamiento de la señal (Breen y Jackson, 1998), aunque esto también puede degradar la señal con la consiguiente pérdida de la calidad sintética (Black, 2003; Toda y Tokuda, 2005). En general, se puede afirmar que cuanto más *alejado* se encuentre el dominio de síntesis del dominio disponible en el corpus, mayor modificación de la señal será necesaria y, en principio, la calidad sintética esperada será menor (dependiendo del sistema de modificación de la señal utilizado, p.ej. TD-PSOLA no permite grandes modificaciones prosódicas de la señal sin que la calidad de esta se degrade notablemente (Peng, Zhao y Chu, 2002; Iriondo et al., 2003)).

3.3.1. Designación de dominio a partir de texto

Así pues, estructurar el corpus de voz en un conjunto de *subcorpus* implica disponer de algún método que indique al CTH el dominio más adecuado sobre el que llevar a cabo el proceso de selección de unidades, es decir, un módulo de clasificación de dominios. Este módulo puede ser externo al CTH —selección manual o supervisada—, o puede formar parte de él —clasificación automática—, como se propone en este trabajo y se describe a continuación. Normalmente, la información referente al dominio suele incorporarse al texto a sintetizar utilizando algún lenguaje de marcas, generalmente tipo XML (p.ej. ver (Alías et al., 2005)), tanto para indicar cual es el estilo de locución más adecuado para una determinada expresión (Johnson et al., 2002), como para indicar la emoción deseada (Hofer, Richmond y Clark, 2005).

Elección manual

En el caso más sencillo, la selección de dominio la realiza el propio usuario de forma manual mediante una interfaz gráfica que permite escoger el estilo de locución deseado (emoción, expresividad,...) entre los disponibles (Iida et al., 2000; Campbell, 2002). Esta opción fue la utilizada en el marco del desarrollo de un locutor virtual[9], en la que se creó una interfaz que permitía escoger entre tres tipos de síntesis (por difonemas) expresiva: alegre, neutra o triste (ver apéndice D.2).

Designación supervisada

La selección del dominio viene indicada por algún módulo o proceso previo a la conversión de texto en habla (p.ej. un sistema de generación de lenguaje natural), guiando la síntesis del mensaje para darle el formato más adecuado (p.ej. el estilo de locución (Yamagishi et al., 2003; Yamagishi et al., 2005)), por lo que la tarea de detección de dominio queda fuera de las atribuciones del CTH. Entre otros sistemas que utilizan esta estrategia, se encuentran algunos sistemas de diálogo —los cuales, en lugar de trabajar con un CTH,

[9]Proyecto financiado por el MCyT (FIT-150500-2002-410).

suelen trabajar con un conversor de concepto en habla (CCH) (*concept-to-speech* o CTS, en inglés) (Taylor, 2000; Wahlster, Reithinger y Blocher, 2001)— o algunas aplicaciones audiovisuales con mensajes sintéticos controlados (Johnson et al., 2002; Alías et al., 2005). En este contexto supervisado, se minimiza la ambigüedad del texto de entrada a sintetizar, ya que éste está enriquecido mediante información que permite guiar mejor la síntesis de voz[10] (Taylor, 2000), gracias al *conocimiento* del contenido y el contexto del mensaje (Sproat, 1997), a diferencia de los CTH que deben estimar toda esta información únicamente a partir del texto de entrada; texto que no siempre ofrece la suficiente información lingüística para obtener una elevada calidad sintética (Sproat, 1997). Sin embargo, con la CTH-MD se pretende profundizar en el análisis del texto de entrada al incorporar un módulo de clasificación automática de dominio para mejorar la naturalidad de la respuesta de estos sistemas (sistemas de diálogo, interfaces persona-máquina, robots,...), cuyo estilo de locución neutro (lectura de noticias) está lejos de satisfacer a los usuarios (Sagisaka, Yamashita y Kokenawa, 2005).

Clasificación automática

Como se ha comentado en el capítulo correspondiente a la introducción del presente trabajo de investigación, un sistema de CTH está formado por dos bloques fundamentales: el bloque de análisis del texto (o bloque de procesamiento del lenguaje natural) y el bloque de generación de la señal oral sintética (o bloque de procesamiento digital de la señal). Típicamente, el bloque de análisis del texto debe resolver las ambigüedades inherentes en el texto escrito para producir una representación lingüística precisa de la frase a sintetizar (Taylor, 2000). Sin embargo, para incorporar el concepto de *dominio* dentro de la arquitectura de un CTH, resulta necesario disponer de un sistema automático de detección y clasificación de dominios para que el sistema de CTH-MD conozca, durante el proceso de conversión de texto en habla, qué dominio es el más adecuado para sintetizar el mensaje requerido de la forma más apropiada posible (Black, 2002). En este contexto es en el que se sitúa el presente trabajo de investigación, donde se aborda esta cuestión mediante un análisis del texto algo más detallado de lo habitual —yendo un poco más allá de las funcionalidades típicas del módulo de procesamiento del lenguaje natural de un CTH clásico (enumeradas en la sección 2.1.2). Concretamente, se aborda el problema de la clasificación de dominio para síntesis mediante técnicas procedentes del ámbito de la clasificación de textos, partiendo, hasta el momento, del único dato disponible a la entrada del CTH: el texto a sintetizar[11].

Bajo este enfoque, a continuación se presenta una breve introducción a la teoría de la clasificación automática de textos, haciendo un pequeño repaso de las aproximaciones o estrategias de clasificación más utilizadas. Posteriormente, se describe el bloque de clasificación de textos que se ha diseñado en el ámbito de la CTH-MD, haciendo énfasis en las

[10] Por ejemplo, en el ámbito de los CCH es posible incluir atributos prosódicos de tipo semántico o pragmático, difíciles de obtener mediante los sistemas de CTH convencionales, gracias al conocimiento del contexto de la comunicación en el que el mensaje se debe emitir (Taylor, 2000; Nakatani y Chu-Carroll, 2000)

[11] En un futuro, se pretende estudiar si incorporar más información del texto, p.ej. la función morfosintáctica de las palabras, puede ayudar a optimizar el funcionamiento del módulo desarrollado.

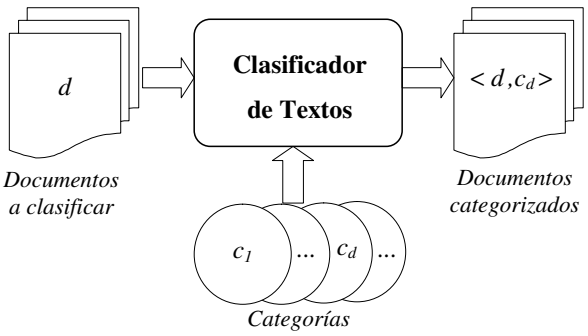


Figura 3.4: Esquema del funcionamiento de un algoritmo de clasificación de textos sobre una colección de documentos a etiquetar mediante el conjunto de categorías predefinido.

particularidades del mismo.

3.3.2. Clasificación automática de textos

La categorización o clasificación de textos (CT) (*Text Classification* en inglés) consiste en la asignación de un texto o documento a una o más categorías, según un proceso manual (realizado por un experto) o automático. Estas categorías pueden indicar, por ejemplo, los temas que conforman el documento analizado, la relevancia del contenido del documento según el perfil del usuario, el género o el autor del texto, entre otras —diferenciándose claramente las aplicaciones de naturaleza *temática* de las *no temáticas* (Sebastiani, 2002; Sebastiani, 2005). Así pues, dada una colección de documentos $\mathcal{D}$ y un conjunto predefinido de categorías $\mathcal{C}$, el algoritmo de CT debe asignar cada documento $d \in \mathcal{D}$ a la categoría (asignación única o *hard classification*), o al grupo de categorías $c_d \supseteq \mathcal{C}$ (asignación múltiple o *soft classification*) que más se adecuan a su contenido (Sebastiani, 2002) (ver figura 3.4).

Conceptualmente, la clasificación de textos automática es una disciplina que surge de la intersección de dos áreas de investigación: la recuperación de información (RI, o *Information Retrieval* en inglés) y el aprendizaje artificial (AA, adaptación de la expresión inglesa *Machine Learning*). Por una parte, en el contexto de los sistemas de RI se define la representación de los datos utilizados por los sistemas de clasificación de textos, y por otra, las técnicas de AA tratan de modelar y generalizar toda la información que interviene en el proceso de clasificación. Para el caso de la clasificación de textos, el proceso de aprendizaje se alimenta de un conjunto de datos de entrenamiento, que, generalmente, están etiquetados —por lo que se trata de un proceso supervisado. En este caso, pues, se parte de una colección de documentos de entrenamiento $\mathcal{D}^e$ que han sido agrupados previamente en un conjunto predefinido de categorías $\mathcal{C}$, para, a continuación, aprender y generalizar la correspondencia entre estos dos conjuntos de datos utilizando técnicas de representación del texto y de extracción de características (ver figura 3.5). En cambio, si no se dispone de datos etiquetados previamente, la clasificación de textos se convierte en una agrupación de textos (o *Text*

clustering en inglés) (Sebastiani, 2002). Éste es un problema de aprendizaje no supervisado, y se basa en la agrupación (*clustering*) de los documentos de entrenamiento ($\mathcal{D}^e$), según la similitud de los datos analizados, con el fin de obtener el conjunto de categorías $\mathcal{C}$ en las que organizar los textos. El número de categorías $|\mathcal{C}|$ puede estar prefijado de antemano o bien puede ser el propio algoritmo el encargado de definirlo durante el proceso de entrenamiento, utilizando algún criterio de homogeneidad de los datos en los grupos obtenidos (ver figura 3.6). Esta información puede ser utilizada para la tarea de clasificación en sí misma, o bien, como proceso previo a la creación de un algoritmo de clasificación de textos.

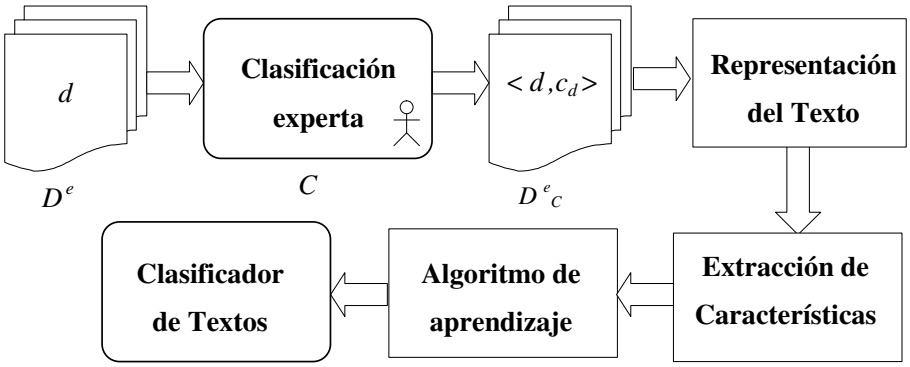


Figura 3.5: Proceso supervisado para la generación de un algoritmo de clasificación de textos.

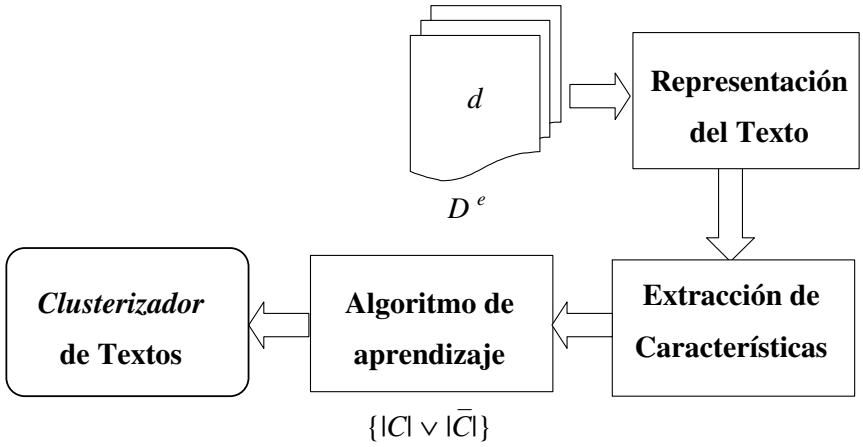


Figura 3.6: Proceso no supervisado para la generación de un algoritmo de agrupación de textos. Se puede indicar el número de categorías deseado ($|\mathcal{C}|$) o dejar que el algoritmo de análisis utilizado lo determine ($|\overline{\mathcal{C}}|$).

En general, la fase de aprendizaje de un algoritmo de CT no se realiza directamente sobre el propio texto que contienen los documentos, sino que éste es previamente parametrizado

y modelado mediante técnicas que provienen del ámbito de la RI. Para esta tarea, y según se muestra en las figuras 3.5 y 3.6, los algoritmos de CT suelen incorporar dos módulos de preprocesamiento del texto previos al algoritmo de aprendizaje. El primero es el módulo de *Representación del texto*, que se encarga de adecuar la representación de los textos a las necesidades del algoritmo de aprendizaje y al proceso de clasificación en sí. El segundo es el módulo de *Extracción de características*, que es el encargado de transformar el contenido de los textos en el conjunto de parámetros más significativos para la tarea considerada —denominados en la literatura de CT como términos, $\mathcal{T}$ (Sebastiani, 2002). De esta manera, el texto de los documentos se mapea sobre un nuevo espacio de datos donde el algoritmo de aprendizaje es capaz de funcionar de forma más eficiente.

Representación del texto

Como ya se ha comentado, las disciplinas de la CT y la RI están interrelacionadas entre sí, ya que ambas se enmarcan dentro del mundo del procesamiento automático del texto, aunque persiguen objetivos algo distintos. Los métodos de RI buscan representar, almacenar y organizar los datos para permitir el acceso a un determinado *ítem* de información (Salton y McGill, 1983) —en el presente trabajo, este ítem de información es el texto analizado. A partir de la consulta de un usuario (*query* en inglés), el sistema de recuperación de información se encarga de hallar el conjunto de documentos que mejor satisfagan los criterios de búsqueda del usuario (Frakes y Baeza-Yates, 1992), normalmente ordenado según alguna pauta que indique el grado de acierto o relevancia de la respuesta (p.ej. el resultado de una consulta en un buscador por Internet tipo Google, Yahoo!, etc.). Para realizar esta búsqueda de forma eficaz, es crucial disponer de una buena representación de los textos *electrónicos* analizados para poder ser tratados de forma eficiente en el dominio informático. Asimismo, procesos como la extracción, la recuperación o la clasificación de información pueden hacer uso de esta parametrización del texto, definida en el ámbito de la RI, para tratar el contenido de los documentos a procesar.

Por otra parte, cabe puntualizar la diferencia existente entre las técnicas de RI y las técnicas de extracción de información (EI) (*Information Extraction* en inglés). Los algoritmos de EI persiguen encontrar los fragmentos de un documento o de una colección de documentos que mejor se ajusten a la información requerida por el usuario (Cowie y Lehnert, 1996; Baeza-Yates y Ribeiro-Neto, 1999). Por ejemplo, la consulta temática de las películas de un determinado director dentro de su filmografía, o la búsqueda de unas declaraciones políticas dentro de un debate trascrito, entre otras. A veces, las técnicas de EI pueden hacer uso de un algoritmo de CT como preprocesamiento para refinar el subconjunto de documentos donde aplicar el proceso de selección de contenidos.

El formato más sencillo de representación de los textos es el que trata a los documentos como una simple colección de palabras o *bolsa* de palabras (*bag-of-words* en inglés), sin considerar sus relaciones ni el orden en qué éstas aparecen en el texto. Aunque se puede trabajar con estructuras lingüísticas complejas, p.ej. frases, la palabra es la unidad de información básica más utilizada en el ámbito de los algoritmos de RI o CT (Sebastiani, 2002). El motivo principal de esta parametrización es que la mayoría de algoritmos desarrollados

Tabla 3.1: Ejemplo ilustrativo de una colección formada por 6 documentos ($\mathcal{D}= \{d_{k=1:6}\}$) que contiene 4 términos distintos ($\mathcal{T}= \{A, B, C, D\}$).

Documento	**Términos**
d_1	$D\ B\ C\ C$
d_2	$A\ A$
d_3	$A\ C\ D$
d_4	$D\ D\ B\ B$
d_5	$C\ A$
d_6	A

hasta el momento han primado encontrar la temática del documento antes que su estructura, tomando más importancia la palabra en sí que su ordenación. En este contexto, frases como *"La casa de la chica"* y *"La chica de la casa"*, pasan a ser representadas del mismo modo al contener las mismas palabras, aunque su significado sea distinto. No obstante, también existen estrategias de clasificación de textos no temáticas que sí fijan su atención en parámetros estilísticos del texto, como la riqueza del vocabulario, la longitud media de las frases y las palabras, etc. (Stamatatos, Kokkinakis y Fakotakis, 2000; Sebastiani, 2005).

A continuación se presenta una breve descripción de las propuestas más importantes para la representación del texto a partir de las palabras $\mathcal{P}$ que lo forman, las cuales han sido definidas como base de muchos de los sistemas actuales de análisis de documentos. Los modelos más conocidos son el *índice inverso*, el *espacio vectorial* y la *representación probabilística*. Ésta última se presentará conjuntamente con el método de aprendizaje basado en la Regla de Bayes, descrito más adelante dentro de esta misma sección (ver la ecuación (3.15)).

- **El modelo de índice inverso:** es una representación que equivale al índice de palabras (glosario) de un libro de texto. Para cada palabra se obtiene la lista de documentos en los que éste aparece (indexación inversa). Por ejemplo, en el contexto de la EI, el usuario determinaría el grupo de palabras clave que quiere encontrar en los documentos, y una vez realizada la búsqueda, el sistema le presentaría el subconjunto de documentos que mejor se adecuara a la consulta. Se trata de una representación poco robusta de los datos, por lo que a menudo necesita de distintos accesos (filtrados sucesivos) para que el usuario pueda obtener los resultados deseados (Frakes y Baeza-Yates, 1992). En la tabla 3.2 se presenta el resultado de esta representación para el ejemplo de colección de documentos representado en la tabla 3.1.

- **El modelo de espacio vectorial (MEV):** es uno de los modelos de representación de textos más utilizados en el ámbito de la RI (Salton y Buckley, 1988; Salton, 1989), y, por extensión, en el ámbito de la clasificación automática de textos (Sebastiani, 2002). Según esta estrategia, cada documento d_k está representado por un vector

Tabla 3.2: Representación del contenido de la colección de documentos $\mathcal{D}$ de la tabla 3.1 mediante el modelo de índice inverso. Cada palabra se describe por un vector de parejas (d_k, número de apariciones de la palabra en d_k).

Término	**Índice inverso**
A	$\{(d_2,2),(d_3,1),(d_5,1),(d_6,1)\}$
B	$\{(d_1,1),(d_4,2)\}$
C	$\{(d_1,2),(d_3,1),(d_5,1)\}$
D	$\{(d_1,1),(d_3,1),(d_4,2)\}$

de ponderaciones de términos (parámetros) dentro del espacio vectorial definido por todos los términos $\mathcal{T}$ presentes en la colección de documentos —ver expresión (3.1).

$$\vec{d_k} = (\omega_1^k, \omega_2^k, \ldots, \omega_{|\mathcal{T}|}^k) \in \mathbb{R}^{|\mathcal{T}|} \tag{3.1}$$

donde $|\mathcal{T}|$ denota el número de términos (o parámetros) que representan el contenido de la colección de textos y ω_i^k es un escalar (ponderación) que representa el término i en el documento d_k (ver apartado de "*Ponderaciones y medidas de similitud*" en esta misma sección). En este contexto, cada término puede ser interpretado como una dimensión en el espacio multidimensional definido. Gracias a esta representación vectorial, se podrán aplicar los operadores (distancias) del mundo del Álgebra Lineal para la clasificación de los textos (como se verá en el apartado de "*Ponderaciones y medidas de similitud*" que se describe en esta misma sección y en el apartado 3.3.3 en el que se describe el método de CT propuesto).

A partir de la colección de $\mathcal{D}^e$ se define el MEV de $\mathbb{R}^{|\mathcal{T}|\times|\mathcal{D}^e|}$ sobre el que se llevará a cabo la tarea de la clasificación, donde $|\mathcal{D}^e|$ es el número total de documentos de entrenamiento. Asimismo, este MEV puede representarse de forma matricial mediante la expresión (3.2).

$$\mathbf{A} = (\mathrm{a}_{ik}) = \omega_i^k, \forall\, (1 \leq i \leq |\mathcal{T}|, 1 \leq k \leq |\mathcal{D}^e|) \tag{3.2}$$

Generalmente, la matriz $\mathbf{A}$ contendrá un número elevado de posiciones nulas debido a que no todas las palabras, y en consecuencia sus parámetros derivados, aparecen en todos los documentos (Aas y Eikvil, 1999). Por lo tanto, se trata de una matriz de carácter disperso (en inglés, *sparse*). Esta cuestión es abordada en el contexto de los sistemas de clasificación temática de documentos mediante la aplicación de técnicas de "*Extracción de características*", que pasan a describirse a continuación.

Extracción de características

El gran tamaño junto al carácter disperso de la matriz **A** dificultan y ralentizan cualquier proceso de análisis del texto que la utilice, tanto para clasificación como para recuperación de información, sobretodo en el contexto del análisis de grandes volúmenes de datos (p.ej. los buscadores de Internet trabajan con billones de páginas *web*). Como consecuencia, el rendimiento del sistema se reduce de forma considerable. En el contexto de los sistemas de clasificación *temática* de documentos, el módulo de extracción de características se encarga de reducir el volumen de información a tratar (ver figuras 3.5 y 3.6). Este módulo incorpora dos procesos: primero, se preprocesa el texto y luego se reduce el espacio de búsqueda, eliminando los datos que son superfluos para la clasificación temática de los textos. De este modo, se consigue disminuir el coste computacional del proceso de clasificación (entrenamiento y explotación), permitiendo que el algoritmo de aprendizaje se concentre en la información más relevante del texto temáticamente hablando. En contraposición, los sistemas de CT no temáticos o estilísticos necesitan de toda la información presente en el texto, por lo que el módulo de extracción de la información tendrá otra funcionalidad. En este caso se encarga de obtener los parámetros necesarios para llevar a cabo la tarea definida (p.ej. riqueza de vocabulario, número y tipo de frases, etc.) (Stamatatos, Kokkinakis y Fakotakis, 2000). Por lo tanto, en general, el número de parámetros considerado $|\mathcal{T}|$ será menor que $|\mathcal{P}|$, tanto para la clasificación temática —debido a la extracción de características— como para las no temáticas —centradas en determinados parámetros del texto—, donde $|\mathcal{P}|$ indica el número de palabras de los textos analizados.

La primera fase del módulo de extracción de características se encarga de **preprocesar el texto**. Este proceso puede ser tan sencillo como eliminar caracteres extraños del texto —como en el presente trabajo de investigación— (p.ej. comillas, retornos de carro, tabuladores, etc.), hasta filtrar los datos que contienen los documentos en el caso de la clasificación temática de textos. En este segundo caso, las dos técnicas aplicadas habitualmente son:

- **Eliminación de palabras vacías (o *stop words*):** la mayoría de documentos están llenos de palabras que aportan muy poca información sobre la temática del texto, p.ej. artículos, preposiciones o conjunciones. Éstas se suelen agrupar en listas de palabras frecuentes (lista de parada o *stop list*, en inglés), para ser eliminadas del texto como paso previo a la clasificación. El grupo de palabras a eliminar del texto se puede obtener a partir de una lista general de palabras para el idioma en cuestión, o bien, mediante un análisis estadístico del corpus a clasificar, eliminando aquellas palabras con una frecuencia superior a la fijada por un umbral. A partir de este filtrado, se pueden alcanzar reducciones muy importantes de la cantidad de términos a considerar —del orden de un 50 % (Sebastiani, 2002)—, sin afectar significativamente a la fiabilidad de la clasificación temática.
- **Extracción del radical (o *stemming*):** consiste en eliminar la variabilidad morfosintáctia de las palabras, extrayendo las terminaciones flexivas (p.ej. diseños → diseño) y derivativas (p.ej. diseñador → diseño) (ver cap. 6 de (Martí et al., 2003)). Por ejemplo, todas las conjugaciones de un verbo se representan por uno mismo lexema, o

palabras como *producto* y *producción* pasan a ser representadas por su raíz (*stem*, en inglés) *produc-*. De este modo, se consigue reducir el tamaño del espacio de búsqueda, minimizando los efectos de la flexión del idioma en la caracterización de los documentos. Uno de los algoritmos más utilizados para el inglés es el que se presentó en (Porter, 1980). Sin embargo, este algoritmo no puede aplicarse a cualquier idioma, ya que no todos los idiomas presentan una variabilidad morfológica similar, por lo que el diseño de este tipo de algoritmos debe estar adaptado al idioma de trabajo. Es importante destacar que la morfología flexiva del castellano y el catalán es mucho más compleja que la del inglés (Vilares, Barcala y Alonso, 2001), cuestión que dificulta la adaptación e implementación del algoritmo de *stemming* para estos idiomas.

Además de estas dos técnicas, en el contexto de los sistemas de clasificación temática de textos se suelen aplicar otras estrategias para reducir aún más el espacio de búsqueda mediante la eliminación de la información menos significativa desde un punto de vista temático. De este modo, se obtiene un nuevo espacio formado por el conjunto de términos (parámetros) $|\mathcal{T}'| \ll |\mathcal{T}|$ de mayor representatividad para la tarea. Existen varios trabajos donde se presentan diversas comparativas, más o menos exhaustivas, en lo referente a métodos de reducción del espacio de términos, entre los que destacan (Yang y Pedersen, 1997; Sebastiani, 2002). En estos trabajos se definen dos de los enfoques más relevantes para abordar la reducción del tamaño del espacio de búsqueda:

- **Selección de palabras:** consiste en escoger el subconjunto de palabras que aportan *mayor* información sobre los contenidos de los documentos, desechando el resto de palabras, sin cambiar de espacio de representación. Algunas de las técnicas más empleadas para determinar el grado de información aportado por las palabras se basan en el cálculo de la ganancia de información, la estadística chi-cuadrado (χ^2) o la información mutua de los términos, entre otras (Yang y Pedersen, 1997). Todas ellas, se aplicarán con el objetivo de escoger el mínimo conjunto de términos que mejor represente la información contenida en la colección de documentos, según el compromiso existente entre el tamaño del espacio de términos obtenido y el número de conceptos (información) representados.
- **Extracción de términos:** estas técnicas no sólo reducen la cantidad de datos a considerar, sino que además definen un nuevo espacio de búsqueda sobre el que aplicar la clasificación de los nuevos documentos. Entre estos métodos destacan la agrupación (o *clustering*) de términos (parámetros) y el algoritmo de indexación de semántica latente (o *latent semantic indexing* o LSI, en inglés) (Deerwester et al., 1990). El primero, agrupa los términos en conjuntos según su similitud, utilizando sólo a un representante de cada grupo (*centroide*) para la clasificación. El segundo, se basa en la descomposición en valores singulares (*Singular Value Decomposition* o SVD, en inglés) del espacio vectorial de términos, definiendo una base ortogonal del espacio de partida, donde cada dimensión representa una dirección que aporta un significado máximo de la colección. En la misma línea, pero fundamentados en el análisis en componentes independientes (*Independent Component Analysis* o ICA, en inglés),

se presentan otras alternativas que buscan, además, la independencia estadística de las componentes que representan los textos (Isbell y Viola, 1999; Kaban y Girolami, 2000). Siguiendo esta filosofía, a lo largo del presente trabajo de investigación, se ha colaborado con Xavier Sevillano en el diseño y posterior desarrollo de un sistema de clasificación de textos basado en ICA (ver (Alías et al., 2003; Sevillano, Alías y Socoró, 2004; Alías et al., 2004b) para más detalles), del que se presentan algunos resultados en el apartado de experimentos descrito en la sección 3.4.

Ponderaciones y medidas de similitud

El contenido del espacio de búsqueda (p.ej. el espacio vectorial) queda definido por la información con la que se parametriza el texto de los documentos, a partir de las palabras que lo componen. En el ámbito de la CT, esta información deberá permitir discernir la categoría de cada documento d_k respecto al resto de documentos que constituyen la colección $\mathcal{D}$. Por este motivo, en la literatura de CT se han definido distintas ponderaciones que permiten cuantificar el grado de discriminación que presentan las palabras de un documento según la tarea de clasificación en la que se utilicen. Los métodos de ponderación de los términos (*term weighting* en inglés) han sido estudiados y propuestos en muchos trabajos, así como en la literatura clásica de la RI (Salton y McGill, 1983).

Antes de pasar a describir las funciones de ponderación, es necesario destacar dos de los parámetros que intervienen en la mayoría de los métodos de ponderación de textos que se describirán seguidamente:

- ***tf***: es el *term frequency* o número de apariciones de un término en un documento. Asocia el grado de importancia de un término a su frecuencia de aparición en el documento. Entonces, para cada término i y para cada documento d_k se obtiene el correspondiente valor de tf_i^k.

- ***idf***: es el *inverse document frequency* y parametriza el carácter discriminativo entre categorías del término, según su distribución a lo largo de toda la colección de documentos. Por ejemplo, en el ámbito de la CT temática, si dentro de un texto se encuentra la palabra "*gol*", parece claro que este documento tratará de deportes, pero si otro documento contiene la palabra "*ganar*", ese documento puede tratar de deportes, economía, política, etc. Entonces, cuanto más discriminativo sea un término, mayor importancia deberá tomar para la clasificación del documento, respecto del resto de términos del texto. Su cálculo se presenta en la ecuación (3.3), donde n_i es el número total de documentos en los que aparece el término i, dentro de la colección de $|\mathcal{D}|^e$ documentos. Por lo tanto, cuanto mayor sea n_i (el término aparece en un mayor número de documentos), menor será el valor de idf_i, es decir, menor será el grado de discriminabilidad del término dentro de la colección de documentos.

$$idf_i = log\left(\frac{|\mathcal{D}^e|}{n_i}\right), \forall\, n_i > 0 \tag{3.3}$$

Una vez realizado este inciso, se pasa a presentar una breve descripción de algunas de las ponderaciones de términos (parámetros) (ω_i^k en las ecuaciones (3.1) y (3.2)) más utilizadas en el ámbito de la clasificación de textos (Aas y Eikvil, 1999), algunas de las culaes se basan en las dos definiciones anteriores:

- **Booleana:** consiste en ponderar el término i con dos posibles valores: *cierto* ('1') si éste aparece en el documento d_k, o *falso* ('0') si no es así:

$$w_i^k = \begin{cases} 1, & si\ tf_i^k > 0 \\ 0, & \text{de lo contrario} \end{cases} \tag{3.4}$$

- **Frecuencia de los términos:** consiste en ponderar cada término según su número de apariciones en el texto. Es decir, se asigna $w_i^k = tf_i^k$ en la ecuación (3.1). Para el ejemplo presentado en la tabla 3.1, con $|\mathcal{D}|= 6$ y $|\mathcal{T}|= 4$, se obtendría la matriz (3.5) utilizando el MEV descrito anteriormente en la expresión (3.2):

$$\mathbf{A} = \left(\begin{array}{c|cccccc} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \hline A & 0 & 2 & 1 & 0 & 1 & 1 \\ B & 1 & 0 & 0 & 2 & 0 & 0 \\ C & 2 & 0 & 1 & 0 & 1 & 0 \\ D & 1 & 0 & 1 & 2 & 0 & 0 \end{array}\right) \tag{3.5}$$

  Si en lugar de trabajar con esta ponderación se trabajara con la ponderación booleana, la matriz (3.5) pasaría a contener sólo valores binarios. El valor de las casillas nulas se mantendría, pero el resto de posiciones cambiarían su valor por un '1' (aplicando la ecuación (3.4)).

- ***tf* × *idf*:** es el resultado de la propuesta hecha en (Salton y McGill, 1983) y una de las más utilizadas en la literatura. Esta ponderación combina el número de apariciones del término en el documento y su grado de discriminación en toda la colección:

$$w_i^k = tf_i^k \times idf_i \tag{3.6}$$

  Existen dos variantes principales de esta ponderación. La primera, denominada tfc (del inglés *term frequency collection*), aplica la distancia del coseno para normalizar el producto $tf \times idf$ respecto a todos los términos del documento (ecuación (3.7)) (Salton y Buckley, 1988). Así se elimina el efecto de la diferencia de longitud de los documentos en la comparación[12].

$$w_i^k(tfc) = \frac{tf_i^k \times idf_i}{\sqrt{\sum_{s=1}^{|\mathcal{T}|} \left(tf_s^k \times idf_s\right)^2}} \tag{3.7}$$

[12]Esta normalización se realiza implícitamente cuando se utiliza la distancia del coseno en la comparación entre los vectores del MEV para realizar la clasificación, como se verá en el apartado de "*Distancias*" descrito a continuación.

La segunda variante, denominada *ltc* (del inglés *length term collection*), presentada por los mismos autores, y que modifica el cálculo de la frecuencia del término aplicándole también la función logaritmo (ecuación (3.8)) para mejorar la separabilidad de los datos gracias al recorrido de esta función.

$$w_i^k(ltc) = \frac{log\left(tf_i^k + 1\right) \times idf_i}{\sqrt{\sum_{s=1}^{|\mathcal{T}|}(log\left(tf_s^k + 1\right) \times idf_s)^2}} \tag{3.8}$$

- **Entropía:** la ponderación de los términos está basada en conceptos extraídos del mundo de la Teoría de la Información, caracterizando cada término por su grado de incertidumbre (entropía) dentro de la colección. Para una descripción más detallada, ver (Aas y Eikvil, 1999).

Una vez detalladas las ponderaciones más usadas para caracterizar los términos de la colección de documentos, se pasa a describir algunas de las medidas de similitud más utilizadas en el contexto de la clasificación automática de textos. Por un lado, destacan las distancias aplicadas sobre el MEV descrito anteriormente, y por el otro, las medidas de similitud basadas en la probabilidad de los datos (Willet, 1983):

- **Distancias:** miden el grado la similitud entre los documentos sobre un espacio de comparación común (MEV). Entre ellas, se encuentra la *distancia euclidiana* (ecuación (3.9)), la cual determina la diferencia entre dos vectores de documentos ($\vec{d_1}$ y $\vec{d_2}$) dentro del espacio multidimensional generado por los documentos y sus términos.

$$Dist_{euc}(\vec{d_1}, \vec{d_2}) = \sqrt{\sum_{i=1}^{|\mathcal{T}|}\left(w_i^{d_1} - w_i^{d_2}\right)^2} \tag{3.9}$$

Asimismo, la similitud entre los documentos se puede evaluar a partir de su *producto escalar* (ecuación (3.10)).

$$Dist_{esc}(\vec{d_1}, \vec{d_2}) = \sum_{i=1}^{|\mathcal{T}|}\left(w_i^{d_1} \cdot w_i^{d_2}\right) \tag{3.10}$$

El problema fundamental de calcular la similitud mediante este tipo de distancias reside en el hecho de no normalizar respecto a la longitud de los documentos comparados. Por este motivo, son medidas poco usadas en el ámbito de la clasificación de textos, ya que puede suceder que documentos poco parecidos presenten cierto grado de similitud, simplemente, por contener un número elevado de palabras. Para evitar este problema, se suelen utilizar distancias normalizadas, como por ejemplo: *Dice*, *Jaccard*, *Tanimoto*, entre otras (ver Apéndice A de (Tombros, 2002)). En este contexto, una de las distancias más utilizadas en la literatura de la CT es la *distancia del coseno*, que

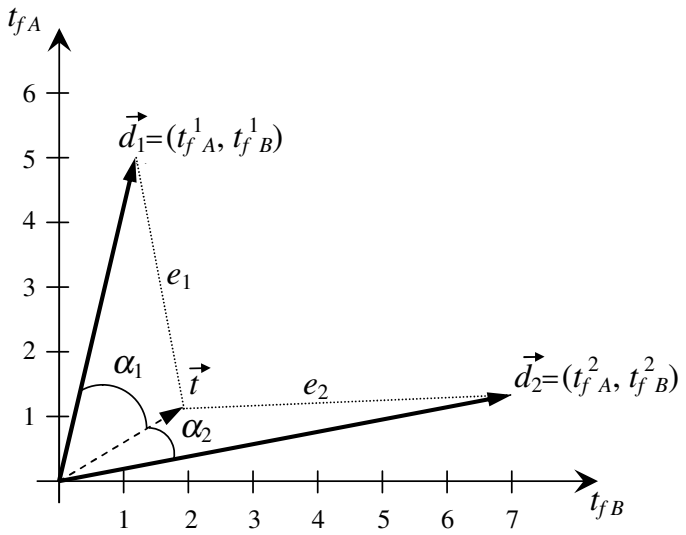


Figura 3.7: Representación vectorial de dos documentos (d_1 y d_2) y un documento de test (t), junto con los ángulos que forman (α_1 y α_2) y las distancias euclidiana (e_1 y e_2), sobre un modelo de espacio vectorial de dos dimensiones definido por el t_f del término A y del término B —adaptada de (Tombros, 2002).

mide el ángulo entre los vectores de términos que definen los documentos, considerando que las direcciones de los vectores son un mejor indicador del grado de similitud que la distancia que los separa (ver ecuación (3.11)). Por lo tanto, dos documentos idénticos describirán la misma dirección (ángulo de 0°) y dos documentos que no se parezcan en nada corresponderán a vectores ortogonales (ángulo de 90°).

$$Dist_{cos}(\vec{d_1}, \vec{d_2}) = \frac{< \vec{d_1}, \vec{d_2} >}{||\vec{d_1}|| \cdot ||\vec{d_2}||} = \frac{\sum_{i=1}^{|\mathcal{T}|} \left(w_i^{d_1} \cdot w_i^{d_2} \right)}{\sqrt{\sum_{i=1}^{|\mathcal{T}|} (w_i^{d_1})^2} \sqrt{\sum_{i=1}^{|\mathcal{T}|} (w_i^{d_2})^2}} \tag{3.11}$$

Por ejemplo, en la figura 3.7 se muestra un espacio vectorial simplificado a dos dimensiones (2D), es decir, definido únicamente por dos términos ($\mathcal{T} = \{A, B\}$). En esta figura se representa el documento $\vec{d_1}$ con 5 apariciones de A y una de B, $\vec{d_1} = (5, 1)$, el documento $\vec{d_2}$ con una aparición del término A y 7 de B, $\vec{d_2} = (1, 7)$, y el documento de test $\vec{t}$ con 1 de A y 2 de B, $\vec{t} = (1, 2)$. Sobre esta representación 2D se puede observar que el documento $\vec{t}$ es más próximo al segundo documento que al primero en lo que se refiere al ángulo, ya que $\alpha_2 < \alpha_1$, según la distancia del coseno (ver el cálculo en la expresión (3.12)). Así pues, si el documento $\vec{t}$ se tratara de una consulta en el contexto de la RI, el sistema tendría que devolver al usuario el documento d_2, que es el más se cercano a su consulta. En cambio, si el sistema fuera un método de CT, el documento de test tendría que ser asignado a la misma categoría a la que pertenece el documento d_2. Nótese que, en términos de distancia euclidiana, el documento $\vec{t}$ se

asignaría a la misma categoría del documento d_1 ($e_1 < e_2$), al tratarse de una distancia no normalizada respecto a la longitud del vector (ver el cálculo de la expresión (3.13)), aunque la distribución de los términos del documento de test es más cercana al documento d_2 (aparece más el término B que el A).

$$\begin{aligned} Dist_{cos}(\vec{d_1}, \vec{t}) &= \frac{5 \cdot 1 + 1 \cdot 2}{\sqrt{5^2 + 1^2}\sqrt{1^2 + 2^2}} = 0{,}614 = \alpha_1 \\ Dist_{cos}(\vec{d_2}, \vec{t}) &= \frac{1 \cdot 1 + 7 \cdot 1}{\sqrt{1^2 + 7^2}\sqrt{1^2 + 2^2}} = 0{,}506 = \alpha_2 \end{aligned} \tag{3.12}$$

$$\begin{aligned} Dist_{euc}(\vec{d_1}, \vec{t}) &= \sqrt{(5-1)^2 + (1-2)^2} = \sqrt{5} = e_1 \\ Dist_{euc}(\vec{d_2}, \vec{t}) &= \sqrt{(1-1)^2 + (7-2)^2} = 5 = e_2 \end{aligned} \tag{3.13}$$

- **Probabilidades:** en los métodos de análisis del texto basados en un modelo probabilístico, la similitud entre documentos se determina según la probabilidad de pertenencia de un documento a cada categoría (para CT), o como el conjunto de documentos que presentan una mayor probabilidad de dar respuesta a la consulta del usuario (para RI). En este segundo caso, el documento d_k se presenta al usuario si su probabilidad de ser relevante es mayor que la probabilidad de no serlo (probabilidades estimadas durante el proceso de aprendizaje del sistema). En ciertas situaciones, estas probabilidades condicionadas se pueden descomponer en factores sencillos que miden la probabilidad de aparición de los términos del documento de test o de la consulta en los documentos de la colección (ver cap. 6 de (Martí et al., 2003) para más detalles).

Todas estas medidas de similitud, descritas en el ámbito de la CT y la RI, también pueden ser utilizadas por los algoritmos de agrupación de textos (*text clustering*, en inglés) que necesitan incorporar el grado de similitud de los documentos de la colección durante el proceso de agrupación no supervisada de los mismos. En este contexto, documentos que contengan información similar se agruparán en regiones del espacio de representación similares. Por ejemplo, para el caso de la clasificación temática sobre un modelo de espacio vectorial utilizando la distancia del coseno, los documentos de una cierta clase presentarán direcciones semánticas o temáticas próximas.

Técnicas de clasificación

Una vez definidas la representación y la ponderación de los textos para el problema de la clasificación automática de documentos, se pasa a describir algunas de las técnicas de clasificación más utilizadas en el contexto de la CT. Esta tarea suele partir de un conjunto de documentos de entrenamiento $\mathcal{D}^e = \{d_1, d_2, \ldots, d_{|\mathcal{D}^e|}\}$ extraídos de la colección de documentos $\mathcal{D} = \{\mathcal{D}^e \cup \mathcal{D}^t\}$, donde $\mathcal{D}^t = \{d_{|\mathcal{D}^e|+1}, d_{|\mathcal{D}^e|+2}, \ldots, d_{|\mathcal{D}|}\}$ representa el conjunto de documentos de test —por lo que $|\mathcal{D}^t| = |\mathcal{D}| - |\mathcal{D}^e|$. En la clasificación supervisada, todos los

documentos de colección $\mathcal{D}$ estarán etiquetados según un grupo predefinido de categorías $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$, donde $|\mathcal{C}|$ indica el número de categorías considerado. A partir de estos datos y, mediante un proceso de aprendizaje, se modelará la información dada por el experto (utilizada para clasificar la colección de documentos dentro del conjunto de categorías considerado) para definir la función f que describa la siguiente aplicación:

$$f : \mathcal{D}^e \rightarrow \mathcal{C} \tag{3.14}$$

Esta función se encarga de proyectar cada documento del conjunto de entrenamiento $\mathcal{D}^e$ en la categoría (*hard clasification*) o conjunto de categorías correspondientes (*soft classification*) (Sebastiani, 2002). De este modo, se pretende modelar la información disponible durante la fase entrenamiento, expresada mediante correspondencias del tipo $< d_k, c_n >$ o $< d_k, c_{n_1}, c_{n_2}, \ldots >$, respectivamente. Es decir, cada documento de entrenamiento $d_k \in \mathcal{D}^e$ $(1 \leq k \leq |\mathcal{D}^e|)$ tiene asociada una o diversas categorías c_n $(1 \leq n \leq |\mathcal{C}|)$[13], relaciones que el método de CT deberá aprender y generalizar para poder clasificar los documentos de test $d_k \in \mathcal{D}^t$, con $|\mathcal{D}^e| < k \leq |\mathcal{D}|$, en adelante documentos o textos t_k —aunque también pueden denominarse documentos consulta . En este contexto es donde se aplican los algoritmos de aprendizaje artificial para modelar y generalizar las relaciones de los datos de entrenamiento y poder utilizar esta información durante la clasificación de nuevos documentos. Existe una gran variedad de técnicas aplicadas al mundo de la clasificación automática de textos. A continuación se presenta una pequeña descripción de algunas de ellas, detallando brevemente el fundamento teórico que las sustenta.

- **El algoritmo de Rocchio:** es un algoritmo basado en el aprendizaje lineal. Utiliza $tf \times idf$ para la ponderación de los datos que representan los documentos, y se define sobre el MEV construido a partir de los términos de la colección de documentos. El algoritmo se basa en el cálculo de un vector *prototipo* obtenido del promediado de los vectores de los documentos de entrenamiento, como representante de cada categoría (*centroide*). Durante el proceso de clasificación, cada documento es asignado a la categoría que tiene el vector prototipo más parecido a él (habitualmente, según la distancia del coseno). Es un algoritmo sencillo y poco robusto, pero que ha sido de los más utilizados en los primeros sistemas de clasificación automática de documentos. Dentro del mismo grupo de algoritmos de aprendizaje se puede encontrar el algoritmo de Widrow-Hoff o variantes basadas en el gradiente exponenciado, como el algoritmo de Kivinen-Warmuth, entre otros (García, Martín y Ureña, 2001).

- **Inducción de reglas y árboles de decisión:** una categoría puede ser descrita a partir de los atributos más relevantes de los datos que la conforman, que, una vez listados adecuadamente, pueden ser utilizados para llevar a cabo de forma simple la clasificación. A partir de la generalización de esta lista de parámetros se pueden obtener las reglas de clasificación simples (inducción de reglas), o bien, si existen

[13] En el caso que las parejas sean únicas, del tipo $< d_k, c_n >$, cada documento tendrá una única proyección en el espacio de categorías, obteniendo una aplicación f exhaustiva. Sin embargo, en el caso que los documentos se asignen a más de una categoría, deja de tratarse de una aplicación estrictamente hablando.

dependencias entre ellos, obtener reglas estructuradas de forma jerárquica, organizadas mediante árboles de decisión, habitualmente. En el caso de la clasificación binaria, cada nodo de uno de estos árboles de decisión está conectado a un conjunto de nodos mutuamente excluyentes. En el proceso de entrenamiento se evalúa la diversidad de los elementos de cada nodo durante la fase de construcción del árbol (normalmente basada en su *entropía*). A la hora de clasificar un nuevo documento, sólo será necesario recorrer todos los nodos que aportan respuestas positivas a la consulta hasta llegar al nodo final, que en el caso de la CT, indicará la categoría asignada al documento de test. Algunos de los algoritmos más utilizados para la construcción de árboles de decisión son C4.5 (Quinlan, 1993) o CART (*Classification and Regression Tree*) (Breiman et al., 1984) y, para inducción de reglas, RIPPER (Cohen, 1995), entre otros.

- ***Nearest neighbour*:** es una técnica de clasificación directa sobre un espacio de comparación común a partir de la agrupación de los datos en los conjuntos definidos. La categoría a la que pertenecerá el documento a clasificar será la que corresponda al ejemplo del conjunto de entrenamiento al que éste más se parezca. La similitud se suele determinar a partir de una distancia euclidiana o del coseno (extraída del modelo de espacio vectorial). Uno de los algoritmos más populares para este tipo de clasificación es el *k-nearest neighbour* (Friedman, 1994), en el que se consideran las k instancias más próximas al ítem analizado, clasificándose éste en la categoría más votada. Normalmente, el valor de k se ajusta de forma empírica según las particularidades de la aplicación en la que se enmarca.

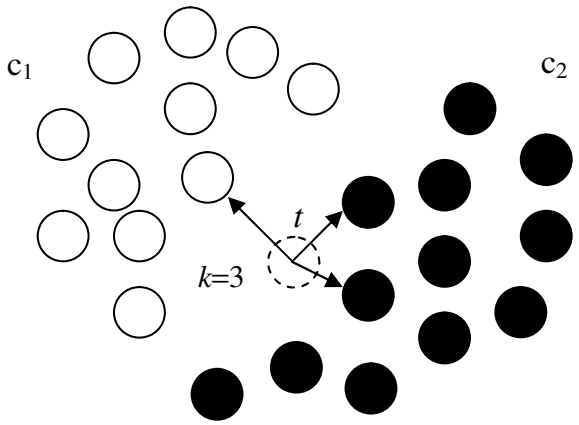


Figura 3.8: Clasificación de un documento de test (t) con un algoritmo de 3-*nearest neighbour* en un espacio de clasificación definido por 2 categorías: c_1 y c_2.

En la figura 3.8 se presenta un ejemplo del funcionamiento de este algoritmo sobre 2 categorías (c_1 y c_2) distintas, junto a la clasificación de un documento de test t teniendo en cuenta los 3 vecinos más próximos ($k = 3$). La clasificación puede realizarse por simple cuenteo (en el ejemplo, $\{2 \text{ vecinos} \in c_2, 1 \text{ vecino} \in c_1\} \rightarrow t \in c_2$) o a partir de un sumatorio ponderado de similitudes, por lo que la clasificación dependerá de las

distancias y los pesos aplicados (*weigthed-sum voting scheme*, en inglés).

- **Redes neuronales:** es un modelo de análisis y representación de los datos basado en el funcionamiento del cerebro humano, ya que tiene un funcionamiento no lineal y una estructura inspirada en las neuronas y sus sinapsis. A partir de un conjunto de unidades de proceso conectadas entre sí según un determinado peso, la red es capaz de encontrar y modelar determinados patrones presentes en el conjunto de datos de entrenamiento. Dada la generalidad de esta aproximación, las redes neuronales han sido aplicadas a infinidad de campos, incluyendo la visión artificial, el reconocimiento de voz, la predicción de índices financieros, entre otros. En el contexto de la clasificación automática de textos, la red tomará los términos que componen los documentos como datos de entrada del proceso de entrenamiento —en el ámbito de la CT se suelen utilizar perceptrones multicapa entrenados, normalmente, mediante el algoritmo de *backpropagation* (Sebastiani, 2002). Durante el proceso de explotación de la red, su salida indicará cuál es la categoría o conjunto de categorías asignadas al documento de test. En el caso de abordar el problema de la CT mediante redes neuronales, resulta necesario disponer de un conjunto de entrenamiento de tamaño considerable para modelar satisfactoriamente las relaciones existentes entre los términos de la colección, y, así, permitir que la red pueda generalizarlas adecuadamente. El conocimiento aprendido por la red se representa mediante los pesos de las conexiones entre las neuronas de la red (que pueden estar divididas en diversas capas de neuronas) (Sebastiani, 2002).

- **Métodos probabilísticos:** son métodos que permiten representar la causalidad de los datos analizados basándose en la probabilidad de su aparición. Uno de los métodos más utilizados es el basado en el *Teorema de Bayes* (Peng y Schuurmans, 2003): a partir de un evento aleatorio a y de una determinada evidencia e, se puede predecir la probabilidad *a posteriori* de que suceda a dado e, siempre que se conozca la probabilidad condicionada entre ambos (en este caso de e condicionado por a) y su probabilidad de ocurrencia individual. Para el caso de la CT, el documento a clasificar actuará como evidencia mientras que la variable aleatoria corresponderá a la clase a la que puede pertenecer. Por lo tanto, la ecuación (3.15) permitirá estimar la probabilidad que el documento d_k pertenezca a la clase c_n dentro de la colección de categorías $\mathcal{C}$.

$$P(c_n|d_k) = \frac{P(c_n)P(d_k|c_n)}{P(d_k)} \tag{3.15}$$

  Entonces, de entre las $|\mathcal{C}|$ categorías posibles, el documento d_k será asignado a la categoría que presente un $P(c_n|d_k)$ máxima, o bien, a aquel subconjunto de categorías $\mathcal{C}_{d_k} \subseteq \mathcal{C}$ que presenten una probabilidad superior a un determinado umbral de pertenencia: $P(c_n|d_k) > \gamma$. Durante la fase de entrenamiento del CT basado en redes Bayesianas se calcularán las probabilidades condicionadas a partir del conjunto de documentos ejemplo (conjunto de entrenamiento).

  Por otra parte, en este contexto, normalmente se asume que las componentes del vector d_k (los pesos ω_i^k de los términos que lo componen) son estadísticamente inde-

pendientes, obteniendo la versión de los métodos probabilísticos más utilizada en el ámbito de la CT: los clasificadores *Bayesianos Naif* (ecuación (3.16)). El carácter *naif* del clasificador es debido al hecho de que, generalmente, la hipótesis de independencia entre las variables no es del todo cierta a la práctica (Sebastiani, 2002).

$$P(d_k|c_n) = \prod_{i=1}^{|\mathcal{T}|} P(\omega_i^k|c_n) \tag{3.16}$$

En el ámbito de los métodos probabilísticos, además de las redes Bayesianas, los modelos de lenguaje basados en *n-gramas* también han sido aplicados a la clasificación de textos, tanto a nivel de carácter (Cavnar y Trenkle, 1994) como de palabra (Peng y Schuurmans, 2003). Los modelos de lenguaje se aplican a la tarea de la CT mediante una formulación similar a la utilizada por los clasificadores Bayesianos Naif (ver (Peng y Schuurmans, 2003) para más detalles), pero con la diferencia que en este caso se modela la dependencia Markoviana entre palabras consecutivas (Peng y Schuurmans, 2003). Durante el proceso de entrenamiento se halla la probabilidad condicionada de aparición de cada *n-grama* para cada una de las categorías consideradas (c_i), según la probabilidad de aparición de un ítem (carácter o palabra) condicionada a la aparición previa de $n-1$ ítems, con sus correspondientes ponderaciones w_i^j en la ecuación (3.17).

$$P(d_k|c_n) = \prod_{i=1}^{|\mathcal{T}|} P_{c_n}(\omega_i^k|\omega_{(i-1)}^k), \ldots, \omega_{(i-n+1)}^k \tag{3.17}$$

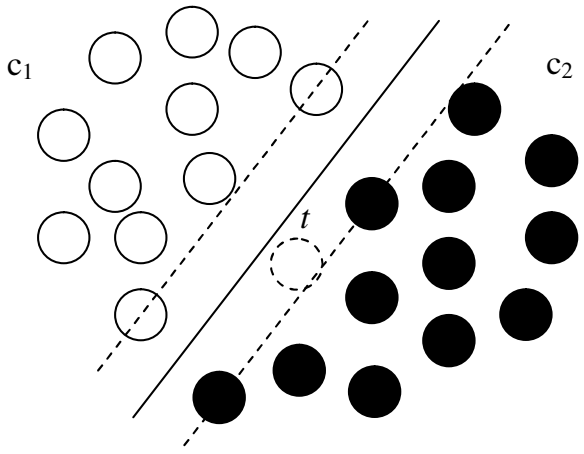


Figura 3.9: Clasificación de un documento de test (t) mediante una SVM en un espacio de clasificación definido por 2 categorías: c_1 y c_2. Los ejemplos sobre la línea discontinua representan los vectores de soporte.

- **Máquinas de soporte vectorial:** del inglés *support-vector machines* o SVM. Es uno de los métodos más utilizados (y que han aportado mejores resultados) en el ámbito de la CT temática de textos sobre grandes colecciones de documentos (Joachims, 1998; Joachims, 2002). Las SVM fueron definidas para ser aplicadas a tareas de clasificación binaria ($\mathcal{C} = \{-1, 1\}$), ya que tiene como objetivo encontrar la superficie que divida de forma *óptima* (máxima separación de los datos) el espacio de datos de entrenamiento en dos zonas independientes: la formada por los ejemplos positivos y la formada por los negativos. La superficie de separación se obtiene maximizando el área de separación entre estas dos zonas del espacio. Los *vectores de soporte* son aquellos vectores equidistantes que delimitan cada una de las áreas del espacio mediante una superficie paralela a la de clasificación (Sebastiani, 2002) —ver el ejemplo de la figura 3.9. Para el caso de la clasificación automática de textos en varias categorías, el problema se tiene que modelar como una sucesión de problemas dicotómicos. Durante la fase de clasificación sólo será necesario tener en cuenta los vectores de soporte, reduciendo drásticamente la memoria del clasificador al no tener que considerar todos los datos del espacio (como sucede, p.ej. en *nearest-neighbour*). De este modo, la aproximación de SVM integra tanto la reducción de dimensiones, como la propia tarea de clasificación (Aas y Eikvil, 1999).

  En la figura 3.9 se presenta el mismo ejemplo de la figura 3.8, pero esta vez resuelto mediante la aplicación de SVM. En ambos casos, el documento de test t sería asignado a la misma categoría, ya que en este caso este documento se sitúa dentro de la zona del espacio que corresponde a la categoría c_2 (mínima distancia a sus vectores de soporte).

Por otra parte, existen sistemas de clasificación de textos que no se basan en una única estrategia de aprendizaje, sino que obtienen la clasificación combinando las decisiones de varios clasificadores independientes. El sistema de CT global se obtiene del entrenamiento de un mismo sistema contra distintos conjuntos de test, o bien, a partir de la combinación de diferentes algoritmos de clasificación basados en estrategias distintas (aportan soluciones parciales al problema). Las dos filosofías más importantes dentro de este tipo de sistemas de clasificación por votación se denominan con los vocablos ingleses *bagging* y *boosting* (ver (Aas y Eikvil, 1999) para una breve descripción). Dos de los algoritmos más conocidos de este tipo de clasificadores son *AdaBoost* (Freund y Schapire, 1996) y *Boostexter* (Schapire y Singer, 2000).

Medidas de evaluación

Una vez construido el modelo de CT a partir de los datos de entrenamiento ($< \mathcal{D}^e, \mathcal{C} >$), se procede a su explotación. Para ello se utilizan un conjunto de documentos de test ($t_k \in \mathcal{D}^t$), que el experto también ha asignado a alguna de las categorías de la colección $\mathcal{C}$. Una vez realizado todo el proceso de clasificación automática para todos los documentos de la colección de test ($|\mathcal{D}^t|$), es necesario evaluar la eficiencia del algoritmo de CT desarrollado. La tabla 3.3 presenta un cuadro que relaciona las soluciones obtenidas por el CT respecto a las decisiones tomadas por el experto, para una determinada categoría de documentos.

Tabla 3.3: Distribución de los resultados obtenidos al clasificar los documentos en una determinada categoría c_n de la colección.

	Experto	
Clasificador	*Cierto*	*Falso*
Cierto	α_n	β_n
Falso	γ_n	δ_n

α_n: # aciertos de clasificación
β_n: # errores de clasificación
γ_n: # errores de desclasificación
δ_n: # aciertos de desclasificación

Los algoritmos de CT perseguirán la maximización del número de aciertos en la clasificación (α_n y δ_n, en la tabla 3.3) junto a la minimización del número de errores (β_n y γ_n, en la tabla 3.3). Las medidas de evaluación del funcionamiento de un CT suelen ser adaptaciones de las usadas en el mundo de la RI o del AA. A continuación se describen algunas de las más utilizadas para la evaluación de los sistemas de clasificación de textos, a partir de los datos descritos en la tabla 3.3:

- **Precisión: (*precision*)** evalúa la probabilidad de acierto del clasificador en asignar un documento de test a una determinada categoría. Es decir, de los documentos asignados a una determinada categoría, qué porcentaje pertenece realmente a esa categoría. Siguiendo la tabla 3.3, la precisión del clasificador se obtiene mediante la ecuación (3.18).

$$P_n = \frac{\alpha_n}{\alpha_n + \beta_n} \tag{3.18}$$

- **Cobertura (*recall*):** evalúa la probabilidad de haber clasificado un documento en la categoría a la que pertenece. Es decir, de los documentos que deberían haber sido clasificados en una determinada categoría, qué porcentaje del total ha sido finalmente asignando a esa categoría. A partir de la tabla 3.3, el grado de cobertura del clasificador se obtiene mediante la ecuación (3.19).

$$R_n = \frac{\alpha_n}{\alpha_n + \gamma_n} \tag{3.19}$$

- **Exactitud (*accuracy*):** evalúa el grado de acierto general del clasificador respecto a todos los documentos clasificados para la categoría analizada. Es una medida poco utilizada en el ámbito de la CT debido a que, como se presenta en la ecuación (3.20) y se indica en (Yang y Liu, 1999), el elevado valor que el denominador suele tomar hace que esta medida sea menos sensible que la precisión y la cobertura a las variaciones

del número de aciertos (α_n y δ_n), además de otros factores descritos en (Sebastiani, 2002).

$$A_n = \frac{\alpha_n + \delta_n}{|\mathcal{D}^t|} \tag{3.20}$$

donde $|\mathcal{D}^t| = \alpha_n + \beta_n + \gamma_n + \delta_n$.

A partir de esta medida, también se puede obtener el error del clasificador para cada categoría c_n como: $E_n = 1 - A_n$.

Una vez descritas las medidas más utilizadas para evaluar la eficiencia de los algoritmos de clasificación para cada una de las categorías de $\mathcal{C}$, resulta necesario obtener un resultado global del funcionamiento del sistema de CT, a partir del promediado de los resultados parciales (por categoría). En la literatura, se presentan dos variantes para la obtención de una tasa que indique la eficiencia global del clasificador:

- **micro-promediado (μ):** el valor global de la medida de evaluación se obtiene como la media de los aciertos por documento tratados de forma individual, a lo largo de todas las categorías, independientemente del número de documentos de cada categoría. Las ecuaciones (3.21) y (3.22) presentan el cálculo de las medidas de precisión y de cobertura según el micro-promediado. Como resultado, se obtendrá la precisión y la cobertura total del clasificador, dado el resultado de la clasificación del conjunto de documentos de test considerado.

$$P^{\mu} = \frac{\sum_{n=1}^{|\mathcal{C}|} \alpha_n}{\sum_{n=1}^{|\mathcal{C}|} (\alpha_n + \beta_n)} \tag{3.21}$$

$$R^{\mu} = \frac{\sum_{i=1}^{|\mathcal{C}|} \alpha_n}{\sum_{n=1}^{|\mathcal{C}|} (\alpha_n + \gamma_n)} \tag{3.22}$$

- **Macro-promediado (M):** es el resultado de promediar los resultados obtenidos por categoría, respecto al número total de categorías $|\mathcal{C}|$ considerado. En este caso, las medidas de precisión y de cobertura con macro-promediado se expresan según las ecuaciones (3.23) y (3.24), que, a su vez, se calculan a partir de los resultados obtenidos mediante las ecuaciones (3.18) y (3.19).

$$P^{M} = \frac{\sum_{n=1}^{|\mathcal{C}|} P_n}{|\mathcal{C}|} \tag{3.23}$$

$$R^{M} = \frac{\sum_{n=1}^{|\mathcal{C}|} R_n}{|\mathcal{C}|} \tag{3.24}$$

Aunque ambas opciones de promediado permiten evaluar globalmente el funcionamiento de un algoritmo de CT a lo largo de las categorías de clasificación, existe una diferencia importante entre ellas. El micro-promediado da la misma importancia a cada uno de los documentos clasificados, mientras que el macro-promediado da el mismo peso a cada categoría. En la literatura de CT se pueden encontrar resultados utilizando ambos tipos de medias. Cuando se dispone de grandes colecciones de textos, generalmente, la más utilizada es el micro-promediado, ya que considera a todos los documentos individualmente, independientemente de la categoría a la que pertenezcan (las categorías están muy pobladas). En cambio, cuando se trabaja con colecciones más modestas o con categorías con número de documentos muy distintos (algunas muy poco pobladas), se suele recurrir al macro-promediado, ya que éste valora cada categoría por separado, independientemente del número de documentos que ésta contenga. De este modo, se minimiza el efecto de enmascaramiento que la no uniformidad de la distribución de los datos puede producir utilizando micro-promediado —p.ej. cuando se obtiene una tasa de error elevada para una categoría con pocos documentos.

Por otra parte, y para considerar tanto la precisión como la cobertura del sistema de clasificación —que persigue la maximización de ambas medidas—, se definen distintas funciones que permiten obtener una visión más general del funcionamiento del clasificador de textos. Una de las más utilizadas es la función F_β (Sebastiani, 2002), que combina los valores de precisión (P) y cobertura (R) en función del parámetro $\beta \in [0, +\infty)$ (ver ecuación (3.25)). Para los valores extremos de este intervalo, la función coincide con las medidas de precisión ($\beta = +\infty$) o de cobertura ($\beta = 0$). Entonces, según el valor que asignamos al parámetro β, se dará mayor importancia a una medida o a la otra.

$$F_\beta = \frac{(\beta^2 + 1)\, P \cdot R}{\beta^2 R + P} \tag{3.25}$$

En la literatura de CT (Sebastiani, 2002), se trabaja habitualmente con un valor de $\beta = 1$ para dar la misma importancia a las dos medidas, obteniendo la función $F_1 \in [0, 1]$ que se define como la media armónica de la precisión y la cobertura de la clasificación (ver ecuación (3.26)) —con la particularidad que penaliza valores bajos de los dos parámetros que considera, a diferencia de una media aritmética, donde, por ejemplo, valores altos de cobertura podrían compensar valores bajos de precisión.

$$F_1 = \frac{2\, P \cdot R}{R + P} \tag{3.26}$$

Bajo este mismo enfoque, existen otras funciones que permiten agrupar la información de la precisión y la cobertura de la clasificación (pero que son menos utilizadas), como por ejemplo el *breakeven point*, que es el valor por el que se cumple que $P = R$, o el *11-point average precision*, en el que se realiza un promediado de los valores de la cobertura a partir de un barrido de once valores ($[0, 1]$, con un incremento $\Delta = 0.1$) de los valores de la precisión en la clasificación (Sebastiani, 2002).

3.3.3. Red Relacional Asociativa adaptada a la CTH-MD

En este apartado del trabajo se describe la propuesta desarrollada como módulo clasificador de dominios dentro de la arquitectura de CTH-MD presentada. Este bloque se encargará de indicar el dominio (o, en un futuro, el conjunto de dominios) al que pertenece el texto de entrada a sintetizar, a partir del modelo extraído de un conjunto de textos de entrenamiento. Los requisitos de partida para el diseño de este módulo —además de perseguir la máxima eficiencia de clasificación — consistieron en:

- Diseñar un sistema de clasificación de textos adaptado a las particularidades de los datos disponibles en el marco de la síntesis multidominio, haciendo hincapié en la clasificación de textos extremadamente cortos (p.ej. 1 frase por documento).
- Incorporar un algoritmo sencillo y rápido a la arquitectura clásica de la CTH, evitando que la clasificación de dominios ralentice en exceso el proceso de conversión de texto en habla.

En comparación con los sistemas de RAH que incorporan detección temática de dominios, parece interesante que la clasificación de dominio en el ámbito de la CTH-MD vaya un poco más allá de este tipo de clasificación y contemple la secuencialidad de las palabras de los textos, tomando en consideración, así, la inherente naturaleza secuencial del habla. Además, aunque el tamaño de los datos de entrada puede ser similar al de los sistemas de RAH (consultas de entre 10 y 20 palabras en (Asami, Takezawa y Kikui, 2002)), el volumen de los datos de entrenamiento diferirá considerablemente (ver p.ej. (Iyer et al., 2002)), en el caso de que sólo se utilizan los datos del corpus de voz para el entrenamiento del clasificador —como es el caso del presente trabajo de investigación—, ya que los corpus para RAH suelen ser de tamaño muy superior a los de CTH.

Según estas premisas, y después de estudiar con detalle la formulación clásica para este tipo de sistemas de clasificación de documentos, a continuación se plantea una nueva técnica de clasificación de textos adaptada a las necesidades planteadas por la CTH-MD, haciendo énfasis en el hecho que el objetivo final de la técnica desarrollada no es tanto la de ser un excelente clasificador de textos, sino que pretende ser un módulo más de un sistema de CTH-MD que tiene como objetivo aumentar la flexibilidad de los sistemas CTH, con una calidad equivalente a la de los CTH-DR —minimizando la pérdida de calidad debida a los errores de asignación de dominio.

Para validar la propuesta de sistema de CT desarrollado, se ha implementado un sistema de CTH-MD basado en selección de unidades que trabaja sobre un corpus de voz multidominio, donde se define claramente la correspondencia entre los contenidos de cada dominio y el estilo de locución utilizado en la grabación —relaciones más complejas, quedan para futuras investigaciones (ver la sección 3.5 para más detalles). Asimismo, por el momento, la información del dominio del texto de entrada ha sido utilizada para escoger el subcorpus donde llevar a cabo la búsqueda de las unidades de voz (módulo de selección de unidades), así como, para determinar la prosodia de las mismas (módulo de modelado prosódico), según los experimentos realizados hasta el momento (ver sección 3.4).

Representación del texto

La mayor parte de la investigación en el ámbito de la clasificación automática de textos se centra en el etiquetado de grandes colecciones de documentos según su distribución temática (p.ej. en textos periodísticos: política, economía, deportes, sociedad, etc.). En este contexto, la mayor parte de sistemas de clasificación de textos (o documentos) se basan en modelar el texto como un conjunto de palabras aisladas (aproximación de la *bolsa de palabras*), como ya se ha comentado en la sección 3.3.2. De este modo, el texto queda reducido sólo a las palabras que lo constituyen, ignorando su orden o cualquiera de las relaciones que puedan existir entre ellos. Así pues, este enfoque clásico de la clasificación automática de textos deja de parametrizar dos fenómenos importantes (Sebastiani, 2002):

- **Polisemia**: una misma palabra puede aportar diferentes significados al texto según el contexto en el que se encuentre. Por lo tanto, al considerar la palabra como un elemento aislado, la capacidad de desambiguación entre sus significados gracias al contexto se elimina. En este caso, cada palabra equivale a un ítem de información, sea cuál sea su significado.

- **Sinonimia**: diferentes palabras pueden aportar el mismo significado al texto. Por lo tanto, dos textos que representen la misma información, pero expresada mediante palabras sinónimas, serán considerados, de entrada, como textos distintos.

A pesar de no contemplar esta información, esta aproximación es una de las más utilizadas para la clasificación de textos, por su sencillez y la facilidad de trabajar con grandes colecciones de textos. De todos modos, a menudo se intentan mejorar estos aspectos incorporando procesos que aborden el problema de la polisemia y la sinonimia, como por ejemplo, la reducción de las dimensiones del espacio de búsqueda basada en el algoritmo *LSI* (ver apartado "*Extracción de características*" de la sección 3.3.2).

Además de la CT temática sobre la que ha girado el grueso de la descripción realizada hasta el momento, existen otras líneas de trabajo *no temáticas* que también pertenecen al ámbito de la clasificación automática de textos. Entre ellas destacan (Stamatatos, Kokkinakis y Fakotakis, 2000): *i)* la *atribución de la autoría* de los textos (pretende determinar si un texto es de un autor u otro, con aplicaciones literarias o forenses), y *ii)* la *determinación del género* del texto (p.ej. literario, periodístico, científico, etc.), entre otras (ver (Sebastiani, 2005)). En este contexto, es necesario modelar el contenido del texto siguiendo un enfoque distinto al de los CT temáticos, ya que, independientemente de la temática del texto, el sistema de clasificación debe ser capaz de detectar el género o el autor del mismo. Para ello, resulta necesario considerar otro tipo de información adicional para parametrizar el texto, como por ejemplo, el tipo y número de artículos, preposiciones y adjetivos utilizados en el texto (análisis del género), la riqueza de vocabulario, el tamaño medio de las palabras y la longitud media de las frases (para determinar la autoría de un documento), incluyendo a menudo información morfosintáctica del texto (*part-of-speech tagging* -POS- en inglés), entre otras (Stamatatos, Kokkinakis y Fakotakis, 2000; Sebastiani, 2005). Así pues, las necesidades de los CT no temáticos o estilísticos hacen que no sea habitual preprocesar el texto

para extraer las palabras vacías ni eliminar la variabilidad morfológica de los términos, cuestiones típicas de la CT temática[14].

La clasificación de dominios para la CTH-MD se encuentra, de algún modo, entre el problema de la clasificación temática y la estilística (no temática). Por un lado, la información temática es útil para organizar los textos que forman el corpus, pero por otro lado, fijarse únicamente en esta información no parece suficiente para determinar la mejor manera de pronunciarlos (prosodia, pausado, énfasis, etc.), por lo que resulta interesante incorporar información que tome en consideración la estructura y secuencialidad del texto, con un enfoque algo distinto al utilizado en los CT no temáticos o estilísticos—ya que éstas suelen buscar el estilo global del texto, mientras que el CT desarrollado se fija en información de estilo local (secuencialidad y estructura del texto). Por lo tanto, en el contexto de la CTH, el problema de la sinonimia y la polisemia pasa a segundo término ante el problema de modelar el texto como una simple colección de palabras, ya que se deja de considerar dos factores que parecen, a priori, relevantes en el ámbito de la conversión de texto en habla, tales como:

- **Secuencialidad**: en el contexto de la CTH, el texto de entrada se convierte en una secuencia de unidades consecutivas (fonemas, difonemas, ...) a sintetizar. Por lo tanto, si el texto de entrada se representa sólo mediante palabras aisladas como elementos de clasificación, se perdería la inherente naturaleza secuencial del mensaje hablado (por ejemplo, el efecto de la coarticulación entre palabras consecutivas no sería considerado). Así pues, parece necesario considerar las relaciones entre las palabras (coocurrencias).
- **Estructura**: el modo de pronunciar un texto (estilo de locución y prosodia: ritmo, tono, energía, etc.) depende de la estructura del mismo (orden de las palabras) así como de los signos de puntuación que contiene (p.ej. una exclamación tiene un estilo de locución distinto al de una frase enunciativa, aunque ambas puedan contener las mismas palabras). Bajo este punto de vista, parece necesario considerar la secuencialidad de las palabras así como de los signos de puntuación del texto dentro del CT desarrollado.

Por los motivos que se acaban de describir, resulta imprescindible encontrar una representación del texto que permita codificar ambas informaciones, es decir, que considere tanto las palabras (y signos de puntuación) del texto como sus relaciones —a diferencia de los CT temáticos que aplican estrategias de extracción de características (descritas en la sección 3.3.2)—, permitiendo modelar el texto de entrada como una secuencia de términos interrelacionados en lugar de una simple colección de palabras aisladas. Después de analizar las estrategias más utilizadas en CT, se decidió representar los textos mediante una **Red**

[14]Por ejemplo, en el análisis de la riqueza del vocabulario se estudia el *hapax legomena* y *dislegomena*, o lo que es lo mismo, el número de palabras que aparecen sólo una o dos veces en el texto, respectivamente (Stamatatos, Kokkinakis y Fakotakis, 2000). En contraposición, estas palabras son generalmente filtradas (a partir de un determinado umbral) en el contexto de los CT temáticos, debido a su baja representatividad estadística

Relacional Asociativa (RRA, del inglés *Associative Relation Network*) (Rennison, 1994). Esta topología de red se define como un *grafo* ponderado de nodos interconectados, formado por tantos nodos como símbolos tenga que modelar, con los nodos entrelazados mediante conexiones ponderadas (ver figura 3.10) (Alías, Sevillano y Socoró, 2006).

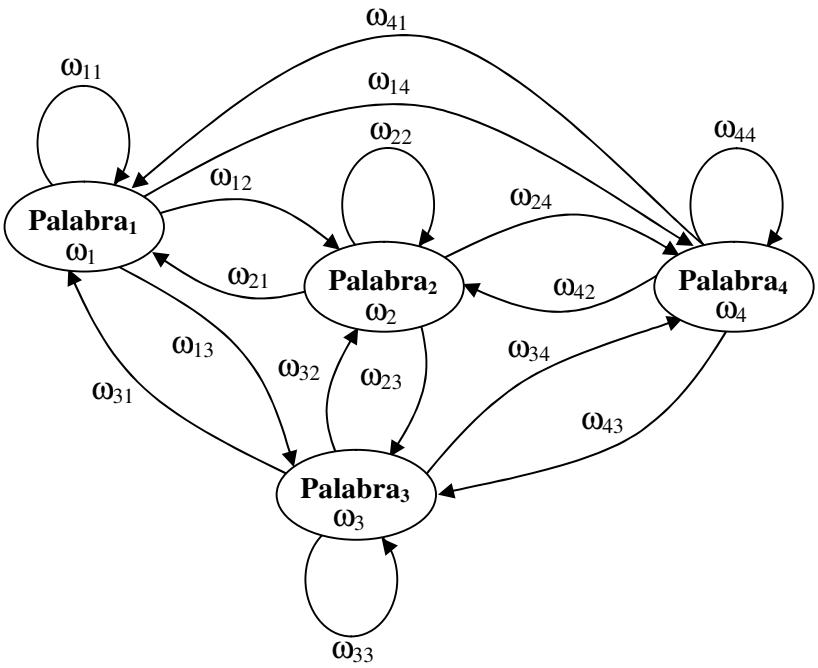


Figura 3.10: Topología básica de la Red Relacional Asociativa a nivel de palabra, inspirada en (Rennison,1994).

En el contexto de la clasificación de textos, este modelo ha sido aplicado al ámbito de la representación visual de la información de los documentos. Esta línea de investigación busca representar visualmente el contenido de una colección de documentos de la forma más eficiente e informativa posible, con el objetivo que el usuario pueda explorar los contenidos de los documentos. Esta navegación se basa en interrelacionar los documentos de la colección a partir de los símbolos (palabras clave, lugar de la acción, eventos temporales, etc.) que definen sus contenidos. En este contexto es fundamental disponer de un modelo que tome en consideración las *relaciones* entre los símbolos que definen a los documentos. En el trabajo de Rennison (1994) se introduce la representación visual denominada *Galaxy of News* (o galaxia de noticias), modelada mediante una RRA que se construye a partir de las coocurrencias de los símbolos que definen a los artículos periodísticos de la colección —relaciones que se actualizan dinámicamente a medida que el usuario navega a través de la galaxia de noticias. Así pues, las conexiones entre los símbolos dentro de la colección definirán las relaciones entre los subconjuntos de documentos sobre los que el usuario está navegando en cada instante.

A diferencia del trabajo de Rennison (1994), en el contexto de la clasificación de textos para CTH-MD, la RRA se ha utilizado para representar todas las palabras del texto modelado (y los signos de puntuación), cuyas conexiones se definen a partir del número de veces que aparecen estas palabras emparejadas dentro del texto. Es decir, se trata de una RRA a nivel de palabra, a diferencia de la RRA basada en símbolos de Rennison (1994). De este modo, la topología de RRA permite codificar el efecto de la secuencialidad de las palabras (continuidad), la temática del texto y, sobre todo, la estructura del texto (incluyendo los signos de puntuación), elementos fundamentales para el sistema de síntesis multidominio desarrollado (ver el ejemplo presentado en la figura 3.11). Para ello, esta red contendrá para cada d_k (ver figura 3.10):

- Información temática: relacionada con la importancia (peso) de cada una de las palabras (y signos de puntuación) del texto, ω_i^k.
- Información de estructura y secuencialidad:
  - La fortaleza de las relaciones entre las parejas de palabras consecutivas del texto (coocurrencias), ω_{ij}^k.
  - El sentido en la dirección de las conexiones entre las palabras del texto — generalmente, $w_{ij}^k \neq w_{ji}^k$.

  Toda esta información es fundamental para codificar la secuencialidad de las palabras del mensaje y la estructura del texto. Información que, para los sistemas de representación visual de documentos no es tan importante, ya que sólo necesitan determinar el peso de los símbolos que definen a los documentos (relacionados con los conceptos que tratan los documentos) y la fortaleza de sus relaciones (representan dependencias entre los símbolos), sin interesarles su orden ($w_{ij} = w_{ji}$ (Rennison, 1994)).

A efectos prácticos, la red está indexada mediante una lista (*array*) de las palabras que contiene, la cual permite el acceso directo a cada uno de los nodos del modelo, acelerando el acceso a la información que contiene cada RRA. Esta información corresponde a las líneas discontinuas que aparecen entre los diferentes nodos del *grafo* presentado en al figura 3.11. De esta manera, se agiliza el proceso de búsqueda a través de los nodos para obtener toda la información que contiene la red.

Parametrización del texto

Como se ha comentado, el primer módulo de la gran mayoría de clasificadores de texto es el que se encarga de preprocesar el texto (ver sección 3.3.2). Para la clasificación temática, este proceso consiste en la eliminación de las palabras que no aportan información temáticamente discriminante (lista de parada o *stop words*), la disminución de la variabilidad morfológica del texto (lematización o *stemming*) y la reducción de la dimensión del espacio de búsqueda mediante técnicas de extracción de características. Sin embargo, en el contexto

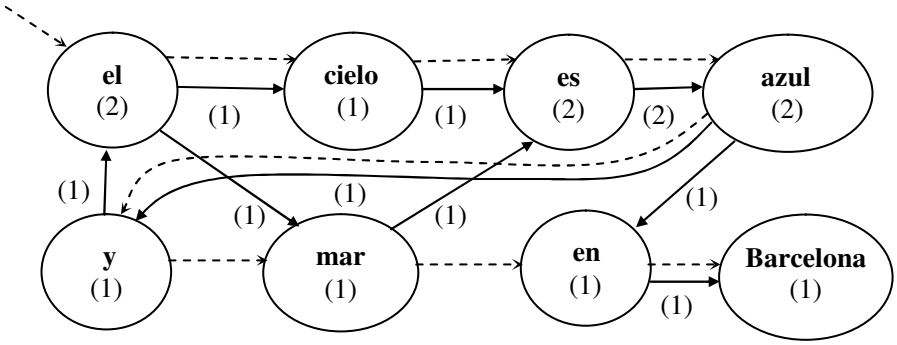


Figura 3.11: Red relacional asociativa que se obtiene a partir del texto $t = \{$*El cielo es azul y el mar es azul en Barcelona*$\}$. Sobre la red se dibuja una línea discontinua que representa el acceso directo a los nodos mediante la correspondiente lista de palabras. Además, en este ejemplo, los pesos de las palabras corresponden a sus frecuencias de aparición, mientras que las conexiones están ponderadas por sus frecuencias de coocurrencia.

de la CT para CTH-MD, surge la necesidad de analizar toda la información del texto[15], así como clasificar textos de tamaño muy reducido (p.ej. una frase —unidad típica de la CTH) en comparación con la CT temática, donde, por un lado, se busca sólo el tema del texto, y por otro, el ítem de información suele ser un documento completo. Por estos motivos, en este trabajo, el preprocesamiento del texto sólo se encargará de eliminar los caracteres *superfluos* del texto (p.ej. comillas, contrabarras, retornos de carro, etc.), ya que resulta necesario mantener toda la información aportada por el texto a clasificar para disponer de toda la estructura del texto, así como, abordar la clasificación de textos extremadamente cortos (si se eliminan palabras, el CT se quedará con demasiado poca información para llevar a cabo la clasificación). No obstante, no eliminar palabras provocará que el espacio de búsqueda tenga un mayor tamaño y un contenido más disperso de lo habitual (como se ha descrito en la sección 3.3.2). En el presente trabajo, el módulo de CT diseñado parte del texto de entrada al CTH (ver figura 3.3), una vez normalizado todo a minúsculas.

Por otra parte, como también se ha comentado en la sección 3.3.2, existen diversas aproximaciones para la ponderación del texto de entrada: desde los modelos booleanos, pasando por los probabilísticos, hasta aproximaciones más sofisticadas como las que se basan en la entropía de los ítems de información. Sin embargo, todas estas aproximaciones, debido a que parten del modelo de bolsa de palabras o *bag-of-words*, no consideran ni el orden de aparición de los términos ni como están organizados en el texto (Sebastiani, 2002). En este trabajo se consideran ponderaciones obtenidas de parámetros que incorporan esta información en la clasificación de textos (*parámetros estructurales*), además de utilizar algunos de los parámetros típicos de parametrización temática del texto (*parámetros temáticos*). Por un lado, los parámetros temáticos considerados son:

[15] Así como sucede en los sistemas de clasificación estilísticos (Stamatatos, Kokkinakis y Fakotakis, 2000).

- **Frecuencia del término × Frecuencia inversa del término en los documentos** (*term frequency × inverse document frequency, tf × idf*): considera el número de apariciones de cada término en el texto ponderado por su singularidad a lo largo de la colección de documentos de entrenamiento para la clasificación (ver sección 3.3.2). Es una de las ponderaciones más típicas de la clasificación temática de textos (Sebastiani, 2002). Esta información se incluye en cada uno de los nodos de la RRA, ponderando el peso de cada palabra del texto modelado (ω_i^k) —en el ejemplo de la figura 3.11 sólo se incluye el tf de las palabras del texto, al depender *idf* de la colección de documentos.

- **Frecuencia inversa de la palabra** (*inverse word frequency, iwf*): se trata de un nuevo parámetro introducido en el presente trabajo de investigación, que se define según la ecuación:

$$iwf_i^k = log\left(\frac{|\mathcal{P}^k|}{tf_i^k}\right), \forall\, tf_i^k > 0 \tag{3.27}$$

  donde $|\mathcal{P}^k|$ es el número de palabras del documento d_k y tf_i^k el número de veces que el término i aparece en ese documento. El parámetro iwf se puede interpretar como una aproximación local de *idf*, ya que pondera cada término según su importancia dentro de *cada* texto (o documento), en lugar de considerar su distribución a lo largo de *toda* la colección de textos (o documentos) de entrenamiento.

Por otro lado, además de considerar los **signos de puntuación**, la RRA contiene, por su propia definición, información sobre las relaciones de las palabras, que puede ser incorporada en el proceso de clasificación de textos mediante los siguientes parámetros:

- **Frecuencia de coocurrencia de las palabras** (*co-occurrence frequency, cof*): se define como el número de veces que dos palabras i y j aparecen consecutivamente dentro del texto y en este mismo orden (primero i y luego j). Este parámetro permite codificar la importancia de las relaciones entre las palabras en el texto (Rennison, 1994; Mittendorf, Mateev y Schäuble, 2000). En la RRA, este parámetro se incluye como el peso de la conexión entre dos nodos conectados del grafo (ω_{ij} en la figura 3.10). Por ejemplo, en la frase presentada en la figura 3.11, la palabras *"es"* y *"azul"* aparecen dos veces emparejadas dentro del texto, por eso su peso de conexión será $w_{es}^{azul} = 2$. Debido a que se trata de un grafo direccional, es necesario puntualizar que $w_{azul}^{es} = 0$, a diferencia de la aproximación de Rennison (1994), donde $w_{ij} = w_{j\,i}$ entre los símbolos i y j de la red.

No obstante, la arquitectura de representación del texto introducida por la RRA permite considerar además la similitud de los textos comparados en tiempo de clasificación (ver el siguiente apartado que describe las "*Medidas de similitud*" utilizadas) mediante un nuevo parámetro denominado:

- **Longitud del patrón** (*pattern length*, PL): este parámetro se define, en el contexto del presente trabajo de investigación, para analizar el impacto en la clasificación de la coincidencia de las secuencias consecutivas de palabras que aparecen en los textos comparados (texto a clasificar y modelo del dominio). El PL puede medir la longitud máxima de palabras consecutivas coincidentes entre ambos, o bien, el número acumulado de palabras consecutivas, denominado como cPL (*cumulative pattern length* o longitud del patrón acumulada). Existirán situaciones en las que PL = cPL, pero existen otras muchas situaciones en las que ambos parámetros difieren claramente (ver ejemplos de la figura 3.13 y los cálculos de las expresiones (3.39) y 3.40)). Por ejemplo, puede darse la situación que, dadas dos frases de 10 palabras cada una, ambas tengan un PL = 2 —respecto a un determinado modelo de dominio—, pero en condiciones distintas: que la $frase_1$ sólo contenga 2 palabras consecutivas dentro de la RRA del modelo, mientras la $frase_2$ dispone de 3 grupos de 2 palabras consecutivas dentro del modelo. Parece más lógico, pues, que la $frase_2$ tenga un grado de pertenencia mayor al modelo analizado que la $frase_1$. Esta cuestión se discute en el apartado de los experimentos (ver sección 3.4).

Por lo tanto, gracias a la topología de red introducida por la RRA y la inclusión de parámetros estructurales para el análisis del texto, se puede modelar el texto más allá de la aproximación clásica de bolsa de palabras —se incorpora la secuencialidad de las palabras (medida mediante las coocurrencias de las palabras y/o la longitud del patrón), además de los parámetros clásicos relacionados con la frecuencia de aparición y la distribución de los términos del texto.

Modelado del texto

Toda esta información que se extrae del texto —tanto de los textos utilizados para el entrenamiento como del texto a clasificar— tiene que ser modelada para poder ser tratada de forma eficiente por el método de clasificación. En este trabajo se ha optado por extender el modelo de espacio vectorial (MEV) definido en (Salton, 1989) (descrito en el apartado de "*Representación del texto*" de la sección 3.3.2), al incorporar información de las coocurrencias de las palabras en la representación de los textos —no obstante, también existen otras opciones para la explotación de los datos de la RRA (ver sección 3.5 para más detalles). En este caso, los documentos d_k estarán representados según la expresión (3.28), definida como generalización de la expresión (3.1), al incorporar los pesos de las coocurrencias de la colección de palabras $\mathcal{P}$ de los textos considerados (obtenidas de la tipología RRA - ver figura 3.10).

$$\vec{d_k} = \left(\omega_1^k, \omega_{11}^k, ..., \omega_{1|\mathcal{P}|}^k, ..., \omega_i^k, \omega_{i1}^k, ..., \omega_{ii}^k, ..., \omega_{i|\mathcal{P}|}^k, ..., \omega_{|\mathcal{P}|}^k, \omega_{|\mathcal{P}|1}^k, ..., \omega_{|\mathcal{P}||\mathcal{P}|}^k\right) \in \mathbb{R}^{(|\mathcal{P}|+1)\cdot|\mathcal{P}|} \tag{3.28}$$

donde $|\mathcal{P}|$ representa el número total de palabras considerado. En este caso, el espacio multidimensional $\mathbb{R}^{|\mathcal{T}|}$ definido en la expresión (3.1) pasa a ser $\mathbb{R}^{(|\mathcal{P}|+1)\cdot|\mathcal{P}|}$, ya que incorpora

todas las palabras (y signos de puntuación) junto a sus coocurrencias en el texto, de manera que el número de parámetros será $|\mathcal{T}| = (|\mathcal{P}| + 1) \cdot |\mathcal{P}|$ (cada palabra teóricamente aparecerá acompañada del resto de las —P-1— palabras y de sí misma). En este caso, a partir de la colección de documentos de entrenamiento se define el MEV de $\mathbb{R}^{\left((|\mathcal{P}|+1)\cdot|\mathcal{P}|\right)\times|\mathcal{D}^e|}$, donde los clasificadores de textos basados en RRA llevarán a cabo la representación y clasificación de documentos.

Por ejemplo, si la representación vectorial incluye la frecuencia de aparición de las palabras del texto, ($\omega_i^k = tf_i^k$) junto a sus frecuencias de coocurrencia ($\omega_{ij}^k = cof_{ij}^k$), la expresión (3.29) representaría el vector $\vec{t}$ correspondiente a la RRA del texto t del ejemplo presentado en la figura 3.11, considerando que todos los términos aparecen consecutivamente en el espacio vectorial (si no fuera así, el vector contendría posiciones nulas intercaladas, como se muestra más adelante en la expresión (3.31)) e incluyendo sólo las coocurrencias existentes entre las palabras del texto en el vector. Se puede observar como el vector $\vec{t}$ (expresión (3.30)) sigue el formato indicado por el vector $\vec{d}_k$ (3.28) para representar su contenido.

$$\vec{t} = \big(2, 1, 1, 1, 1, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1\big), \tag{3.29}$$

donde los valores numéricos de este vector corresponden a los siguientes valores parámetros:

$$\begin{aligned}\vec{t} = \big(&tf_{el}, cof_{el}^{cielo}, cof_{el}^{mar}, tf_{cielo}, cof_{cielo}^{es}, tf_{es}, cof_{es}^{azul}, tf_{azul}, cof_{azul}^{y},\\ &cof_{azul}^{en}, tf_{y}, cof_{y}^{el}, tf_{mar}, cof_{mar}^{es}, tf_{en}, cof_{en}^{Barcelona}, tf_{Barcelona}\big)\end{aligned} \tag{3.30}$$

Como se puede observar después de analizar el contenido de estos vectores, el parámetro PL (o cPL) no está incluido en la representación vectorial de los textos. Esto es debido a que se trata de un parámetro que no puede ser considerado como una dimensión más del espacio de búsqueda, ya que aporta una información global del grado de pertenencia del texto de entrada (texto a sintetizar por el CTH-MD) sobre el dominio considerado, según el número de secuencias de palabras coincidentes entre los dos textos. Por este motivo, esta información ha sido incluida como ponderación de las medidas de similitud utilizadas durante la fase de clasificación, y que se describen más adelante en el apartado de "*Explotación del sistema de clasificación*".

Entrenamiento: generación de los modelos de los dominios

El objetivo del proceso de entrenamiento del CT desarrollado consiste en la generación de las redes relacionales asociativas que modelan a los $|\mathcal{C}|$ dominios considerados, construidas a partir del conjunto de documentos de entrenamiento $\mathcal{D}^e$. Estos modelos se pueden generar mediante un proceso de entrenamiento supervisado o no supervisado ("*Algoritmo de Aprendizaje*" en el esquema de las figuras 3.5 y 3.6, donde el bloque de). El primero obtendrá los modelos de los dominios a partir de textos que han sido clasificados de forma manual en un conjunto predefinido de $\mathcal{C}$ categorías (etiquetados por un experto, normalmente). El segundo determina automáticamente la distribución de los textos en categorías a partir de la medida de similitud utilizada (S_i), definiendo *a priori* el número de categorías

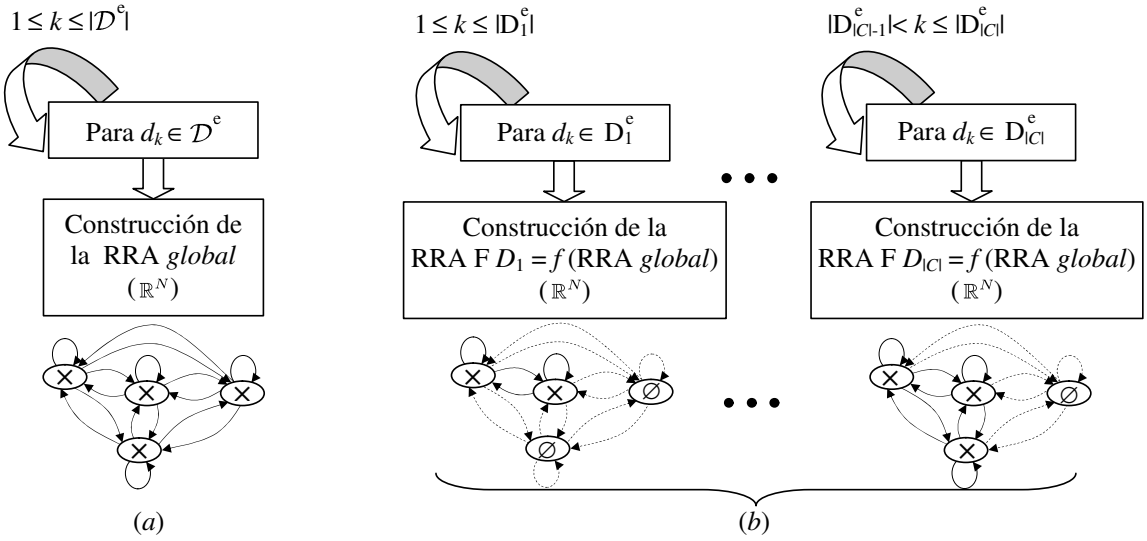


Figura 3.12: Proceso de generación de (a) la RRA *global* y (b) las RRA F de dominio, desde D_1 hasta $D_{|\mathcal{C}|}$, referenciadas a la RRA *global* de dimensión $\mathbb{R}^N$ y construidas a partir de los documentos de entrenamiento de cada dominio $\mathcal{D}^e = \{D^e_1, \ldots, D^e_{|\mathcal{C}|}\}$. En los grafos, "×" indica nodo ocupado, mientras "∅" representa los nodos vacíos. Además, las conexiones discontinuas denotan coocurrencias inexistentes.

deseado ($|\mathcal{C}|$) o dejándolo abierto según un criterio heurístico (p.ej. basado en una medida de entropía de los grupos). De esta manera, se puede determinar de forma no supervisada cómo agrupar los datos de la colección de entrenamiento ($\mathcal{D}^e$), siguiendo la filosofía clásica de un algoritmo de *clustering* (p.ej. ver (Sevillano, Alías y Socoró, 2004)). En este trabajo, centrado en la CT, se ha aplicado el proceso supervisado de entrenamiento del clasificador de textos, que parte de la asociación directa entre los textos de entrenamiento $\mathcal{D}^e$ y cada uno de los dominios (categorías) $\mathcal{C}$, como base para la construcción de la RRA de cada dominio. Como resultado de este proceso se obtendrá la red relacional asociativa de cada uno de los dominios que forman parte del corpus de voz multidominio.

Para poder disponer de un único espacio de representación y comparación único y coherente para todos los modelos RRA de dominio, resulta necesario generar previamente una RRA *global* (de $\mathbb{R}^N$) que contemple todas las palabras de cada uno de los dominios, así como sus relaciones, totalizando N parámetros o dimensiones del MEV considerado (ver figura 3.12(a)). Teóricamente el espacio vectorial potencial definido a partir de los $|\mathcal{T}|$ términos (parámetros) obtenidos al considerar las apariciones y las coocurrencias de las palabras del texto sería $\mathbb{R}^{|\mathcal{P}|(|\mathcal{P}|+1)}$. No obstante, debido a que se parte de un conjunto finito de documentos de entrenamiento, la dimensionalidad del espacio será $N \leq |\mathcal{P}|(|\mathcal{P}|+1)$, ya que, generalmente, resultará imposible observar todas las coocurrencias de las palabras (todas contra todas) por no disponer de suficientes datos de entrenamiento y por las restricciones del propio lenguaje. Así pues, el grafo global irá aumentando de tamaño a medida que vaya

incorporando nuevos textos a la RRA que lo modela, considerando el orden de los mismos en su construcción y activando sólo las coocurrencias observadas. La RRA global simplemente define la identidad y el orden de comparación de los datos, definiendo las dimensiones del MEV utilizado, sin considerar la información de la ponderación de los parámetros considerados en el modelado de los textos.

Una vez generada la red global, se procede a modelar los textos de cada uno de los dominios según ésta, obteniendo una RRA *Full*[16] (de $\mathbb{R}^N$) para cada dominio o RRA F D_n (categoría c_n, $n = 1 \dots |\mathcal{C}|$) (ver figura 3.12($b$)). El proceso de entrenamiento finaliza después de representar cada RRA F de dominio mediante un vector patrón[17] ($\vec{p_n} \in \mathbb{R}^N$, $n = 1 \dots |\mathcal{C}|$), obtenido por los $d_k \in ||D_n^e|$ documentos de entrenamiento que le corresponden (ver figura 3.12(b)). De este modo, cada vector patrón $\vec{p_n}$ contendrá toda la información de ese dominio en un único vector representado según el MEV definido por la RRA global (ver ejemplo de la tabla 3.5) —consiguiendo una representación consistente de los datos a lo largo de las RRA F D_n de los dominios.

Siguiendo con el ejemplo de la figura 3.11, si esta RRA se toma como representación de la RRA global de comparación para un texto $t_1 = \{el\ cielo\ en\ Barcelona\}$ (asociado a un cierto dominio, por ejemplo), quedaría representado según la expresión (ver expresión 3.31).

$$\vec{t_1} = \big(1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1\big) \tag{3.31}$$

donde estos valores numéricos corresponden a los siguientes parámetros, según la expresión 3.30 que define el orden de comparación de los parámetros considerados:

$$\begin{aligned}\vec{t_1} = \big(&tf_{el} = 1, cof_{el}^{cielo} = 1, cof_{el}^{mar} = 0, tf_{cielo} = 1, cof_{cielo}^{es} = 0, tf_{es} = 0,\\ &cof_{es}^{azul} = 0, tf_{azul} = 0, cof_{azul}^{y} = 0, cof_{azul}^{en} = 0, tf_{y} = 0, cof_{y}^{el} = 0,\\ &tf_{mar} = 0, cof_{mar}^{es} = 0, tf_{en} = 1, cof_{en}^{Barcelona}, tf_{Barcelona} = 1\big)\end{aligned} \tag{3.32}$$

Por el hecho de no eliminar palabras (no se aplica lista de parada) ni reducir la flexión de las mismas (no se extrae el radical) —para poder, así, incorporar las relaciones entre palabras y su secuencialidad—, el espacio multidimensional definido por los vectores que caracterizan la red global tendrá un tamaño considerable (proporcional al número de palabras de los textos de la colección de dominios si se trabaja sin coocurrencias, y como máximo de orden cuadrático si aparecen todas las coocurrencias de todas las palabras). Este será uno de los factores clave del sistema de CT desarrollado, tal y como se describe a lo largo del resto de secciones del presente capítulo.

[16] Sus componentes seguirán el orden dictado por el MEV completo (global) o *Full space*, ya que contempla todas las palabras —y signos de puntuación— junto a sus coocurrencias existentes en todo el corpus de textos de entrenamiento, de ahí su nombre (*full*≡ completo).

[17] Representación vectorial de toda la información del conjunto de documentos que constituye la RRA F D_n, a diferencia de otras aproximaciones, donde se representa el conjunto de vectores por un único representante o *centroide* (documento promedio).

Jerarquización: Una vez obtenidos los modelos de dominio, el sistema de clasificación de textos puede necesitar de la organización jerárquica de los mismos —según las características del sistema de CTH-MD desarrollado. Para ello, es necesario agrupar los dominios de contenido similar a distintos niveles (p.ej. macro-dominios), obteniendo una base de datos multidominio estructurada jerárquicamente (ver figura 3.3). Los objetivos fundamentales de este proceso son:

1. Obtener niveles intermedios de clasificación.
2. Estructurar organizadamente la información del corpus multidominio.
3. Definir una arquitectura lo más flexible posible.

Gracias a este proceso, el sistema de CT puede disponer de distintos niveles para la organización de los datos del corpus. De esta manera, si durante el proceso de asignación de dominios, el CT encuentra distintas soluciones posibles, es decir, existen diferentes dominios en los que se puede clasificar el texto de entrada, el CT puede pasar a un nivel jerárquico superior para analizar si el texto pertenece con más claridad a algún macro-dominio (en la línea de otros sistemas orales multidominio (Lane et al., 2005)) —no obstante, esta cuestión queda fuera del alcance del presente trabajo de investigación. Por ejemplo, si el CT no es capaz de discernir entre los modelos básicos a la hora de clasificar un texto (tienen grados de similitud muy parecidos), la jerarquía permite subir un nivel en la estructura multidominio (ver figura 3.3) para, o bien, repetir la búsqueda, o bien, asignar el texto al macrodominio correspondiente a los n dominios con mayor grado de pertenencia. Este proceso es iterativo y finaliza cuando se llega al nivel superior de la jerarquía, momento, en el que el CTH-MD, o bien, consideraría todas las unidades del corpus, o bien, utilizaría sólo información de propósito general, según esté diseñado.

En el presente trabajo de investigación, se ha realizado un estudio preliminar de la jerarquización de textos en el contexto de la CTH-MD. Concretamente, se ha abordado el problema mediante el **Análisis en Componentes Independientes** (*Independent Components Analysis* o ICA, en inglés) del contenido del corpus en colaboración con Xavier Sevillano (Alías et al., 2003; Sevillano, Alías y Socoró, 2004). ICA es una técnica estadística de propósito general fundamentada en un modelo generativo de variables latentes (Hyvärinen, Karhunen y Oja, 2001). El modelo lineal de 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) (para relaciones más complejas, p.ej. no lineales, ver (Hyvärinen, Karhunen y Oja, 2001)). Así pues, los algoritmos ICA hallarán de forma no supervisada las CI de los datos, así como los coeficientes de su combinación lineal, partiendo únicamente de las observaciones disponibles —generalmente, un mínimo de tantas como número de componentes a detectar.

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 temáticos* (Isbell y Viola, 1999; Kaban y Girolami, 2000). Es decir, un documento es producto de la interacción de un conjunto de variables ocultas independientes que lo generan, variables que estan

asociadas íntimamente a la temática o temáticas que trata el documento. El uso de ICA como CT está muy vinculado a la técnica de indexado de la semántica latente de los textos o *latent semantic indexing* (LSI) (Deerwester et al., 1990). Esta técnica proyecta los documentos en un espacio ortogonal de dimensionalidad reducida, extrayendo las K direcciones principales del espacio que mejor representan la información más significativa de los datos. La aplicación de ICA sobre el espacio LSI permite descubrir las K temáticas independientes que generaron los documentos, lo que permite su clasificación (Kaban y Girolami, 2000). Según la relación entre K y el número de categorías $|\mathcal{C}|$ en las que un experto ha dividido la colección, se pueden obtener los distintos niveles jerárquicos del corpus, mediante un proceso semisupervisado[18]:

- Si $K = |\mathcal{C}|$, se obtienen tantos grupos de datos (o *clusters*) como categorías, clasificando los documentos en los ámbitos temáticos correspondientes a las categorías definidas a priori.
- 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 —en este caso, más parecidas temáticamente.
- Si $K > |\mathcal{C}|$, los textos pertenecientes a cada una de las categorías pueden separarse en subcategorías. A veces, el número de temáticas presentes en una colección de documentos puede ser mayor que el considerado por el experto (Kaban y Girolami, 2000). Es decir, dentro de una temática principal pueden existir diversas subtemáticas que no han sido detectadas de antemano y que tienen la suficiente entidad como para ser consideradas como subdominios. Así pues, en este contexto, ICA permite descubrir subcategorías que tienen una cierta consistencia (o homogeneidad) temática. La posterior supervisión experta de los documentos que corresponden a estas subcategorías permite la asignación de una etiqueta significativa del conjunto de documentos que representan. Por ejemplo, en textos periodísticos de política pueden existir subdominios que traten de política internacional o nacional.

Parece lógico pensar, pues, que cuanto más homogénea sea la temática de un determinado dominio del corpus, sus textos tenderán a agruparse en un único grupo, que 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 descubiertos cuando se incremente la dimensionalidad del análisis. Estas subtemáticas deberán ser analizadas a posteriori para validar la coherencia de sus contenidos, junto a la distribución de los datos. En este contexto, será necesario aplicar sucesivamente el algoritmo ICA con valores crecientes de K para obtener la estructura jerárquica del corpus, analizando cada vez la agrupación de los textos pertenecientes a cada dominio. No obstante, se deja para trabajos futuros el análisis exhaustivo de técnicas de jerarquización de los

[18]Una vez aplicado el algoritmo ICA, que es un proceso no supervisado, la asignación de etiquetas por componente independiente se realiza de forma supervisada, asociando cada componente independiente a la categoría de la mayoría de documentos para los que esta componente se ha activado.

contenidos del corpus, junto a su combinación con el análisis de su estilo introducido por la representación del texto basada en la RRA descrita. Asimismo, se deja para un trabajo futuro la explotación de esta jerarquía en el contexto de la CTH-MD, ya que en el presente trabajo los experimentos de síntesis se han realizado sobre un corpus que no está organizado jerárquicamente (ver sección 3.4.2).

Test: explotación del sistema de clasificación

Una vez generados los modelos de los dominios a partir de la red global de representación de los textos (RRA F $D_n = f(RRA\ global)$, $n = 1 \dots |\mathcal{C}|$, en la figura 3.12) y obtenidos sus vectores patrón $\vec{p}_n \in \mathbb{R}^N$, se procede a explotar el sistema de clasificación de textos. Para ello, primero, será necesario modelar también el texto a clasificar t_k según la RRA global, obteniendo el vector $\vec{t}_k \in \mathbb{R}^N$ (ver el ejemplo de la tabla 3.5), tomando el formato definido por el MEV construido a partir de la red global. A continuación, se compara el vector $\vec{t}_k$ con cada uno de los vectores patrón que representan los dominios considerados, utilizando un medida de similitud S_i basada en la distancia del coseno (ver ecuación (3.11)) —las variantes definidas se describen a continuación en las ecuaciones (3.34), (3.35) o (3.37). Finalmente, el texto de entrada t_k se asignará a la categoría (dominio $\hat{D}_n$) respecto a la que presente una menor distancia aplicando la ecuación (3.33) —es decir, se trata de una asignación única o *hard classification*[19]. En este caso, pues, una menor distancia equivaldrá a una mayor proximidad (similitud) de los datos comparados.

$$\hat{D}_n = \underset{1 \le n \le |\mathcal{C}|}{Argmin} \Big(S_i(t_k, D_n) \Big) \tag{3.33}$$

A continuación se presentan las distintas medidas de similitud estudiadas, incluyendo aquellas diseñadas para incluir información de la secuencialidad del texto para mejorar la eficiencia del CT en el contexto de aplicación en el que se ha definido: la CTH-MD. Este estudio se verá reflejado en la sección 3.4, donde se presentan los resultados obtenidos para las distintas distancias consideradas.

Designación de dominio y medidas de similitud: Para poder llevar a cabo la tarea de la clasificación de los textos (designación de dominio), es necesario definir la medida que evalúe el grado de similitud entre el vector del texto de entrada ($\vec{t}_k$) y el vector patrón del dominio ($\vec{p}_n$). Dado que el modelo de clasificación de textos escogido se basa en un MEV, la primera de las medidas de similitud considerada es la distancia del *coseno*, medida típica de la clasificación de textos mediante representación vectorial (Sebastiani, 2002). Para el caso del algoritmo de CT diseñado, este cálculo se realiza según la ecuación (3.34), donde N corresponde a la dimensión del espacio vectorial global ($\mathbb{R}^N$).

[19]Se deja para un trabajo futuro la asignación múltiple o *soft classification* (Sebastiani, 2002), mediante la cual se podría explotar plenamente la estructura jerárquica del corpus, como se ha descrito anteriormente.

$$S_1\big(t_k, D_n\big) = \frac{< \vec{t}_k, \vec{p}_n >}{\|\vec{t}_k\| \cdot \|\vec{p}_n\|} = \frac{\sum_{i=1}^{N} \left(v_i^{t_k} \cdot p_i^n \right)}{\sqrt{\sum_{j=1}^{N} \left(v_j^{t_k} \right)^2} \sqrt{\sum_{j=1}^{N} \left(p_j^n \right)^2}} \tag{3.34}$$

A partir de esta distancia, en el presente trabajo de investigación, se define una nueva medida que incorpora la longitud del patrón (PL) como ponderación de la distancia del coseno (ver ecuación (3.35)). De este modo, la información global de la pertenencia del texto t_k al dominio D_n, codificada por este parámetro y complementaria a la de las coocurrencias, puede ser considerada en el proceso de designación del dominio del texto de entrada.

$$S_2\big(t_k, D_n\big) = \mathrm{PL}(\vec{t}_k, \vec{p}_n) \cdot S_1\big(t_k, D_n\big) \tag{3.35}$$

donde, dada la secuencia de índices de $I(\vec{t}_k) = \{i_1^k, i_2^k, ..., i_{|\mathcal{P}^k|}^k\}$ de las palabras del texto t_k ordenadas según el MEV de representación común considerado[20] y siendo $|\mathcal{P}^k|$ el número de palabras —y signos de puntuación— del texto t_k, $\mathrm{PL}(\vec{t}_k, \vec{p}_n)$ puede definirse como:

$$\mathrm{PL}(\vec{t}_k, \vec{p}_n) = \frac{1}{|\mathcal{P}^k| - 1}\, max\Big\{ \mathcal{O} \;\Big|\; \prod_{r=l}^{l+\mathcal{O}-1} (\omega^{\vec{t}_k}_{i_r^k, i_{r+1}^k}) \neq 0, \forall\, 1 \leq l \leq |\mathcal{P}^k| - \mathcal{O} \Big\} \tag{3.36}$$

donde $\mathcal{O}$ será el índice de cuenteo del número de coocurrencias consecutivas coincidentes entre el texto a clasificar y el dominio comparado, es decir, $0 \leq \mathcal{O} \leq |\mathcal{P}^k| - 1$, y $\omega^{\vec{t}_k}_{ij}$ es la frecuencia de coocurrencia entre las palabras i y j del texto de entrada t_k una vez representado en el espacio de comparación común.

Posteriormente, se modificó la medida de similitud S_2 sustituyendo el parámetro PL por su versión acumulada, cPL (ver ecuación (3.37)), con el objetivo de mejorar los resultados obtenidos mediante S_2, tal y como se describe en el apartado que describe los experimentos (ver sección 3.4).

$$S_3\big(t_k, D_n\big) = \mathrm{cPL}(\vec{t}_k, \vec{p}_n) \cdot S_1\big(t_k, D_n\big) \tag{3.37}$$

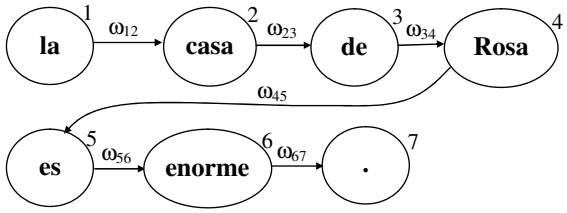
donde $\mathrm{cPL}(\vec{t}_k, \vec{p}_n)$ se define a partir del número acumulado de segmentos de palabras que ocurren en el mismo orden en el texto t_k respecto al dominio D_n de comparación, una vez representados en un espacio de comparación común, representado en el cálculo por sus vectores $\vec{t}_k$ y $\vec{p}_n$:

$$\mathrm{cPL}(\vec{t}_k, \vec{p}_n) = \frac{\sum_{i,j=1}^{|\mathcal{P}^k|} \omega^{\vec{t}_k}_{ij}}{|\mathcal{P}^k| - 1} \tag{3.38}$$

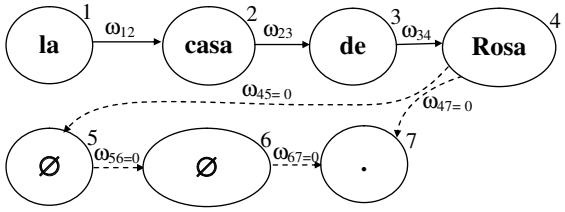
[20]Por este motivo, la secuencia de índices se denota por $I(\vec{t}_k)$, ya que el orden de comparación de las palabras sólo se obtiene una vez representado el texto en la RRA global, generando su representación vectorial $\vec{t}_k$.

Como se puede observar de las ecuaciones (3.36) y (3.38), ambos cálculos se normalizan respecto al número total de coocurrencias de las palabras del texto de entrada, por lo que $0 \leq \{\mathrm{PL}, \mathrm{cPL}\} \leq 1$. Gracias a la inclusión de estos parámetros en la medida de similitud, se puede considerar un factor más para determinar el grado de pertenencia del texto de entrada respecto al modelo de cada uno de los dominios, incorporando información global en lo que se refiere a la secuencialidad del texto y su estructura (p.ej. signos de puntuacion).

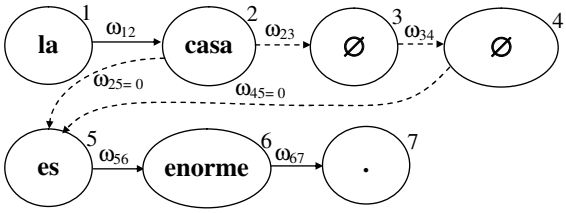
En la figura 3.13 se presenta un pequeño ejemplo de la representación de distintos textos sobre una RRA de dominio (RRA D) definida a partir de una única frase. Como se puede observar en la figura, todo aquello que no ha sido observado durante la fase de entrenamiento (palabras y/o coocurrencias entre palabras del dominio) no está representado en las RRA de los textos t_k a clasificar —ya que la RRA de dominio es la que fija el espacio de comparación. Por ejemplo, la unión "*Rosa-.*"del texto t_1 en la figura 3.13(b), la unión "*casa-es*"de t_2 en la figura 3.13(c) o la palabra "*blanca*" y sus coocurrencias de t_3 en la figura 3.13(d) no quedan representadas en las RRA que corresponden a cada uno de los t_k a clasificar.



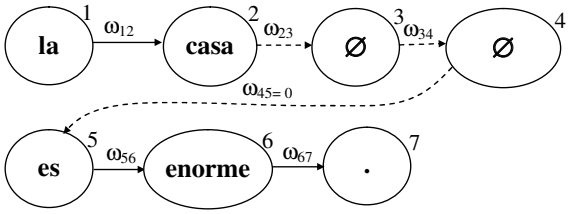
(a) RRA del dominio D ={*La casa de Rosa es enorme.*} o RRA D.



(b) RRA del texto t_1 ={*La casa de Rosa.*} según RRA D.



(c) RRA del texto t_2 ={*La casa es enorme.*} según RRA D.



(d) RRA del texto t_3 ={*La casa blanca es enorme.*} según RRA D.

Figura 3.13: Representación, según la RRA de dominio representada en (a), de tres ejemplos de textos a clasificar (b), (c) y (d) con distintas casuísticas de representación. "∅" representa los nodos vacíos de los grafos de los textos, mientras las conexiones discontinuas indican coocurrencias inexistentes. En este ejemplo, todas las coocurrencias activas son $\omega_{ij}^k = 1$.

Seguidamente, en la tabla 3.4 se presenta el cálculo de PL y cPL para el texto t_2 y el dominio D de la figura 3.13 —cálculo también aplicable al texto t_3 de la misma figura, ya que, aunque se trata de textos distintos, una vez representados según la RRA D, quedan

Tabla 3.4: Ejemplo del cálculo de PL y cPL para el texto t_2 y el dominio D de los ejemplos de la figura 3.13, dada la secuencia de índices $I(\vec{t}_2) = \{1, 2, 5, 6, 7\}$ sobre la RRA del dominio y considerando que el texto t_2 contiene 5 palabras.

$$
\begin{aligned}
l &= 1 \rightarrow \mathcal{O} = 1 \quad \prod_{r=l}^{l+\mathcal{O}-1} (\omega^{t_2}_{i^2_r, i^2_{r+1}}) = \omega^{t_2}_{12} = 1 \\
&\qquad \mathcal{O} = 2 \quad \prod_{r=l}^{l+\mathcal{O}-1} (\omega^{t_2}_{i^2_r, i^2_{r+1}}) = \omega^{t_2}_{12} \cdot \omega^{t_2}_{25} = 0 \\
l &= 2 \rightarrow \mathcal{O} = 1 \quad \prod_{r=l}^{l+\mathcal{O}-1} (\omega^{t_2}_{i^2_r, i^2_{r+1}}) = \omega^{t_2}_{25} = 0 \\
l &= 3 \rightarrow \mathcal{O} = 1 \quad \prod_{r=l}^{l+\mathcal{O}-1} (\omega^{t_2}_{i^2_r, i^2_{r+1}}) = \omega^{t_2}_{56} = 1 \\
&\qquad \mathcal{O} = 2 \quad \prod_{r=l}^{l+\mathcal{O}-1} (\omega^{t_2}_{i^2_r, i^2_{r+1}}) = \omega^{t_2}_{56} \cdot \omega^{t_2}_{67} = 1
\end{aligned}
$$

$$
\mathrm{PL}(\vec{t}_2, \vec{D}) = \frac{1}{5-1}\, max(\mathcal{O} = \{1, 0, 2\}) = \frac{1}{5-1} \cdot 2 = \frac{1}{2} \tag{3.39}
$$

$$
\mathrm{cPL}(\vec{t}_2, \vec{D}) = \frac{\omega^{t_2}_{12} + \omega^{t_2}_{25} + \omega^{t_2}_{56} + \omega^{t_2}_{67}}{5-1} = \frac{1+0+1+1}{4} = \frac{3}{4} \tag{3.40}
$$

representados por la misma red relacional. En la tabla se detalla todo el proceso seguido para obtener el valor de PL a partir de la ecuación (3.36), así como, el resultado de cPL calculado según la ecuación (3.38). El resultado obtenido (expresiones (3.39) y (3.40)) muestra el distinto análisis de la secuencialidad de las palabras considerada por ambos parámetros. Siguiendo el mismo proceso, se puede obtener para el texto t_1 de la figura 3.13 que PL = cPL = $\frac{3}{4}$, ya que en este caso sólo existe una secuencia de cuatro palabras coincidentes (tres coocurrencias consecutivas). Nótese que, aunque coinciden todas las palabras del texto, la longitud del patrón no es 1, debido a que la unión "*Rosa-.*"no está representada en la RRA del dominio, es decir, no tienen el mismo patrón de signos de puntuación[21].

Red Relacional Asociativa Reducida (RRA R)

Uno de los factores críticos del modelado de textos según RRA F es la complejidad computacional del mismo, ya que para clasificar cualquier texto es necesario recorrer antes toda la RRA global para representarlo de forma coherente con los datos de entrenamiento. Además, por el hecho de no eliminar palabras (no se aplica lista de parada), no reducir

[21]De este modo, se está considerando la diferencia de pronunciación que tendrán los textos comparados, cuestión fundamental para la clasificación de textos en el ámbito de la CTH-MD presentado.

Tabla 3.5: Ejemplo ilustrativo de la representación de los vectores patrón de dominio $\vec{p}_n$ (RRA F $D_n, n = 1 \ldots |\mathcal{C}|$) y del texto a clasificar t_k según la RRA global, dados tres dominios D_1, D_2 y D_3 distintos. Los símbolos $\{\omega_A^n, \omega_B^n, \ldots, \omega_Z^n\}$ representan los pesos correspondientes a los términos de los textos modelados.

Datos	**Términos**	**Representación global**
MEV	$\{A, B, C, D, E, F, G, H, I, J\}$	$(\omega_A, \omega_B, \omega_C, \omega_D, \omega_E, \omega_F, \omega_G, \omega_H, \omega_I, \omega_J)$
D_1	$\{A, B, C, D, E, F, G\}$	$\vec{p}_1 = (\omega_A^1, \omega_B^1, \omega_C^1, \omega_D^1, \omega_E^1, \omega_F^1, \omega_G^1, 0, 0, 0)$
D_2	$\{A, B, C, E, H, I, J\}$	$\vec{p}_2 = (\omega_A^2, \omega_B^2, \omega_C^2, 0, \omega_E^2, 0, 0, \omega_H^2, \omega_I^2, \omega_J^2)$
D_3	$\{A, B, H, D\}$	$\vec{p}_3 = (\omega_A^3, \omega_B^3, 0, \omega_D^3, 0, 0, 0, \omega_H^3, 0, 0)$
t_k	$\{C, A, Z\}$	$\vec{t}_k = (\omega_A^k, 0, \omega_C^k, 0, 0, 0, 0, 0, 0, 0)$

la flexión de las mismas (no se extrae el radical), e incluir sus coocurrencias, el espacio multidimensional definido por la RRA global tendrá un tamaño considerable en comparación con el texto a clasificar, lo que reduce la separabilidad entre los distintos dominios (como se verá). Por lo tanto, la representación vectorial $\vec{t}_k$ del texto (o documento) a clasificar será típicamente muy dispersa (muchos elementos nulos), cuestión que puede provocar una reducción de la capacidad de discriminación de los vectores patrón de los dominios, afectando negativamente a la eficiencia del sistema de clasificación de textos. Con el objetivo de minimizar el coste computacional y mejorar la tasa de clasificación del sistema de CT basado en RRA F —cuestiones clave en el contexto de la CTH-MD— se propone una nueva estrategia de clasificación basada en la representación de los textos sobre una Red Relacional Asociativa Reducida (RRA R).

La idea fundamental de la estrategia de CT basada en RRA R consiste en sustituir el espacio de comparación de los textos definido por la RRA global por el espacio vectorial definido a partir del texto de entrada t_k. Por lo tanto, la RRA generada a partir de t_k (en adelante, RRA R) es la que marcará el orden de comparación de los datos, dando lugar a un nuevo modelo de espacio vectorial denominado como MEV$'$. En este caso, su contenido estará representado a partir del vector $\vec{t}'_k$ dentro del espacio vectorial $\mathbb{R}^{L^k}$ definido por la RRA R (siendo L^k el número de parámetros considerados -palabras y coocurrencias- del texto t_k, con $L^k \ll N$, típicamente). Así pues, durante el proceso de clasificación, los dominios D_n deberán representarse según la RRA R ($\mathbb{R}^{L^k}$) (ver ejemplo de la tabla 3.6). Es decir, en el proceso de construcción de las RRA D_n de la figura 3.12, se sustituye la RRA global por la RRA R generada a partir del texto de entrada t_k como referencia para la representación de su contenido. La designación de la categoría del texto de entrada se realiza aplicando también la ecuación (3.33), pero sustituyendo, en este caso, el vector $\vec{t}_k$ por el vector $\vec{t}'_k$ y los vectores patrón $\vec{p}_n$ por sus versiones reducidas $\vec{p}\,'_n$.

Gracias a utilizar el modelo RRA R, se consigue minimizar la complejidad del proceso de representación de los datos en la fase de explotación del CT, al sustituir el tiempo necesario para representar el texto t_k sobre el espacio definido por red global ($\mathbb{R}^N$) (es necesario recorrer toda la RRA global para obtener el orden de comparación de los datos),

Tabla 3.6: Representación de los datos de la tabla 3.5 en el MEV′ definido sobre $\mathbb{R}^3$ ($L^k = 3$) definido por la RRA R del texto a clasificar t_k. Los símbolos $\{\omega_A^n, \omega_B^n, \ldots, \omega_Z^n\}$ representan los pesos correspondientes a los términos de los textos modelados.

Datos	**Términos**	**Representación reducida**
MEV′	$\{C, A, Z\}$	$(\omega_C, \omega_A, \omega_Z)$
t_k	$\{C, A, Z\}$	$\vec{t'}_k = (\omega_C^k, \omega_A^k, \omega_Z^k)$
D_1	$\{A, B, C, D, E, F, G\}$	$\vec{p}\,'_1 = (\omega_C^1, \omega_A^1, 0)$
D_2	$\{A, B, C, E, H, I, J\}$	$\vec{p}\,'_2 = (\omega_C^2, \omega_A^2, 0)$
D_3	$\{A, B, H, D\}$	$\vec{p}\,'_3 = (0, \omega_A^3, 0)$

por el coste de representar cada dominio D_n en el espacio vectorial definido por la RRA R ($\mathbb{R}^{L^k}$), que será, generalmente, mucho menor. No obstante, no hay que olvidar que la representación de los textos mediante la RRA R no es más que una aproximación de la RRA F, cuya justificación teórica se pasa a describir a continuación.

Justificación algebraica de la RRA R: La reducción de dimensionalidad del modelo de espacio vectorial utilizado para representar los textos que implica la estrategia RRA R respecto a la RRA F puede justificarse desde un punto de vista algebraico. Por un lado, la RRA F puede representarse sobre el espacio vectorial $\mathbb{R}^N$, espacio al que pertenecen todos los vectores del conjunto de documentos de entrenamiento $\mathcal{D}^e$. En él se definen tanto los vectores patrón $\vec{p}_n$ de cada dominio, como los vectores $\vec{t}_k$ de los textos a clasificar —con M^k componentes activas[22] y $(N - M^k)$ nulas, donde $M^k \ll N$ generalmente (ver ejemplo de la tabla 3.5, con $M^k = 2$ y $N = 10$). En este contexto, el grado de pertenencia del texto a clasificar respecto a los dominios se calculará utilizando vectores de $\mathbb{R}^N$ (usando las ecuaciones basadas en la distancia del coseno definidas anteriormente). Asimismo, la RRA R también puede representarse algebraicamente en un espacio vectorial $\mathbb{R}^{L^k}$, también con $L^k \ll N$ y además con $L^k \geq M^k$, habitualmente, debido a las palabras de t_k que no aparecen en la RRA global (no observadas durante la fase de entrenamiento).

Por otro lado, dentro del espacio vectorial $\mathbb{R}^N$ definido por la RRA global, también se puede definir un subespacio vectorial $V \subset \mathbb{R}^N$, a partir de una base $B = \{\vec{b_1}, \vec{b_2}, \ldots, \vec{b}_{M^k}\}$ de M^k vectores ortogonales definidos por las componentes activas del vector $\vec{t}_k$. Esta base estará formada por los vectores de la base canónica de $\mathbb{R}^N$ que ocupan las posiciones no nulas del vector $\vec{t}_k$ (ver ejemplo de la tabla 3.7). Utilizando esta base, se puede representar el vector patrón $\vec{p}_n$ sobre el subespacio V como la mejor aproximación de este vector ($\hat{\vec{p}}_n$), utilizando el criterio de los mínimos cuadrados, mediante la ecuación (3.41). Cualquier otra proyección (no ortogonal) de los vectores $\vec{p}_n$ dentro del subespacio V, presentará un error de aproximación mayor, calculado como $\vec{e} = \vec{p} - \hat{\vec{p}}$ (ver figura 3.14).

[22]Siendo M^k el número de parámetros considerados (palabras y relaciones de palabras) del texto de entrada que aparecen en la colección de documentos de entrenamiento.

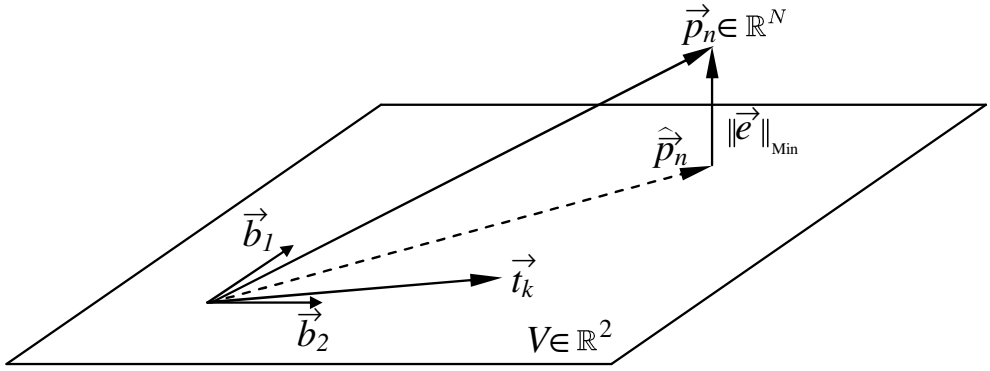


Figura 3.14: Resultado de la proyección del vector patrón $\vec{p}_n$, definido según el espacio vectorial $\mathbb{R}^N$ definido por la RRA F, sobre el subespacio vectorial $V \subset \mathbb{R}^2$, engendrado por la base $B = \{\vec{b}_1, \vec{b}_2\}$ que queda definida a partir de las $M^k = 2$ componentes activas del vector $\vec{t}_k$ en $\mathbb{R}^2$. Se obtiene el vector patrón aproximado $\hat{\vec{p}}_n$ según el criterio de los mínimos cuadrados, resultando un error de aproximación $||\vec{e}||$ mínimo.

Tabla 3.7: Representación de los datos de la tabla 3.5 en el subespacio vectorial V generado a partir de la base ortogonal $B = \{\vec{b}_1, \vec{b}_2\}$ definida por las $M^k = 2$ componentes activas de $\vec{t}_k$, representado según el MEV global. Los símbolos $\{\omega_A^n, \omega_B^n, \ldots, \omega_Z^n\}$ representan los pesos correspondientes a los términos de los textos modelados.

Base del Subespacio Vectorial V		
$\vec{b_1} = (1,0,0,0,0,0,0,0,0,0)$		
$\vec{b_2} = (0,0,1,0,0,0,0,0,0,0)$		
Datos	**Términos**	**Componentes en V**
D_1	$\{A,B,C,D,E,F,G\}$	$\hat{\vec{p_1}} = (\omega_A^1, \omega_C^1)$
D_2	$\{A,B,C,E,H,I,J\}$	$\hat{\vec{p_2}} = (\omega_A^2, \omega_C^2)$
D_3	$\{A,B,H,D\}$	$\hat{\vec{p_3}} = (\omega_A^3, 0)$
t_k	$\{C,A,Z\}$	$\hat{\vec{t_k}} = (\omega_A^k, \omega_C^k)$

$$\hat{\vec{p}}_n = \frac{<\vec{p_n}, \vec{b_1}>}{<\vec{b_1}, \vec{b_1}>}\vec{b_1} + \frac{<\vec{p_n}, \vec{b_2}>}{<\vec{b_2}, \vec{b_2}>}\vec{b_2} + \cdots + \frac{<\vec{p_n}, \vec{b}_{M^k}>}{<\vec{b}_{M^k}, \vec{b}_{M^k}>}\vec{b}_{M^k} \tag{3.41}$$

donde $\{\vec{b_1}, \vec{b_2}, \ldots, \vec{b}_{M^k}\}$ es una base ortogonal del subespacio V definido a partir de $\vec{t}_k$.

Si se compara la representación de los datos dentro del espacio vectorial $\mathbb{R}^{L^k}$ definido por la RRA R (ver tabla 3.6) con la que se obtiene de su proyección en el subespacio vectorial $V \subset \mathbb{R}^N$ definido por la RRA global (ver tabla 3.7), se demuestra que utilizar la estrategia RRA R equivale a haber aproximado, con un error cuadrático mínimo, los vectores patrón de los dominios sobre el subespacio vectorial V. Añadir que, además del

cambio de orden de las componentes, cuestión que no afecta al cálculo de las distancias, sólo existe una pequeña diferencia de $(L^k - M^k)$ posiciones nulas de los vectores patrón debidas a las palabras presentes en el texto a clasificar que no han sido observadas durante la fase de entrenamiento (construcción de la RRA global). Sin embargo, estas posiciones nulas, por un lado, no afectan al resultado del producto escalar de los vectores $< \vec{t}_k, \vec{p}_n >=< \vec{t}'_k, \vec{p}'_n >$, y por otro, afectarán por igual a todas las comparativas en el contexto de las distancias cosenoidales utilizadas (a través de la norma del vector $\vec{t}'_k$).

A continuación se presenta un pequeño ejemplo numérico que permite completar la explicación teórica que se acaba de describir. En este caso, se supone que el texto a clasificar formado por los términos $t_k = \{C, C, A, F, F, F, A, A, F\}$, el cual queda representado por el vector de pesos $\vec{t}'_k = (2, 3, 4)$, donde cada peso indica el número de veces que aparece cada término (tf_i) en el texto, siguiendo el MEV$' = (\omega_C, \omega_A, \omega_F)$ (enfoque RRA R). Asimismo, el conjunto de términos $D_n = \{A, A, C, D, D, D, D, D, F, F, F, F, G, G, G, B, B\}$ define el dominio ejemplo. En este ejemplo, también se considera una red global a partir de la cual se construye el MEV $= \{\omega_A, \omega_B, \omega_C, \omega_D, \omega_E, \omega_F, \omega_G, \omega_H\}$ de $\mathbb{R}^8$, sobre el que los vectores que definen el contenido del texto y el dominio pasan a ser $\vec{t}_k = (3, 0, 2, 0, 0, 4, 0, 0)$ y $\vec{p}_n = (2, 2, 1, 5, 0, 4, 3, 0)$, respectivamente (nótese el cambio de orden de los términos y la inclusión de posiciones nulas en ambos vectores, una vez éstos han sido representados mediante la red de textos global). En este caso, el espacio vectorial global tiene dimensión $N = 8$ y el subespacio V engendrado por la base $\vec{B} = \{\vec{b_1}, \vec{b_2}, \vec{b_3}\} = \{(1, 0, 0, 0, 0, 0, 0, 0), (0, 0, 1, 0, 0, 0, 0, 0), (0, 0, 0, 0, 0, 1, 0, 0)\}$, definida a partir de las componentes activas del $\vec{t}_k$, tiene dimensión $M^k = 3$. Así pues, la mejor aproximación del vector patrón de D_n en el subespacio V, según la ecuación (3.41), será $\hat{\vec{p}_n} = (2, 0, 1, 0, 0, 4, 0, 0)$, o $\hat{\vec{p}} = (2, 1, 4)$ expresado en componentes de $\vec{B}$. En cambio, para RRA R, el vector patrón representado sobre el espacio vectorial de $\mathbb{R}^3$ definido por $\vec{t}'_k$, será $\vec{p}'_n = (1, 2, 4)$, con el orden de componentes designado por el MEV$'$ construido a partir de la RRA R. Como resultado, se obtienen las parejas de vectores texto-dominio: $\{(3, 2, 4), (2, 1, 4)\}$, en componentes del subespacio vectorial V definido sobre el espacio vectorial $\mathbb{R}^8$ de la RRA global y $\{(2, 3, 4), (1, 2, 4)\}$ en componentes del espacio vectorial $\mathbb{R}^3$ definido por la RRA R del textos, por lo que el cálculo del producto escalar será idéntico —además, en este ejemplo, debido a que no hay ningún término de t_k qu no aparezca en el MEV definido por la RRA global, el resultado de calcular la similitud entre el texto y el dominio usando cualquiera de las distancias cosenoidales definidas también será idéntico.

De todos modos, una vez justificado que la RRA R es la mejor aproximación, en términos de mínimos cuadrados, de la RRA F, cabe señalar que la aproximación RRA R implica perder parte de la información contenida en la representación global de los vectores patrón de las RRA F D_n, cuestión que afecta al cálculo de la distancia del coseno a través de la norma de los mismos. En los experimentos que se presentan a continuación, se analiza el impacto de esta aproximación en términos de la eficiencia de la clasificación, así como de su coste computacional.

3.4. Experimentos

A continuación, se presentan los experimentos realizados para analizar los distintos aspectos de la presente propuesta de sistema de clasificación de textos, en lo que se refiere a los parámetros considerados, los modelos estudiados y el coste computacional de los mismos, junto a la evaluación subjetiva de los resultados de la conversión de texto en habla multidominio obtenidos hasta el momento. Los experimentos se dividen en dos grandes bloques. Primero, se presentan los experimentos preliminares realizados sobre corpus de textos (sin señal oral) para evaluar la viabilidad de la propuesta de CTH-MD, utilizando la estrategia de CT basada en RRA F como método de clasificación de textos. A continuación, se describen las pruebas realizadas para estudiar, de forma exhaustiva, las distintas propuestas de sistema de clasificación de textos descritas en el presente trabajo de investigación. Para ello, se hace uso de un corpus de voz publicitario dividido en tres dominios independientes, permitiendo estudiar los resultados sintéticos obtenidos.

En la tabla 3.8 se describen los corpus utilizados en estos experimentos. El primero es una colección de artículos del periódico AVUI en catalán (C_{Cat}), recopilados durante dos periodos de tiempo distintos: a lo largo del año 2000, a partir del trabajo desarrollado en (Guaus y Iriondo, 2000a), y posteriormente ampliados durante el año 2003 mediante el trabajo descrito en (Alías, Iriondo y Barnola, 2003). El segundo corpus es una colección de textos en castellano etiquetados —de tamaño mayor que el anterior—, cedido por el Centro de Lingüística Computacional (CLiC) de la Universidad de Barcelona (UB) (C_{Cast}). Finalmente, el tercer corpus se construyó en el marco de un proyecto de investigación subvencionado[23], en colaboración con el Departamento de Comunicación Audiovisual y Publicidad de la Universidad Autónoma de Barcelona (CAP-UAB), que se encargó se diseñar y grabar un corpus de voz a partir de textos publicitarios, designando tanto los dominios de los textos del corpus como el mejor estilo de locución para cada uno de los dominios considerados.

3.4.1. Experimentos y resultados preliminares

Los resultados que se presentan en este apartado resumen las primeras pruebas que se llevaron a cabo para evaluar la viabilidad de la propuesta de CTH-MD sobre las dos colecciones de textos disponibles en ese instante. Se estudia la eficiencia del clasificador de documentos, así como el impacto teórico de la propuesta en términos de la calidad esperada de la CTH-MD. Excepto en el estudio del algoritmo de jerarquización basado en ICA, donde se trabaja con macro y subdominios, el resto de pruebas se realizan para los dominios considerados a priori, mediante el modelado RRA F, utilizando información temática (*term frequency*) y secuencial (*co-ocurrence frequency*) para representar los textos.

[23] MCyT PROFIT FIT-150500-2002-410

Tabla 3.8: Características de los corpus sobre los que se han llevado a cabo las pruebas en el ámbito de la CT y la CTH-MD.

Corpus	**AVUI**	**CLiC-UB**	**CAP-UAB**
Identificador	C_{Cat}	C_{Cast}	C_{Pub}
Idioma	catalán	castellano	castellano
Tamaño (frases)	202 docs. (9549)	5288 docs.	2.5h (2590)
Número de dominios	4	8	3
Tipo	texto	texto	oral
Dominios	política	política	tecnología
	sociedad	sociedad	educación
	música/cultura	cultura	cosmética
	teatro/literatura	literatura	-
	-	deportes	-
	-	negocios	-
	-	filosofía	-
	-	entretenimiento	-

Experimento 1 - Análisis de la propuesta como clasificador de documentos

Con este experimento se pretende evaluar el funcionamiento del algoritmo de CT basado en RRA F sobre el corpus de textos C_{Cast} del CLiC-UB (ver tabla 3.8) sobre una tarea de clasificación temática. Para ello, se analiza el efecto del número de documentos dedicados al entrenamiento y al test (15 o 20 % del total de documentos de cada dominio) sobre tres particiones distintas del corpus: una con los 4 dominios más poblados, otra con los 6 dominios más poblados, y finalmente, la que incluye los 8 dominios del corpus. La eficiencia del clasificador de textos se mide respecto al etiquetado manual de los dominios correspondientes a cada partición analizada, utilizando la función F_1 (ecuación (3.26)), a partir de los valores micro-promediados de precisión (P^{μ} en ecuación (3.21)) y cobertura (R^{μ} en ecuación (3.22)).

La tabla 3.9 resume los resultados de F_1 obtenidos para los tres conjuntos de pruebas (4, 6 y 8 dominios) respecto a los dos porcentajes de test considerados (15 y 20 %). En la tabla, la fila etiquetada como S_1 corresponde a haber usado la distancia del coseno (ecuación (3.34)) como medida de similitud, mientras que S_2 corresponde a la distancia del coseno ponderada por la longitud del patrón o PL (ecuación (3.35)). En este primer estudio, también se diseñó otra medida de similitud, en este caso basada en la función sigmoidea, que aportaba unos resultados similares a los obtenidos mediante S_2 (ver (Alías, Iriondo y Barnola, 2003) para más detalles). No obstante, se descartó su uso para el resto de pruebas debido a la complejidad del ajuste de los parámetros que la conforman —en un futuro, se seguirá estudiando su uso en el contexto de la CT. Además, en el transcurso de este estudio

todavía no se había estudiado la distancia S_3 (ecuación (3.37)), por lo que no se incluyen sus resultados —su análisis se presenta en la sección 3.4.2.

Tabla 3.9: F_1 obtenida por RRA F mediante las medidas de similitud indicadas sobre seis configuraciones (partición y volumen de entrenamiento) distintas del corpus C_{Cast}.

F_1	**4 dominios**		**6 dominios**		**8 dominios**	
Test	**15 %**	**20 %**	**15 %**	**20 %**	**15 %**	**20 %**
S_1	0.462	0.449	0.368	0.343	0.323	0.322
S_2	0.623	0.608	0.553	0.535	0.521	0.517

Como se puede comprobar en la tabla 3.9, el funcionamiento del algoritmo de CT basado en RRA F presenta unos resultados (medidos con F_1) significativamente superiores a los que se obtendrían de una simple clasificación aleatoria (p.ej. $F_1 = 0.25$ para 4 dominios), reduciendo su eficiencia de clasificación al reducir el volumen de textos dedicados a su entrenamiento (los mejores resultados se obtienen con el 15 % de los documentos dedicados para el test). Asimismo, aumentar el número de dominios de clasificación, por un lado, dificulta notablemente la tarea de la clasificación automática de textos, y por otro, hace que la diferencia en eficiencia de clasificación entre las dos configuraciones de test estudiadas (20 % y 15 %) se reduzca (con valores prácticamente idénticos para 8 dominios).

Para poder validar estos resultados con los presentados en la literatura en lo que se refiere a la tarea de la clasificación temática del texto (ver tabla VI de (Sebastiani, 2002) para un resumen), es necesario utilizar alguna de las grandes colecciones de textos con la que estos algoritmos trabajan normalmente, como por ejemplo la colección Reuters-21758 (Lewis, 1994) o la OHSUMED (Hersh et al., 1994). En este caso, se optó por analizar el funcionamiento del CT basado en RRA F utilizando las distancias S_1 y S_2 sobre 5 de las categorías más pobladas de la colección *Reuters-21758*: *acq*, *earn*, *grain*, *crude* y *trade*, que disponen de 2210, 3776, 574, 566 y 513 documentos, respectivamente —con un número medio de 500 palabras por documento, aproximadamente. Además, en lugar de escoger un determinado porcentaje de textos de entrenamiento y de test, gracias al mayor número de datos disponible, se realiza un barrido de documentos por categoría para evaluar mejor el impacto del volumen de datos de entrenamiento en el funcionamiento del clasificador basado en RRA F. La figura 3.15 presenta los resultados obtenidos. Comentar que los valores de F_1 para más de 500 documentos se obtienen sólo de las dos categorías más pobladas (*acq* y *earn*), mientras que el resto de valores de F_1 se obtienen del macro-promediado de los valores de F_1 por dominio para considerar el tamaño distinto de categorías (ver ecuaciones (3.23) y (3.24)). Como se puede observar, el CT basado en RRA F presenta unos resultados bastante satisfactorios cuando trabaja con S_2, mejorando significativamente respecto los obtenidos mediante S_1 —valores de F_1 del orden de los usuales en la literatura (Sebastiani, 2002), considerando que se trabaja con un número bajo de categorías, cuestión que se compensa por el menor número de documentos de entrenamiento utilizados. Por ejemplo, como se indica en (Sassano, 2003), SVM (con los mejores resultados en la tabla VI de (Sebastiani, 2002)) presenta bajas tasas de clasificación para un número reducido de documentos por dominio,

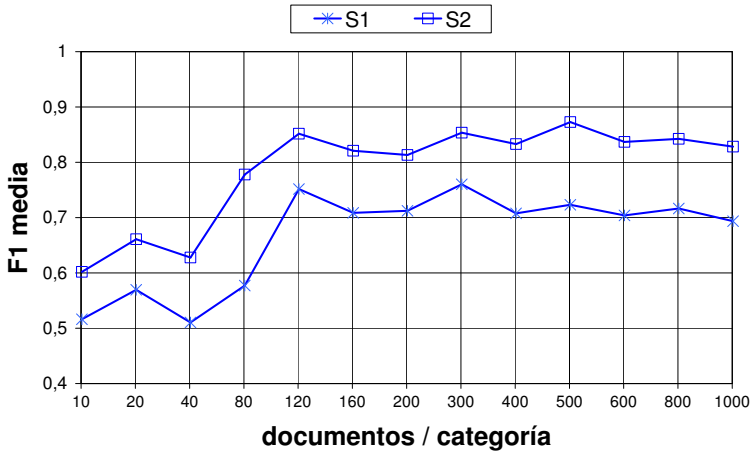


Figura 3.15: Eficiencia de clasificación media (macro-promediada) del clasificador de textos basado en RRA F para distintos tamaños de las cinco categorías estudiadas de la colección *Reuters-21758*.

llegando a dejar de funcionar cuando se dispone de un número de ejemplos de entrenamiento menor a la dimensión del MEV que se está intentando modelar (ver la discusión de la sección 3.5, para más detalles).

Aunque este primer experimento no pretendía ser un análisis exhaustivo de la propuesta en términos de sistema de clasificación temática de documentos, permitió, por un lado, validar la viabilidad de la propuesta basada en una red relacional asociativa, y por otro, analizar el impacto de las medidas de similitud (S_1 y S_2) consideradas para la clasificación de textos. De los resultados obtenidos, se observa que la inclusión del parámetro PL en S_2 permite mejorar de forma sustancial (con un incremento medio del 18 % para C_{Cast} y del 11 % para el experimento basado en la colección Reuters) la eficiencia del clasificador basado en RRA F respecto a utilizar sólo la distancia del coseno (S_1) a lo largo de las pruebas realizadas. Asimismo, el experimento permite observar que cuanto mejor estén modelados los dominios (mayor número de documentos dedicados a entrenamiento), mejor eficiencia de clasificación obtiene el clasificador, con cierta tendencia a estabilizarse a partir de un determinado volumen de datos de entrenamiento.

Experimento 2 - Estudio de viabilidad de la propuesta para CTH multidominio

A continuación se presenta el experimento desarrollado para estudiar la viabilidad de la propuesta para conversión de texto en habla multidominio, incluyendo un sistema de clasificación de textos basado en RRA F. En este experimento se utiliza la colección de textos en catalán C_{Cat} recopilados del periódico AVUI (ver tabla 3.8). Este corpus ha sido diseñado tomando en consideración los criterios de cobertura y balanceo de las unidades de voz típicamente utilizados en la creación de corpus para CTH. Este corpus está formado por 9549 frases divididas en 4 dominios distintos, por lo que permite estimar de forma teórica

el funcionamiento del sistema de CTH-MD propuesto y, así, validar la viabilidad de la propuesta —en este caso, sobre un corpus multidominio formado por dominios independientes (estrategia *tiering*).

Como paso previo, se estudia la eficiencia del CT basado en RRA F sobre este corpus de textos en catalán. En la tabla 3.10 se muestra la eficiencia de clasificación F_1 obtenida para distintos porcentajes de test considerados (sobre el total de documentos), utilizando la distancia S_2, que ha demostrado un mejor comportamiento que S_1 en el experimento anterior. Como se puede apreciar, se obtienen resultados similares a los conseguidos sobre el corpus C_{Cast} para cuatro dominios (ver tabla 3.9), aunque con la particularidad de trabajar con un volumen mucho menor de datos de entrenamiento (ver tabla 3.8).

Tabla 3.10: Eficiencia de clasificación obtenida sobre el corpus C_{Cat} por el CT basado en RRA F, utilizando la distancia del coseno ponderada por PL (S_2) para distintos porcentajes de test.

Test	**10 %**	**15 %**	**20 %**	**25 %**
F_1	0.695	0.636	0.628	0.654

No obstante, el objetivo final de este experimento es estudiar el impacto del coste computacional de la arquitectura multidominio definida (que incorpora el CT), así como la calidad *estimada*[24] de la síntesis para cada dominio. Para ello, se estudia el tiempo medio de ejecución (*Average Execution Time* o AET, en inglés) del proceso que simula la síntesis (incluyendo el tiempo correspondiente a la CT), junto a la longitud media de los segmentos de síntesis (*Average Segment Length* o ASL, en inglés) (Batůšek, 2001; Chu et al., 2002) —también denominado como *Super-Unit Mean Length* en (Guaus y Iriondo, 2000a). Ambos parámetros se obtienen como promedio del resultado parcial de las distintas pruebas realizadas (frases de test).

Por un lado, el AET evalúa el tiempo medio que el sistema necesita para obtener la secuencia de unidades que conformen las distintas frases de test a lo largo de las pruebas. Debido a que en este experimento no se genera el mensaje oral, este parámetro representará el coste computacional de la CT más el tiempo correspondiente a la búsqueda de las unidades en el dominio del corpus indicado (queda fuera de AET el tiempo de procesado de la señal necesario para constuir la señal de voz sintética). Por otro lado, el ASL permite estimar cualitativamente la naturalidad de la síntesis, ya que se ha demostrado que existe una correlación entre esta medida y la calidad de la síntesis evaluada subjetivamente (medida con el *Mean Opinion Score* o MOS, en inglés) (Chu y Peng, 2001; Chu et al., 2002). Concretamente, cuanto mayor sea el ASL, mayor será el tamaño de los segmentos de voz (mayor número de unidades continuas) y menor serán el número de concatenaciones y discontinuidades audibles en la señal generada, cuestión clave para obtener una voz sintética

[24]Se estudia la calidad mediante un parámetro objetivo, ya que no se dispone de señal de voz alguna, al trabajar sólo con la trascripción fonética de las frases del corpus.

natural. Sin embargo, es necesario comentar que la correlación entre ASL y MOS se ha demostrado dentro del ámbito de los sistemas de propósito general, donde las variaciones prosódicas entre las unidades son mínimas[25].

Según las premisas que se acaban de describir, la función de coste del módulo de selección de unidades se ha simplificado para trabajar sólo sobre la trascripción fonética de las frases (el corpus C_{Cat} no contiene información oral). En este contexto, el módulo de selección de unidades, simplemente, buscará la secuencia de unidades que maximice la longitud de los segmentos que formen la frase a sintetizar a partir de las unidades que forman el corpus, es decir maximizar el ASL (Guaus y Iriondo, 2000a). Para ello, la función de coste sólo tendrá en consideración el coste de concatenación (ver ecuaciones (2.2) y (2.3) del capítulo 2), definido de forma booleana (binaria), según la ecuación (3.42) (Chu et al., 2001) —el coste de unidad sólo contempla escoger las unidades indicadas en la trascripción fonética de la frase a sintetizar.

$$C^c(u_i, u_{i+1}) = \begin{cases} 0, & \text{si } u_i \text{ y } u_{i+1} \text{ son consecutivas en el corpus} \\ 1, & \text{de lo contrario} \end{cases} \tag{3.42}$$

donde C^c es el coste de concatenación, u_i la unidad actual y u_{i+1} la siguiente unidad.

Tabla 3.11: Ejemplo del resultado de la selección de unidades para una frase en catalán.

Frase	"Un hivern plàcid"(*Un invierno plácido*)
Transcripción	[uniBERmplasit]
Unidades	[un] [ni] [iB] [BE] [ERm] [mp] [pla] [as] [si] [it]
SuperUnidades (*post-selección*)	[un] [niBERm] [mp] [plas] [sit]

En la tabla 3.11 se presenta un pequeño ejemplo del cálculo del ASL extraído de (Guaus y Iriondo, 2000b). En este caso, las unidades fonéticas están representadas según la notación SAMPA (*Speech Assessment Methods consortium Phonetic Alphabet*) (Wells et al., 1992). Para construir la frase son necesarias 10 unidades (difonemas y trifonemas del catalán), que después del proceso de selección se convierten en 5 *SuperUnidades*, gracias a que algunas de las unidades solicitadas han sido encontradas consecutivamente dentro del corpus. Así pues, una *SuperUnidad* estará formada por la unión de diversos difonemas o trifonemas. Por ejemplo, $[niBERm]$ contiene 4 unidades, mientras $[un]$ corresponde a 1 unidad, número mínimo de unidades contenidas en una *SuperUnidad*. A partir de este ejemplo, se puede calcular el ASL, siguiendo la ecuación (3.43).

[25]No obstante, si se trabaja sobre corpus no neutro (p.ej. corpus expresivos), seleccionar la secuencia de unidades que comporte un menor número de discontinuidades no tiene porque significar que estas unidades presenten los mejores puntos de concatenación ni las características prosódicas más adecuadas (p.ej. ver (Alías et al., 2005) o apéndice D.1), por lo que se descarta su uso en los experimentos finales.

$$ASL = \frac{\#Unidades}{\#SuperUnidades} = \frac{10}{5} = 2 \tag{3.43}$$

Por lo tanto, cuanto mayor sea el valor del ASL menor número de *SuperUnidades* conformarán la señal sintética, y por lo tanto, menos concatenaciones ésta contendrá. Como se discute en (Guaus y Iriondo, 2000a), el ASL depende de diversos parámetros, como son: el tamaño del corpus (el número de unidades que lo forman), la similitud entre el estilo (o dominio) del contenido del corpus respecto al del texto a sintetizar —uno de los elementos básicos de la propuesta CTH-MD—, y la longitud media (en unidades) de las frases presentes en el corpus.

En este caso, las pruebas se llevaron a cabo sobre un PC (PIV 1.6GHz - 256 MB RAM) con sistema operativo Windows 2000 ®, utilizando el compilador de Visual C++ 6.0 ®—los tiempos correspondientes al coste computacional se obtuvieron con redondeo al segundo.

Una vez definidas las medidas de evaluación que se consideran en este segundo experimento preliminar, se pasa a detallar los pasos seguidos para evaluar los resultados obtenidos sobre el corpus C_{Cat} dividido en cuatro dominios (ver tabla 3.8). Por una parte, se estudia el comportamiento de la selección de unidades sobre los distintos dominios del corpus, y por otra, se evalúa la eficacia del binomio clasificación-selección por dominio respecto a realizar la búsqueda *global* sobre un corpus definido a partir de todos los dominios contemplados. Las pruebas se realizan a nivel de frase (1 frase por documento). Una vez finalizadas, se calculan el AET y el ASL resultantes. Concretamente, las frases de test se han obtenido de los ficheros pol01, de 49 frases, soc01, de 45 frases, cul01, con 70 frases y lit01, con 50 frases —ficheros que han sido extraídos de la colección de documentos. Comentar que, cada uno de los subcorpus (dominios) considerados estarán constituidos por dos partes diferenciadas, siguiendo lo descrito en (Guaus y Iriondo, 2000a; Guaus y Iriondo, 2000b):

- **Unidades básicas:** esta parte del corpus asegura la cobertura de todas las unidades del idioma considerado, el catalán en este caso. Concretamente, se trata de 895 difonemas y 312 trifonemas, sumando 1207 unidades en total, obtenidas a partir de un listado de palabras portadoras. Gracias a estas unidades, el sistema de síntesis siempre dispondrá, como mínimo, de una realización de la unidad a sintetizar. Este subconjunto permite la compatibilidad hacia atrás de la propuesta de síntesis basada en selección de unidades con la síntesis basada en difonemas y trifonemas.

- **Realizaciones de las unidades básicas:** la segunda parte del corpus estará formada por las distintas realizaciones de las unidades básicas (difonemas y trifonemas) presentes en las transcripciones fonéticas de las frases del corpus C_{Cat}. De este modo, el módulo de síntesis dispondrá de varias unidades candidatas para seleccionar la secuencia de unidades más apropiada según la función de coste definida.

La primera parte del estudio consiste en realizar un barrido del número de unidades por subcorpus para los 4 dominios de C_{Cat}: política, sociedad, cultura y literatura. Empezando

con un tamaño de 20000 unidades (difonemas y trifonemas) por dominio, se va incrementando en 5000 unidades su tamaño, hasta llegar a 80000 unidades, siguiendo un proceso similar al descrito en (Guaus y Iriondo, 2000a), pero con un objetivo distinto. En aquel trabajo se pretendía determinar el tamaño óptimo de un corpus de voz para selección de unidades, demostrando la importancia del estilo del corpus respecto al ASL obtenido, mientras que en este experimento se pretende evaluar la viabilidad de la propuesta de CTH-MD descrita. La figura 3.16 presenta el resultado obtenido del barrido para los cuatro dominios estudiados. La función $ASL = f(AET)$ resultante presenta un comportamiento similar para todos los dominios estudiados. La línea de puntos discontinua une, de forma alternada, los puntos correspondientes a cada dominio para un mismo tamaño del subcorpus (Alías, Iriondo y Barnola, 2003).

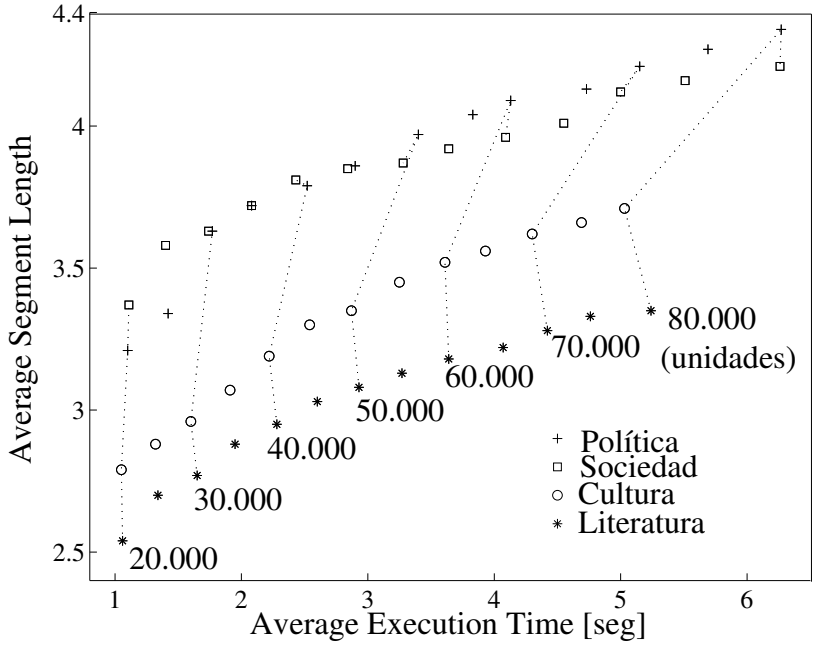


Figura 3.16: $ASL = f(AET)$ para los cuatro dominios del corpus C_{Cat}, para el barrido de 20000 hasta 80000 unidades por dominio. Las líneas discontinuas unen los resultados para un mismo tamaño de corpus de forma alternada.

Del análisis de la figura 3.16, se puede concluir que, a medida que aumenta el tamaño de cada dominio, también aumentan el tiempo de ejecución, así como la longitud de las *SuperUnidades* obtenidas, pero no de forma lineal, sino que crece aproximadamente de forma logarítmica (función convexa) —al igual que sucedía en (Guaus y Iriondo, 2000a). Asimismo, se puede observar que los valores de ASL y AET obtenidos dependen del dominio considerado. Por un lado, los dominios de *política* y *sociedad* presentan los valores de ASL mayores, debido a que estos dos dominios contienen frases de longitud media mayor que el resto, cosa que facilita encontrar *SuperUnidades* más largas. Por otro lado, los resultados para los dominios de *cultura* y *literatura* presentan un crecimiento menor de AET, respecto

a los dominios de *política* y *sociedad*, a medida que el tamaño del corpus aumenta. Este comportamiento se debe al menor número de realizaciones por unidad de estos dominios (contienen frases más cortas), cuestión que provoca una reducción del tamaño del espacio de búsqueda, disminuyendo, consecuentemente, el tiempo de selección —en el experimento, el proceso de selección de unidades toma en consideración todas las realizaciones de las unidades del corpus correspondientes a la frase a sintetizar.

La segunda parte del experimento consiste en comparar los resultados obtenidos, en términos de ASL y AET, al trabajar con el módulo de CT sobre los cuatro subcorpus de dominio (D_n) respecto a hacerlo sobre el corpus equivalente global ($\sum D_n$) (acumulando todos los subcorpus) sin incorporar clasificación de textos. En este caso, este corpus se ha construido acumulando los cuatro dominios estudiados en un mismo corpus, por lo que sus valores de ASL definirán la cota superior de los resultados que se pueden obtener a partir del corpus multidominio (dominios separados según la estrategia *tiering*). Para esta prueba, se ha definido un tamaño para los corpus de los dominios de 20000 unidades, y, consecuentemente, un tamaño de 80000 unidades para el corpus acumulado (dimensión máxima en el estudio anterior). Cada uno de estos corpus estará formado por las 1207 unidades básicas más las realizaciones de las unidades obtenidas del conjunto de frases correspondientes a cada dominio, asegurando así la cobertura del idioma. Asimismo, el corpus acumulado seguirá el mismo formato, pero agrupando las realizaciones de las unidades de todos los dominios.

Tabla 3.12: ASL y AET [seg] para los cuatro dominios C_D (de 20000 unidades cada uno) de C_{Cat}, junto a los valores de estos parámetros obtenidos para el corpus global $\sum C_D$ (de 80000 unidades).

Test	**Literatura**		**Cultura**		**Política**		**Sociedad**	
Corpus	$\sum D_n$	D_n	$\sum D_n$	D_n	$\sum D_n$	D_n	$\sum D_n$	D_n
ASL	3.2	2.6	3.5	2.8	4.0	3.3	4.0	3.4
AET	1.5	1.1	1.6	1.0	2.3	1.1	2.2	1.1

Analizando los resultados obtenidos (ver tabla 3.12), se puede observar que trabajar con el módulo de CT sobre el corpus multidominio (CTH-MD) provoca una reducción media del 15 % en el ASL y del 40 % en AET respecto a trabajar con el corpus acumulado sin CT. De este resultado, se puede deducir que la estrategia CTH-MD, que incorpora la CT antes que la selección de unidades, permite una reducción importante del tiempo de ejecución de la conversión de texto en habla, sin que esto implique una pérdida excesiva de la calidad *esperada* de la señal sintética (siempre que el clasificador de textos acierte el dominio del texto de entrada). Además, cabe destacar que el corpus de referencia no es un corpus de propósito general propiamente dicho, sino que es el resultado de acumular los distintos corpus independientes (similar a la filosofía *blending*), por lo que presenta el valor máximo de ASL que puede alcanzar a partir del corpus multidominio. Esta cuestión invita a suponer que este decremento de la calidad esperada será menor cuando la comparativa se realice

contra un corpus genérico, ya que su cobertura no se diseña a partir de la acumulación de distintos dominios, sino que se busca dar cobertura al idioma considerado (CTH-PG). Este experimento permitió validar, aunque inicialmente sólo de forma teórica, la propuesta de CTH-MD, cuestión que permitió abordar el diseño e implementación de un sistema de CTH-MD basado en corpus y estudiarlo mediante los experimentos que se describen en la sección 3.4.2.

Experimento 3 - Jerarquización del corpus mediante ICA

Este experimento, realizado en colaboración con Xavier Sevillano (Alías et al., 2003), analiza la capacidad del algoritmo ICA para jerarquizar los documentos del corpus C_{Cat}, a partir de un barrido del número de grupos (variando el número de componentes independientes considerado, K) sobre todos los datos disponibles en la colección. Debido a que esta tarea no trata de clasificar sino de organizar los datos del corpus, se ha optado por evaluar el resultado a partir de la exactitud de la agrupación obtenida (ecuación (3.20)), referenciada a los textos previamente etiquetados —a diferencia de los experimentos previos enfocados a evaluar la eficiencia de la clasificación usando la medida F_1. En el experimento se ha utilizado un algoritmo de punto fijo que maximiza el momento de tercer orden de los datos (*skewness*, en inglés), siguiendo el trabajo de (Kaban y Girolami, 2000).

Como primer paso, y debido a que la colección estudiada se divide en cuatro dominios, la primera prueba consiste en hallar $K = 4$ grupos de documentos o *clusters* para organizar temáticamente los documentos del corpus. La tabla 3.13 presenta el resultado obtenido. Se puede apreciar que la agrupación de datos no es perfecta, a excepción de los documentos de teatro que se agrupan en un mismo cluster, el resto de documentos no se agrupan en un único *cluster*. Por ejemplo, el 93.2 % de los documentos de *música* se agrupan en el mismo *cluster*, mientras que sólo el 70.5 % de los documentos de *política* están dentro del mismo grupo. De este primer experimento, se puede extraer la mayor dificultad que presentan los documentos de política y de sociedad para ser agrupados eficientemente en un único grupo de datos.

Tabla 3.13: Exactitud de la agrupación obtenida sobre el corpus C_{Cat} mediante ICA trabajando con $K = 4$ componentes independientes.

Dominio	**Exactitud**
Política	0.705
Sociedad	0.754
Música	0.932
Teatro	1.000

El siguiente paso en el experimento consiste en analizar la capacidad de clasificación y agrupamiento del algoritmo de CT basado en ICA conforme aumenta el número de *clusters*. En la figura 3.17 se muestra la evolución de la exactitud de la clasificación de los textos de

cada dominio para diversos valores de $K \geq 4$. Como puede apreciarse, la tasa de clasificación máxima (exactitud en %) para el dominio de *política* se logra para $K = 7$, con un valor de 96.7 %, mejorando notablemente los resultados obtenidos en este dominio para $K = 4$. El número de *clusters* en el que se agrupan sus textos para $K = 7$, en este caso, es 4, mientras que el dominio de *sociedad* no mejora su exactitud al aumentar el número de componentes independientes —se clasifica con una exactitud del 75.4 % cuando $K = 4$ y $K = 10$, agrupando los documentos en 1 o 3 grupos, respectivamente.

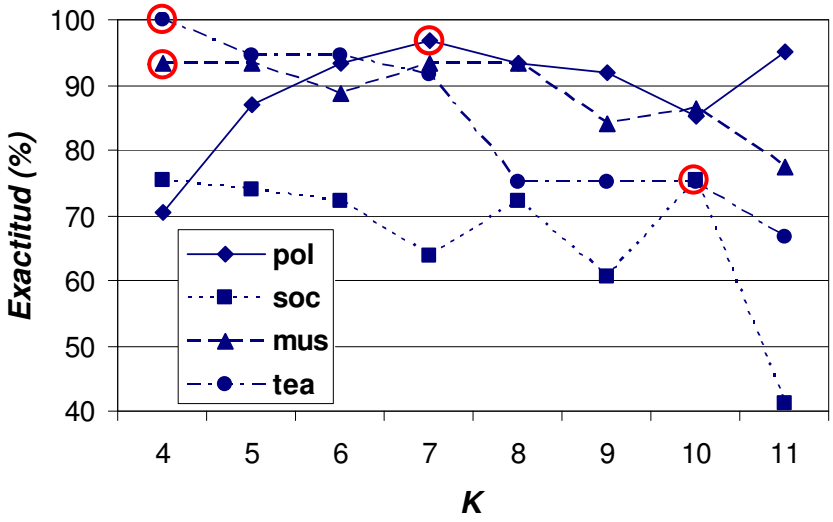


Figura 3.17: Tasa de la clasificación obtenida mediante ICA, para un barrido de componentes independientes $K \in [4, 11]$. El número óptimo de grupos para cada dominio está indicado con un círculo.

Por otro lado, la mejor distribución de los textos de *música* y *teatro* se obtiene para $K = 4$, donde cada componente independiente corresponde a un único dominio, por lo que se deduce que cada uno de estos dominios se puede caracterizar mediante un único grupo de documentos o *cluster*. Una vez revisados los textos, se constata que el contenido de estos dominios contiene textos de una temática bastante homogénea y específica, a diferencia de lo que sucede con los dominios de *política* y *sociedad*, con un contenido más heterogéneo.

El experimento finaliza aplicando el algoritmo ICA para valores de $K < 4$ con el objetivo de generar niveles de jerarquía superior (macrodominios y superdominios). Tal y como se muestra en la figura 3.18, el análisis para $K = 3$ muestra un agrupamiento de los textos de *música* y *teatro* en un macrodominio que podría denominarse como *cultura*, mientras que para $K = 2$ los textos de *sociedad* y *cultura* comparten un mismo superdominio. En esta misma figura se presentan los subdominios obtenidos a partir del barrido con un número de componentes independientes mayor al número de categorías de partida presentado en la figura 3.17.

Este pequeño experimento, se ha descrito con el objetivo de demostrar la viabilidad

del método de CT basado en ICA aplicado a la tarea de la jerarquización del contenido de un corpus de textos, con el objetivo de aplicarlo en un futuro como paso previo a la aplicación del sistema de CT basado en RRA descrito en este trabajo de investigación. Para más detalles sobre otros experimentos en el ámbito de la clasificación de textos basada en ICA, consultar (Sevillano, Alías y Socoró, 2004).

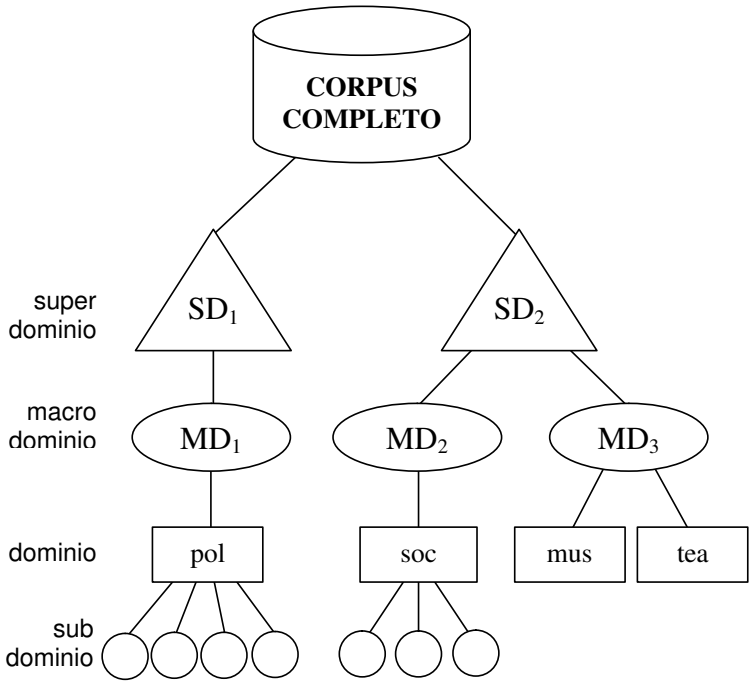


Figura 3.18: Estructura jerárquica de la colección de textos obtenida mediante ICA, como resultado de la agrupación de los contenidos del corpus según barrido de componentes independientes realizado $|\mathcal{C}| < K \leq |\mathcal{C}|$, donde $|\mathcal{C}|$ indica el número de categorías definidas a priori.

3.4.2. Análisis objetivo y subjetivo de la propuesta

En los siguientes experimentos, se analizan las dos propuestas de explotación de la red relacional asociativa presentadas —RRA F y RRA R— sobre el corpus C_{Pub} (ver tabla 3.8), comparando su funcionamiento respecto a un sistema de clasificación de referencia, en este caso, *Nearest Neighbour* (en la sección 3.5 se detallan los motivos). Según esta estrategia, una vez representados todos los vectores mediante un espacio vectorial común, el texto de entrada se asignará a la categoría asociada al documento más cercano[26], según la distancia de similitud empleada —filosofía de *pattern matching*. Todas las pruebas se

[26]En un futuro se pretende estudiar la versión k-Nearest Neighbour, con una asignación por votación de las categorías de los k documentos más cercanos a la consulta.

realizan a partir de un 10-*fold random-subsampling*[27], entrenando los clasificadores con el 80 % de los datos del dominio analizado. Como se discute en (Alías et al., 2004b), es importante realizar un barrido en las pruebas para evitar sesgos en los resultados al trabajar con conjuntos de entrenamiento y de test de tamaño reducido —cuando se trabaja con grandes corpus de textos, la propia variabilidad de los datos de test conlleva disponer de una buena generalidad de los resultados. Para poder comparar de forma consistente los distintos métodos estudiados, se utiliza una *semilla* que permite aleatorizar los ficheros de test, pero compartiendo el grupo de documentos de entrenamiento y de test. En este caso, se evalúa mediante la función F_1 *micro-promediada* que incorpora en el cálculo la precisión y la cobertura del método de clasificación estudiado (ver ecuación (3.26)).

El corpus C_{Pub}, grabado por una locutora profesional del CAP-UAB, está formado por 2590 frases extraídas de una base de datos publicitaria, que se agrupan en tres dominios: educación (EDU: 916 frases), tecnología (TEC: 833 frases) y cosmética (COS: 841 frases). Además, las frases de cada uno de los tres dominios temáticos se han grabado, respectivamente, con tres estilos de locución distintos (Montoya, 1999): *i)* alegre, que corresponde al estereotipo extrovertido/alegre/fascinado, *ii)* estable, que corresponde al estereotipo estable/inteligente/sensitivo y maduro (generalmente conocido como *neutro*) y *iii)* sensual. En cada caso se escogió el estereotipo más adecuado para el contenido de cada dominio según el criterio de los expertos del CAP-UAB, permitiendo clasificar los estilos de locución a través de los contenidos de los textos correspondientes. En un futuro se prevé estudiar e incorporar otro tipo de correspondencias para la designación automática de dominios a partir del texto (ver sección 3.5).

Con el objetivo de evaluar la habilidad de las estrategias de clasificación de texto analizadas, las frases grabadas se agrupan aleatoriamente para generar pseudo-documentos[28] susceptibles de ser clasificados. De este modo, se puede evaluar la eficiencia de los métodos de CT comparados a medida que los pseudo-documentos van reduciendo el número de frases que contienen, pasando desde una situación más cercana al problema de la CT clásica (con muchas frases por documento), hasta el caso extremo de disponer sólo de una frase por documento, situación habitual en el ámbito de la CTH. El resultado de la generación de estos pseudo-documentos se presenta en la tabla 3.14(a), con un conjunto de $\{30, 25, 20, 15, 10, 7, 6, 5, 4, 3, 2, 1\}$ frases por pseudo-documento —la distribución de palabras por pseudo-documento se muestra en la figura 3.19. Por ejemplo, para el caso de 5 frases por pseudo-documento, se obtienen 166 documentos de tecnología, con 3 frases sobrantes que son descartadas. A partir del trabajo presentado en (Alías et al., 2004b), se pudo apreciar que existía un conjunto de frases dentro de la colección con un contenido temático poco definido (ambigüedad de categoría), cuestión que dificulta su clasificación (p.ej. frases tipo "*La mejor solución*", "*De principio a fin*", "*Forma parte de tu vida*", etc.). Esta situa-

[27] El *k-fold random-subsampling* (con $k = 10$, habitualmente) es una estrategia de partición de los datos en conjunto de entrenamiento y test parecida al *k-fold crossvalidation* (Sebastiani, 2002), pero sustituyendo el barrido de k conjuntos disjuntos por k grupos definidos aleatoriamente —por lo que todos los resultados descritos son fruto del promedio de los 10-*folds* realizados. Se utiliza esta técnica debido al reducido número de ejemplos disponible, sobretodo en la zona de mayor agrupación de frases por documento del barrido.

[28] Estos pseudo-documentos son simples agrupaciones de frases del mismo dominio, por lo que no pueden definirse como documentos estrictamente hablando, ya que no siguen ningún hilo argumental.

ción resulta muy evidente para el caso del clasificador basado en ICA (fundamentalmente de carácter temático), como se discute en (Alías et al., 2004b) y se resume en el apartado 3.3.2 de este trabajo de investigación. Por ello, se decide estudiar el impacto en la eficiencia del CT de eliminar estas frases de la colección, dando lugar al *corpus reducido*, cuya agrupación en pseudo-documentos se presenta en la tabla 3.14(b) y cuya distribución de palabras por pseudo-documento se muestra en la figura 3.19. Se puede observar de la figura que la mayoría de las frases eliminadas (manualmente) corresponden a textos cortos, por lo que el número medio de palabras por pseudo-documento (en adelante, documento) es menor en el corpus completo.

Tabla 3.14: Distribución de los pseudo-documentos por dominio, en función del número de frases consideradas en cada documento, para las dos versiones estudiadas del corpus publicitario.

Corpus completo	**Pseudo-documentos por dominio**		
Frases	**EDU**	**TEC**	**COS**
1	916	833	841
2	458	416	420
3	305	277	280
5	183	166	168
6	152	138	140
7	130	119	120
10	91	83	84
15	61	55	56
20	45	41	42
25	36	33	33
30	30	27	28

(a) Agrupación de frases por pseudo-documento para el corpus completo.

Corpus reducido	**Pseudo-documentos por dominio**		
Frases	**EDU**	**TEC**	**COS**
1	527	323	517
2	263	161	258
3	175	107	172
5	105	64	103
6	87	53	86
7	75	46	73
10	52	32	51
15	35	21	34
20	26	16	25
25	21	12	20
30	17	10	16

(b) Agrupación de frases por pseudo-documento para el corpus reducido.

En cuanto a la parametrización del texto (descrita en la sección 3.3.3), se estudia el funcionamiento de los CT analizados trabajando con cuatro configuraciones de pesos distintas. Por un lado, se estudia el impacto de incorporar parámetros estructurales, en este caso, las coocurrencias de las palabras (COF en las figuras), respecto a no incluirlos (NCOF en las figuras) en la representación de los textos. Por otro lado, se analizan la influencia de las ponderaciones de los términos mediante los parámetros temáticos $tf \times idf$ (ecuación (3.6)) (TFIDF en las figuras), como ponderación clásica de términos, e IWF (ecuación (3.27)), como nueva ponderación a nivel de palabra.

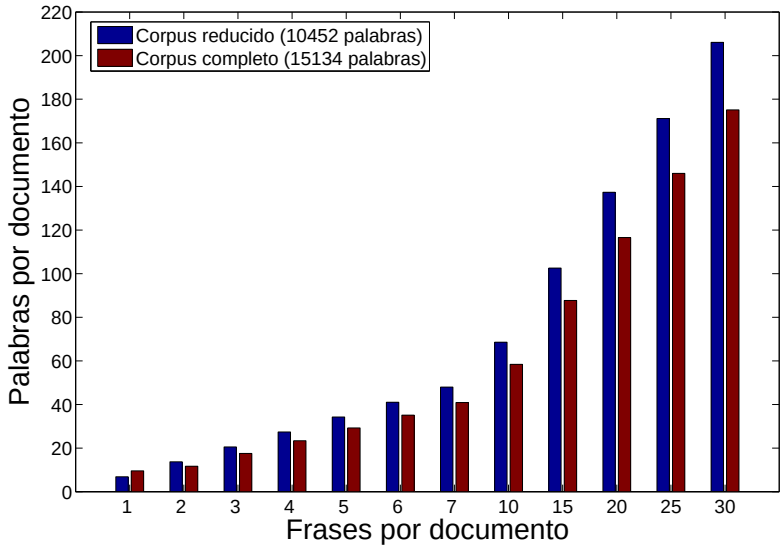


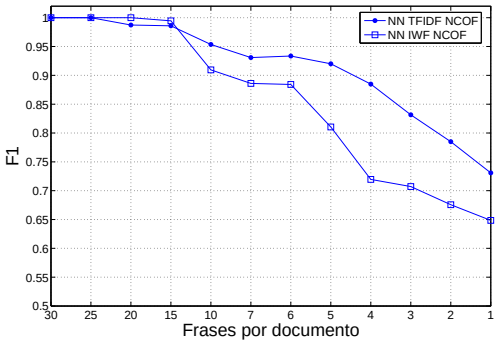
Figura 3.19: Distribución del número medio de palabras por documento para las dos versiones estudiadas del corpus, a lo largo del barrido de frases por documento realizado.

Experimento 4 - Análisis de las propuestas para distintas configuraciones

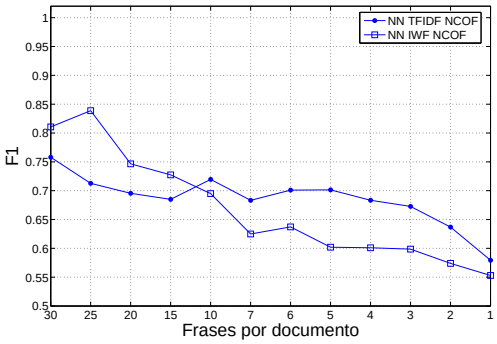
En el primer experimento se analiza el funcionamiento de los tres métodos considerados: el método de referencia *Nearest Neighbour* (NN en las figuras) junto a los métodos de CT basados en RRA (RRA F y RRA R), aplicados a la clasificación de textos tanto para el *corpus reducido* como para el *completo*, según el barrido de frases por documento descrito anteriormente. En las figuras se presentan los resultados de las distintas configuraciones de CT estudiadas, modelando el texto con o sin las coocurrencias de las palabras (COF o NCOF), y representando los términos mediante las ponderaciones TFIDF o IWF. En este experimento, se utiliza la distancia del coseno, S_1 (ecuación 3.34), para la clasificación.

El propósito de la prueba es, por un lado, comparar el funcionamiento de las distintas parametrizaciones de texto en términos de eficiencia de clasificación (medida F_1) promediada para los tres dominios del corpus, y por otro, analizar el impacto del tamaño de los textos a clasificar en la tarea de clasificación de textos planteada, con el objetivo de encontrar la mejor estrategia de clasificación de entre los métodos propuestos en el ámbito de la CTH-MD, estudiando el impacto de la parametrización del texto en este problema.

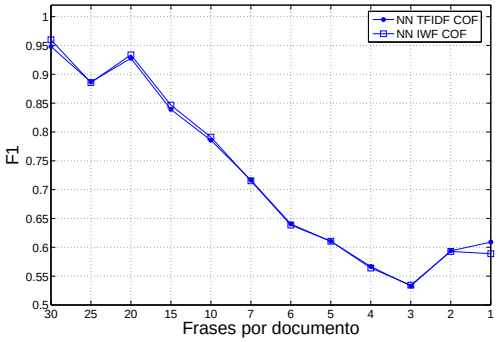
Resultados obtenidos con el método de CT basado en NN: El primer estudio evalúa la relación entre la eficiencia de clasificación del CT basado en *Nearest Neighbour* (NN) (método de referencia), según un MEV sin incluir las coocurrencias (NCOF). Como se puede apreciar en las figuras 3.20(a) y 3.20(b), a medida que se reduce el número de



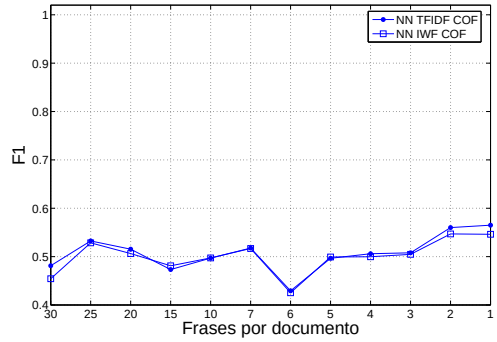
(a) Sin coocurrencias sobre el corpus reducido.



(b) Sin coocurrencias sobre el corpus completo.



(c) Con coocurrencias sobre el corpus reducido.



(d) Con coocurrencias sobre el corpus completo.

Figura 3.20: Eficiencia de clasificación obtenida por el clasificador de textos basado en NN sobre el corpus completo y su versión reducida, para distintas parametrizaciones del texto.

frases por documento, la eficiencia del clasificador disminuye, independientemente del tipo de ponderación temática utilizado. Por ejemplo, para el corpus reducido, se pasa de una $F_1 = 1$ para 30 frases/doc, a una $F_1 = 0.7309$ con TFIDF o una $F_1 = 0.6405$ con IWF para 1 frase/doc. No obstante, la ponderación TFIDF presenta una mejor eficiencia global que IWF a lo largo del barrido (a excepción de la zona inicial del barrido dentro del corpus completo). Por otro lado, si se analiza el funcionamiento del método NN respecto al tipo de corpus, se observa claramente el efecto negativo que tiene para este método el hecho de incluir frases temáticamente ambiguas en la colección. Por otro lado, en las figuras 3.20(c) y 3.20(d) se presentan los resultados obtenidos al incluir las coocurrencias (COF) en la representación de los textos. Se observa como los resultados empeoran significativamente respecto a los obtenidos sin las coocurrencias, sea cual sea el tipo de ponderación temática utilizada (TFIDF o IWF). Asimismo, los resultados del corpus completo presentan una peor

eficiencia global —menores valores de F_1 y comportamiento mas errático— respecto a su versión reducida, con unos resultados muy malos al incluir COF en la parametrización del texto.

Así pues, se puede concluir que para el clasificador basado en NN las representaciones sin coocurrencias junto a la ponderación TFIDF son las parametrizaciones que ofrecen una mejor tasa de clasificación de entre las posibles configuraciones estudiadas. Además, se debe añadir que NN es un método muy sensible a las características de los datos de entrenamiento, como se puede deducir al comparar los resultados obtenidos con el corpus reducido respecto a los conseguidos con el corpus completo.

Resultados obtenidos con el método de CT basado en RRA F: Un análisis inicial de los resultados obtenidos por CT basado en RRA F, se puede observar que este método presenta un comportamiento similar al método de referencia (NN), utilizando la distancia del coseno como medida de similitud (ver figuras 3.20 y 3.21). Esto es debido, fundamentalmente, a que ambos métodos siguen una filosofía de clasificación parecida, modelando todos los datos sobre un MEV común antes de llevar a cabo la clasificación de los textos. Si se comparan globalmente los resultados, se puede comprobar que, en ambos casos, la ponderación temática TFIDF es la presenta una mejor eficiencia de clasificación (F_1), aunque para NN y el corpus completo existe una zona (de 30 a 15 frases/doc) donde IWF presenta mejores (o iguales) resultados que TFIDF. Asimismo, la inclusión de las coocurrencias (ver figuras 3.21(c) y 3.21(d)), provoca que la RRA F empeore claramente su eficiencia de clasificación comparándola con la obtenida sin incluirlas. Parece claro, pues, que para las representaciones globales de la información (NN y RRA F), incluir las coocurrencias en el modelado del texto no sólo no ayuda al CT, sino que empeora sus resultados, trabajando con la distancia del coseno (esto no es así para otras medidas de similitud, cuyas pruebas se describen a continuación). Siguiendo con el modelo algebraico descrito en el apartado 3.3.3, se puede deducir que el hecho de aumentar el tamaño de los vectores de referencia (los vectores patrón para RRA F o los vectores que representan a cada documento en NN) implica aumentar el valor de su norma, ya que contemplan más elementos (los términos más sus coocurrencias). Este elemento afectará decisivamente al cálculo de la distancia del coseno (ver ecuación 3.34), reduciendo el rango de ángulos en la comparativa y aumentando la probabilidad de error en la clasificación. Esta situación se ve agravada en el caso de la RRA F, que representa cada dominio mediante un único vector patrón, a diferencia de la NN que tiene tantos *patrones* como documentos considerados en el entrenamiento. En el caso de la RRA F, el aumento de la norma del vector patrón será mucho mayor (varios órdenes de magnitud) al del vector del texto a clasificar (con un número de términos y coocurrencias mucho más reducido), a diferencia de la NN donde los tamaños serán más comparables. Además de este razonamiento, cabe añadir el problema de la representatividad estadística de los datos, ya que para un mismo conjunto de entrenamiento se están modelando tanto las palabras como sus coocurrencias (relaciones de segundo orden), disponiendo estas segundas de menor robustez estadística.

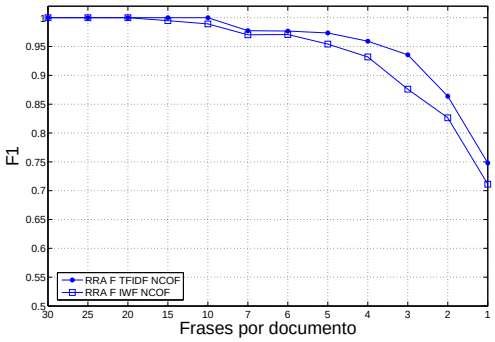
Por otro lado, los resultados obtenidos con RRA F para ambos corpus sin considerar las coocurrencias de las palabras mejoran notablemente los obtenidos con el método de

referencia, utilizando la misma ponderación temática (ver figuras 3.21 y 3.20). De este modo, se demuestra que la representación de los textos en un único vector patrón utilizada en el CT basado en RRA F consigue una mejor eficiencia de clasificación que el uso de los distintos vectores del método NN. Este resultado se observa para cualquier tamaño de documento (aunque para tamaños importantes, presentan F_1 equivalentes), aumentando la eficiencia del CT basado en RRA F respecto a la NN a medida que se reduce el número de frases por documento, así como ante la existencia de frases temáticamente ambiguas en el entrenamiento y en la clasificación (corpus completo). Por lo contrario, cabe destacar que, en general, la inclusión de COF en RRA F provoca un impacto más negativo en este sistema de clasificación que en el sistema basado en NN. No obstante, para el corpus reducido y trabajando simplemente con la distancia del coseno, RRA F presenta mejores (o iguales) resultados que NN, para la misma ponderación temática, dentro de la zona de mayor interés de la clasificación de textos en el contexto de la CTH-MD, es decir, de 6 frases/doc hasta 1 frase/doc.

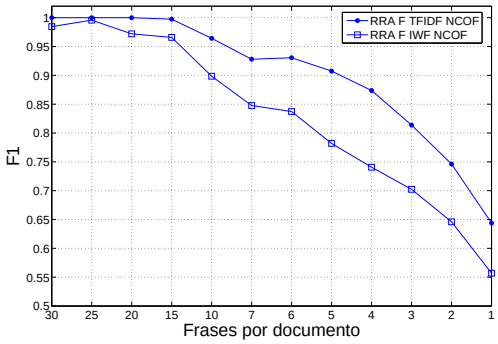
Resultados obtenidos con el método de CT basado en RRA R: El estudio de la versión reducida de la red relacional asociativa, planteada como modelo de representación y clasificación de textos (define el MEV utilizado), presenta un comportamiento distinto respecto a los resultados obtenidos con los CT basados en RRA F y NN. En este caso, tanto para el corpus reducido como para el completo, la ponderación temática IWF aporta una mejor eficiencia de clasificación que la ponderación TFIDF, independientemente de la inclusión o no de las coocurrencias en la representación del texto (ver figura 3.22), a diferencia de lo que sucedía en las estrategias de representación global de los datos (figuras 3.20 y 3.21). En lo que se refiere a los resultados obtenidos sobre el corpus completo (figuras 3.22(b) y 3.22(d)) respecto a los obtenidos sobre su versión reducida (figuras 3.22(a) y 3.22(c)), la CT basada en RRA R presenta un comportamiento claramente mejor respecto a los métodos anteriores, aunque manteniendo la tendencia de obtener peores resultados sobre el corpus completo. Además, la inclusión de las coocurrencias mejora los resultados obtenidos para el corpus completo (con resultados bastante similares a NCOF en el corpus reducido), provocando, a la vez, que los resultados obtenidos con las ponderaciones temáticas TFIDF e IWF sean bastante similares. Como resultado, la configuración IWF COF es la que presenta una mejor eficiencia de clasificación a lo largo del barrido, seguida muy de cerca por IWF NCOF, siendo la peor configuración TFIDF NCOF (curiosamente, la mejor para los métodos de CT globales).

Este comportamiento particular de la CT basada en RRA R está ligado al cambio de enfoque para la CT que esta estrategia propone. Según lo descrito en la sección 3.3.3, la CT basada en RRA R toma el texto de entrada como elemento de referencia para representar los datos a comparar, a diferencia de NN y RRA F que trabajan con un MEV global que tiene en cuenta todos los términos de cada dominio. En este contexto es donde la información global de los datos de entrenamiento tendrá un menor impacto en la eficiencia del CT, al tomar mayor importancia el peso de palabra en el texto (IWF) que su peso y/o singularidad a lo largo de la colección (TFIDF)[29] —resulta menos importante la singularidad de las palabras

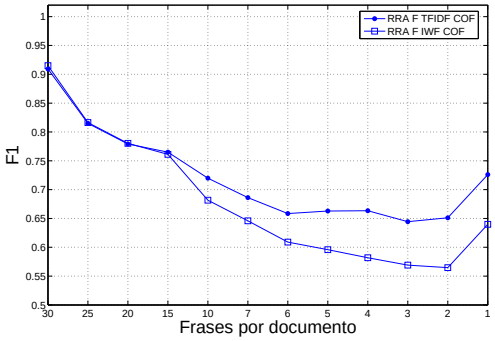
[29]Nótese también que la ponderación TFIDF realiza, de forma indirecta, el filtrado de palabras temática-



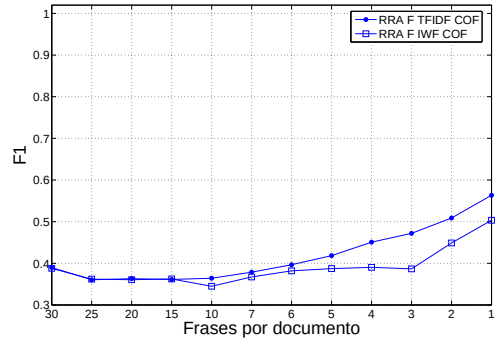
(a) Sin coocurrencias sobre el corpus reducido.



(b) Sin coocurrencias sobre el corpus completo.



(c) Con coocurrencias sobre el corpus reducido.



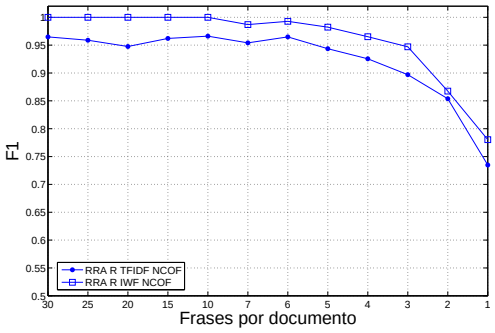
(d) Con coocurrencias sobre el corpus completo.

Figura 3.21: Eficiencia de clasificación obtenida por el clasificador de textos basado en RRA F sobre el corpus completo y su versión reducida, para distintas parametrizaciones del texto.

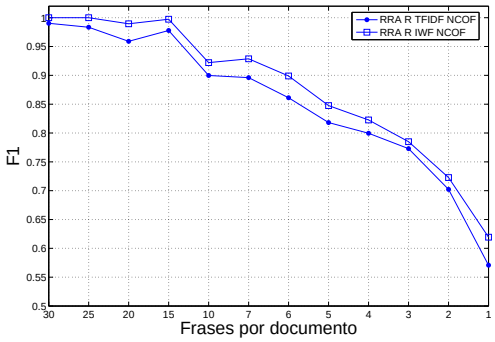
que su mera presencia. Asimismo, la parametrización estructural del texto, representada en este caso mediante la inclusión de las coocurrencias de las palabras, permite mejorar las tasas de clasificación obtenidas, a diferencia del enfoque global donde se observa un empeoramiento de los resultados —utilizando en todos los experimentos la distancia del coseno como medida de similitud.

Resumen: En la figura 3.23 se presenta una visión global de los resultados obtenidos para los tres métodos analizados, promediando los valores de F_1 obtenidos a lo largo del barrido de frases/doc presentado en las figuras 3.20, 3.21 y 3.22. Este promediado simplemente se

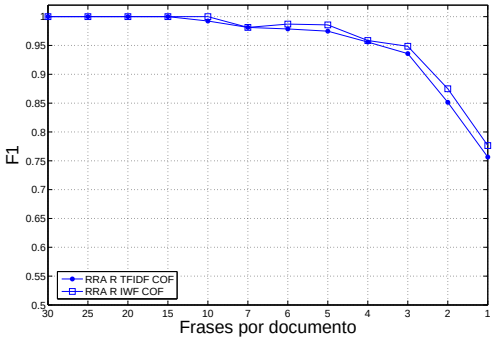
mente poco significativas, al presentar éstas (p.ej. artículos, preposiciones,...) un IDF = 0. Para el caso de utilizar la estrategia RRA R, esto tendrá un mayor impacto en la clasificación, al disponer de un espacio de comparación de tamaño menor.



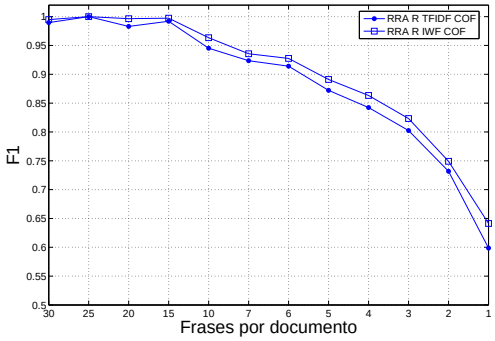
(a) Sin coocurrencias sobre el corpus reducido.



(b) Sin coocurrencias sobre el corpus completo.



(c) Con coocurrencias sobre el corpus reducido.

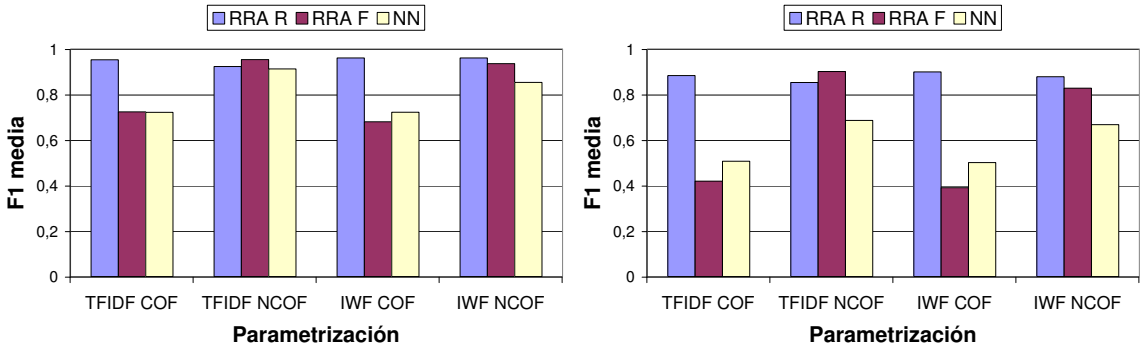


(d) Con coocurrencias sobre el corpus completo.

Figura 3.22: Eficiencia de clasificación obtenida por el clasificador de textos basado en RRA F sobre el corpus completo y su versión reducida, para distintas parametrizaciones del texto.

presenta para tener una visión global del comportamiento de todos los métodos respecto al conjunto de datos de entrenamiento (corpus completo o reducido) y a la parametrización del texto utilizada. De este modo, se puede constatar el paralelismo del funcionamiento de los modelos globales (RRA F y NN), con dificultades claras ante el corpus completo, junto a una pérdida evidente de eficiencia al incluir las coocurrencias en la representación del texto (en RRA F provoca un impacto más negativo que en NN, para el corpus completo). Asimismo, se puede observar que el clasificador de textos basado en RRA R presenta un comportamiento mucho más robusto que el resto de CT ante las distintas parametrizaciones del texto utilizadas, con una relación de la eficiencia de clasificación entre el corpus completo y el reducido muy buena —presenta una mayor robustez que el resto de métodos de CT a la presencia de frases de categoría ambigua.

Concretamente, y de forma cualitativa, se puede observar que trabajar sobre el corpus



(a) F_1 media sobre el corpus reducido.

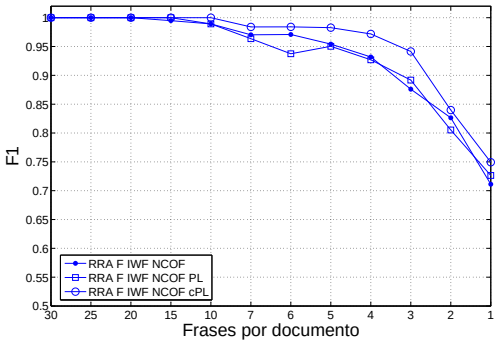
(b) F_1 media sobre el corpus completo.

Figura 3.23: Eficiencia de clasificación media de los métodos de clasificación estudiados dentro del barrido de frases por documento realizado, para distintas parametrizaciones del texto.

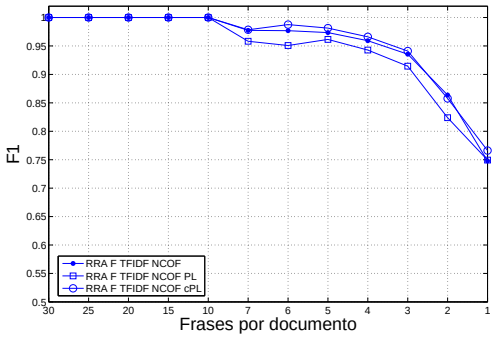
completo provoca un empeoramiento global medio (para las distintas estrategias de clasificación y las parametrizaciones utilizadas) de un 15 % en términos de F_1 respecto a trabajar sobre el corpus reducido. Así pues, disponer de frases de temática y estilo poco claros (difíciles de categorizar), implica un aumento del ruido en los datos de entrenamiento que, a su vez, provoca una reducción del rendimiento de los métodos de clasificación de textos estudiados —debido, fundamentalmente, a que se trabaja directamente sobre los textos, sin ningún tipo de filtrado (preprocesamiento), como suelen aplicar los métodos que sólo buscan la clasificación temática de los documentos. Sin embargo, este impacto es bastante desigual, según el método de CT y la configuración utilizados, yendo desde un empeoramiento del 30 % para RRA F TFIDF COF, pasando por un 22.6 % para NN TFIDF NCOF, hasta un 6.1 % para RRA R IWF COF, entre otros.

Experimento 5 - Estudio del impacto de las medidas de similitud ponderadas sobre el rendimiento de la clasificación de textos basada en RRA

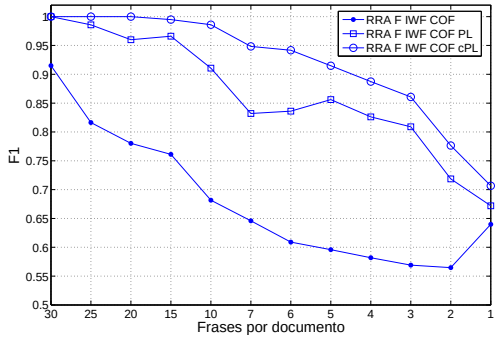
Como se ha descrito en el apartado 3.3.3, en el presente trabajo de investigación se propone incorporar la secuencialidad de los textos a clasificar respecto a los modelos de los dominios (parámetros estructurales), no sólo mediante la parametrización del texto (coocurrencias) sino también incorporando esta información como ponderación de la propia medida de similitud. Para ello, se han definido dos parámetros denominados como *longitud del patrón* y *longitud del patrón acumulada*, PL (ver ecuación (3.36)) y cPL (ver ecuación (3.38)), respectivamente. Estos parámetros, calculados a través de las RRA del texto a clasificar respecto a la RRA del dominio, se incluyen en el proceso de comparación para la clasificación del texto como ponderaciones de la medida de similitud utilizada, en este caso la distancia del coseno (ver ecuaciones (3.35) y (3.37)). A continuación se describe el impacto de estas medidas (S_2 y S_3) en la eficiencia de los métodos de clasificación basados



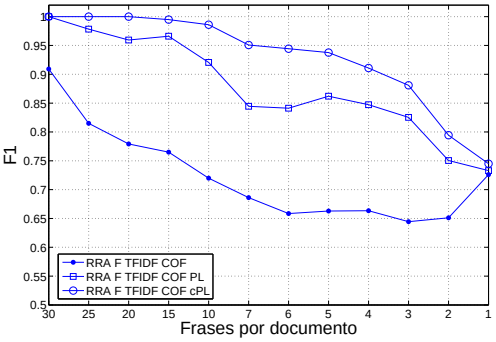
(a) Parametrización IWF sin coocurrencias.



(b) Parametrización TFIDF sin coocurrencias.



(c) Parametrización IWF con coocurrencias.

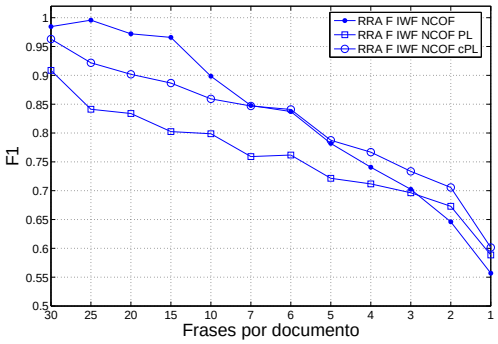


(d) Parametrización TFIDF con coocurrencias.

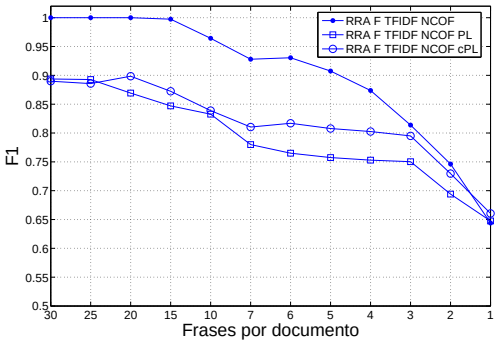
Figura 3.24: Eficiencia de clasificación obtenida por el clasificador de textos basado en RRA F sobre el corpus reducido, para distintas parametrizaciones del texto, según tres distancias de similitud: coseno (S_1), coseno ponderado por PL (S_2) y coseno ponderado por cPL (S_3).

en RRA respecto a trabajar sin ellas (S_1, ecuación (3.34)). Los resultados que se presentan a continuación constatan los primeros resultados obtenidos en los experimentos previos realizados sobre colecciones de textos mayores, donde se consideraba un mayor número de dominios (ver tablas 3.9 y 3.10), utilizando el modelo de representación de textos RRA F.

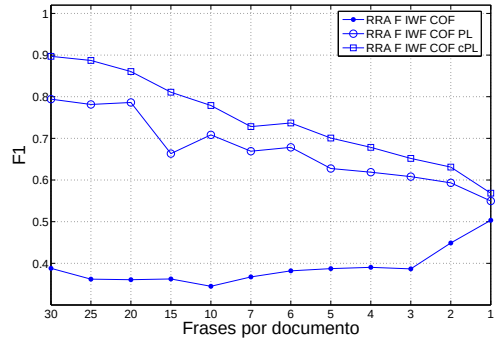
Impacto en el rendimiento de la CT basada en RRA F: En el caso de la clasificación de textos basada en RRA F, la inclusión de PL en la medida de similitud (S_2) mejora claramente el rendimiento del CT cuando se incluyen las coocurrencias (COF) en el modelado del texto, pasando de tasas mediocres de clasificación, a valores bastante buenos, tanto para el corpus reducido como para el completo (ver figuras 3.24(c), 3.24(d), 3.25(c) y 3.25(d)). Sin embargo, este efecto es menor cuando se parametriza el texto sin considerar



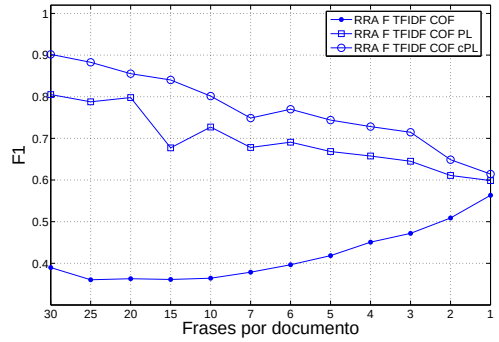
(a) Parametrización IWF sin coocurrencias.



(b) Parametrización TFIDF sin coocurrencias.



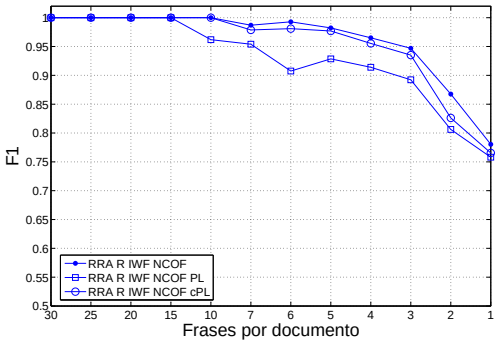
(c) Parametrización IWF con coocurrencias.



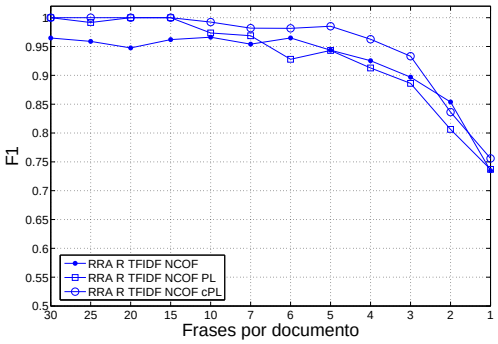
(d) Parametrización TFIDF con coocurrencias.

Figura 3.25: Eficiencia de clasificación obtenida por el clasificador de textos basado en RRA F sobre el corpus completo, para distintas parametrizaciones del texto, según tres distancias de similitud: coseno (S_1), coseno ponderado por PL (S_2) y coseno ponderado por cPL (S_3).

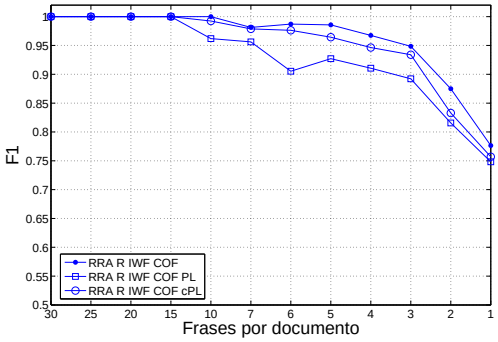
las coocurrencias de los términos que lo componen, obteniendo unos resultados similares a los obtenidos sin PL, en el caso del corpus reducido, y algo peores para el corpus completo, excepto en algunos puntos del barrido de frases/doc (ver figuras 3.24(a), 3.24(b), 3.25(a) y 3.25(b)). No obstante, cabe resaltar la mejora conseguida al trabajar con S_2 ante S_1 en el punto más crítico de la clasificación: 1 frase/doc, independientemente de la parametrización de los textos utilizada (ver figuras 3.24 y 3.25). Por otro lado, sustituir PL (número máximo de términos consecutivos) por su versión acumulada cPL (número total de términos consecutivos) como ponderación de la distancia del coseno (distancia S_3), consigue una mejora significativa de los resultados obtenidos respecto a los conseguidos mediante S_2 (ver figuras 3.24 y 3.25) —incluso para el caso de 1 frase/doc. Asimismo, cabe destacar los resultados obtenidos mediante cPL en el corpus reducido (figura 3.24), que consigue mejorar las con-



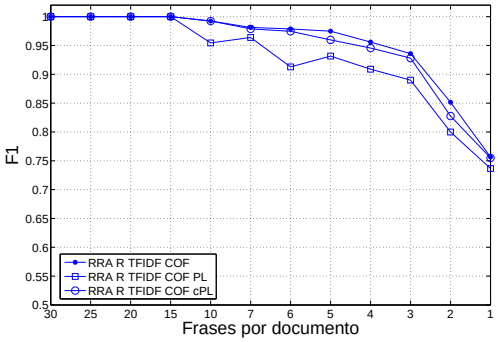
(a) Parametrización IWF sin coocurrencias.



(b) Parametrización TFIDF sin coocurrencias.



(c) Parametrización IWF con coocurrencias.

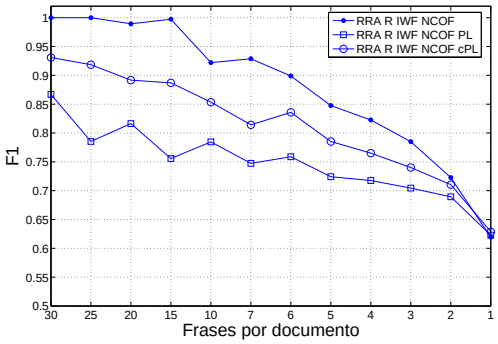


(d) Parametrización TFIDF con coocurrencias.

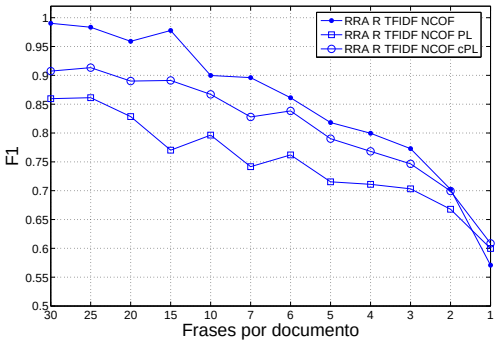
Figura 3.26: Eficiencia de clasificación obtenida por el clasificador de textos basado en RRA F sobre el corpus reducido, para distintas parametrizaciones del texto, según tres distancias de similitud: coseno (S_1), coseno ponderado por PL (S_2) y coseno ponderado por cPL (S_3).

figuraciones óptimas de la CT basada RRA F (NCOF), acercando significativamente los resultados obtenidos con las configuraciones COF y NCOF.

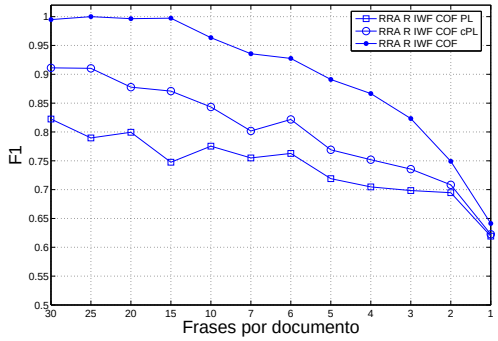
Así pues, la inclusión de la ponderación estructural en la medida de similitud utilizada (distancia del coseno) permite, en el caso de la CT basada en RRA F, compensar el impacto del aumento del tamaño de los vectores al incluir COF en la parametrización, mejorando (ligeramente) su rendimiento al utilizar la ponderación temática TFIDF respecto a IWF, pero sin llegar a ser de gran ayuda en el contexto del corpus reducido respecto a la configuración óptima de la CT basada en RRA F: TFIDF NCOF con la distancia del coseno. De todos modos, cabe destacar la mejora conseguida en el punto del barrido donde se obtienen peores tasas de clasificación (1 frase/doc), cuestión especialmente interesante en el contexto de la CTH-MD.



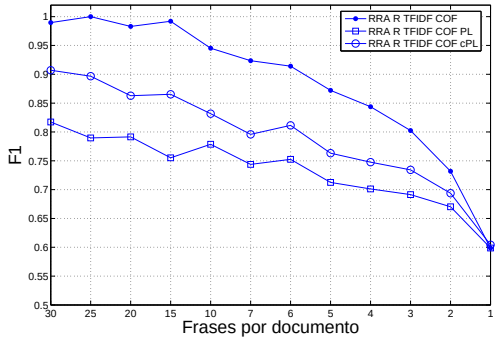
(a) Parametrización IWF sin coocurrencias.



(b) Parametrización TFIDF sin coocurrencias.



(c) Parametrización IWF con coocurrencias.



(d) Parametrización TFIDF con coocurrencias.

Figura 3.27: Eficiencia de clasificación obtenida por el clasificador de textos basado en RRA F sobre el corpus completo, para distintas parametrizaciones del texto, según tres distancias de similitud: : coseno (S_1), coseno ponderado por PL (S_2) y coseno ponderado por cPL (S_3).

Impacto en el rendimiento en la CT basada en RRA R: A diferencia de las mejoras que se obtienen sobre la CT basada en RRA F al incluir las ponderaciones PL (S_2) o cPL (S_3) en la distancia de similitud, el impacto que estas distancias tienen sobre el rendimiento de la CT basada en RRA R es bastante distinto. Como se puede observar en la figura 3.26, los resultados obtenidos sólo mejoran los conseguidos con TFIDF NCOF ponderando la distancia del coseno con cPL (con PL, sólo para más de 7 frases/doc). Para el caso del corpus completo, los resultados son, en general, bastante peores que los obtenidos con S_1 (ver figura 3.27). No obstante, al igual que sucedía con la RRA F, la distancia S_3 con cPL presenta unos mejores resultados que la S_2 con PL, en la comparativa relativa entre ambas. Así pues, queda claro que estas ponderaciones, en el contexto de la CT según el espacio vectorial definido por la RRA R, en general no ayudan —con alguna excepción— a mejorar

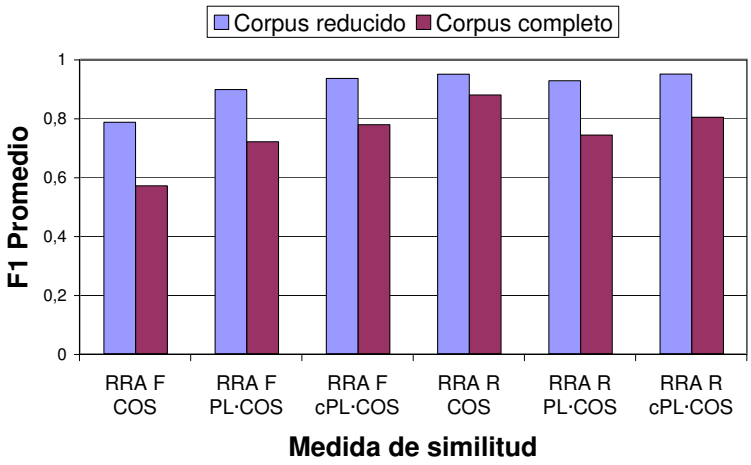


Figura 3.28: Eficiencia de clasificación promedio en el barrido de frases por documento realizado, al incluir los parámetros PL y cPL en la medida de similitud utilizada para los CT basados en RRA F y RRA R, tanto en el corpus completo como en el reducido.

la tasa de clasificación (F_1) respecto a utilizar simplemente la distancia del coseno.

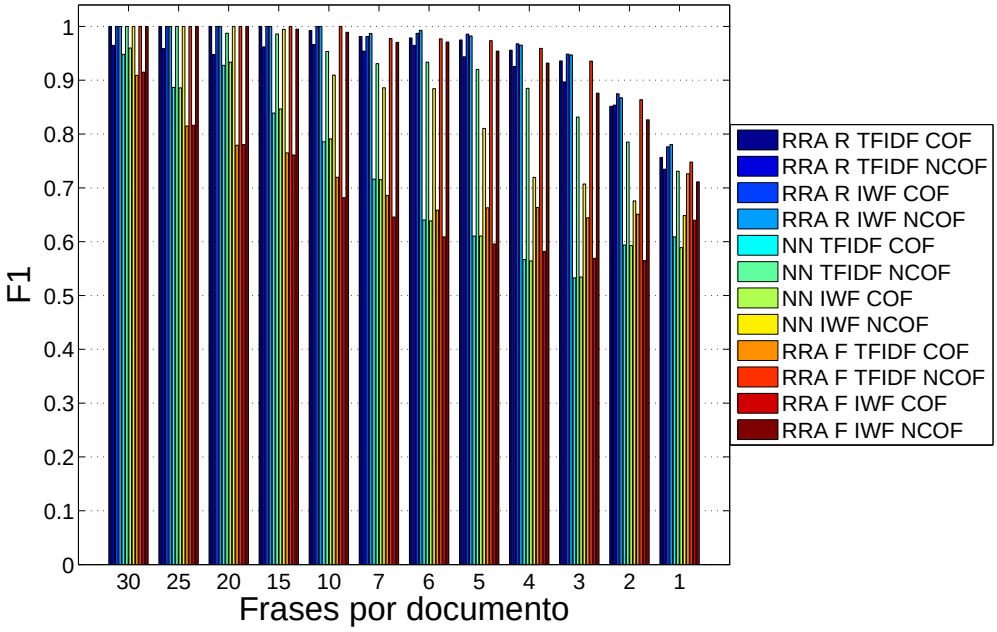
Resumen: Como resumen de este estudio, se presenta una comparativa global (cualitativa) del impacto del uso de las distancias ponderadas (S_2 en ecuación (3.35) y S_3 en ecuación (3.37)) respecto a la distancia del coseno de referencia (S_1 en ecuación (3.34)), promediando los resultados obtenidos para las distintas configuraciones estudiadas de la CT basada en RRA F y RRA R a lo largo del barrido de frases por documento estudiado. En la figura 3.28 se puede observar la tendencia de mejora que se obtiene al incluir primero PL y, a continuación cPL, como ponderaciones de la distancia del coseno en el caso de la CT basada en RRA F, tendencia no observable en el caso de la CT basada en RRA R. Concretamente, en RRA F, incluir PL implica conseguir una mejora relativa en F_1, respecto a trabajar simplemente con la distancia del coseno, del 14.2 % para el corpus reducido y del 20.8 % para el corpus completo. Asimismo, ponderarla mediante cPL consigue unas mejoras del 19 % y el 36.4 %, para los mismos casos. Así pues, se demuestra que para RRA F, la ponderación cPL es más óptima que PL, ya que minimiza el efecto no deseado de los patrones locales de palabras, que son menos significativos para la clasificación.

En cambio, para RRA R el uso de S_2 y S_3 como medidas de similitud, por un lado, prácticamente no afecta a los resultados obtenidos sobre el corpus reducido, mientras que por otro, afecta negativamente a los resultados obtenidos sobre el corpus completo, especialmente en el caso de utilizar la ponderación PL. Por lo tanto, se deduce, para los experimentos desarrollados, que el impacto sobre la CT basada en RRA R de las ponderaciones de patrón sobre la distancia del coseno no tiene el efecto conseguido al utilizar RRA F. Intuitivamente, se puede deducir que, en el caso que las palabras coincidentes entre el vector a clasificar y el modelo del dominio no sean discriminativas (p.ej. preposiciones, artículos, etc.), la distancia

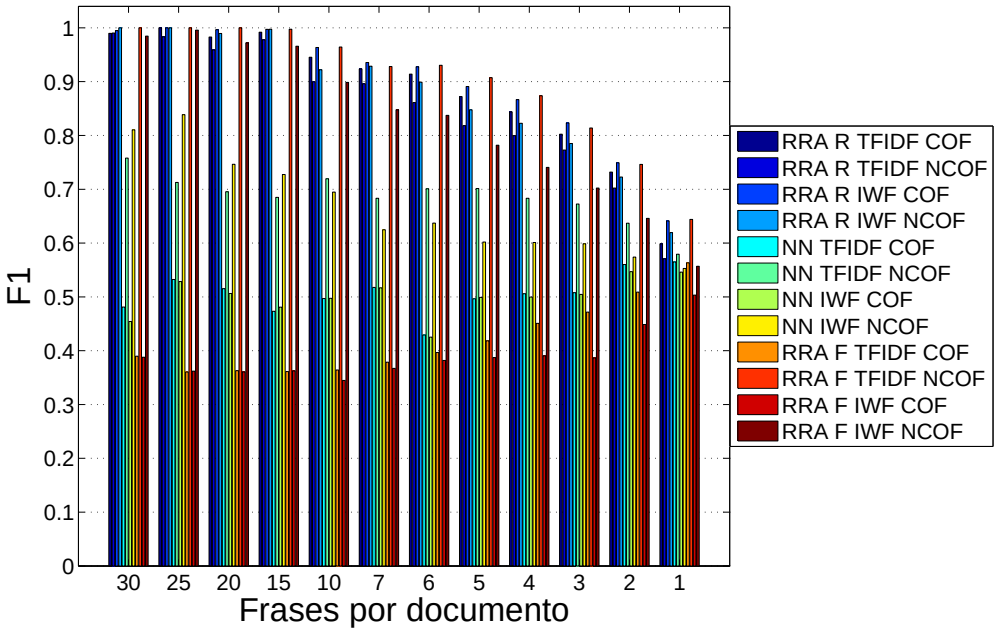
ponderada puede sesgar la clasificación hacia un dominio erróneo en el caso de la RRA R (tamaño del vector reducido), guiado por la coincidencia de términos, que de entrada, parecen poco significativos —cuestión que tiene un impacto mucho menor en el caso de la RRA F debido al mayor tamaño de los vectores patrón utilizados. No obstante, esta cuestión deberá ser estudiada con más detalle en futuras investigaciones.

Análisis final: Finalmente, las figuras 3.29 y 3.30 presentan un resumen de los resultados obtenidos del barrido de frases por documento realizado para todos los métodos de CT y todas las parametrizaciones del texto estudiadas, utilizando la distancia del coseno (S_1) y la distancia del coseno ponderada por la *longitud del patrón acumulada* (cPL) (S_3) como medidas de similitud (S_3 ha presentado mejor comportamiento que S_2 en el estudio previo), sobre las versiones completa y reducida del corpus C_{Pub}. Después de haber detallado el comportamiento de cada método de clasificación para los distintas parametrizaciones y distancias de similitud estudiadas, este análisis sólo pretende dar una visión global de los resultados obtenidos sobre el problema de la CT planteado, dadas las especificaciones de trabajo definidas. En este sentido, se puede observar como las mejores tasas de clasificación (F_1) se consiguen sobre el corpus reducido, en el cual se han filtrado de frases difícilmente asignables a uno u otro dominio (frases temáticamente ambiguas), independientemente del método de clasificación utilizado (ver figuras 3.29(a) y 3.30(a)). Sin embargo, al trabajar sobre el corpus completo, estos resultados empeoran de forma considerable (del orden del 15 % en media), fundamentalmente, debido al aumento del ruido en los datos de entrenamiento, como se ha discutido a lo largo de los experimentos (ver figuras 3.29(b) y 3.30(b)). No obstante, cabe destacar que la CT basada en RRA R demuestra ser el método de clasificación más robusto al trabajar sobre el corpus completo —consiguiendo el CT basado en RRA F también buenos resultados cuando se trabaja con la distancia del coseno ponderada por la longitud del patrón acumulada.

Por otro lado, en lo que se refiere a la medida de similitud utilizada en el proceso de clasificación del texto, se puede observar que, independientemente del corpus considerado, los resultados obtenidos para las distintas parametrizaciones de los CT basados en RRA F tienden a igualar e incluso mejorar sus prestaciones (de forma muy evidente para las peores configuraciones de RRA F), a diferencia de lo que sucede con S_1 donde existen unas parametrizaciones con unas prestaciones claramente superiores a las otras (p.ej. TFIDF NCOF). No obstante, este comportamiento que es muy positivo al trabajar sobre el corpus reducido, tiene un impacto negativo cuando se trabaja sobre el corpus completo, ya que provoca un empeoramiento global del orden del 10 % en la eficiencia de la clasificación respecto a utilizar la distancia del coseno —a excepción de trabajar con 2 o 1 frase/doc, donde los resultados se igualan. A tenor de estos resultados, se puede deducir que se debe ser muy cuidadoso en la elección de los datos de entrenamiento para este tipo de estrategias de clasificación que trabajan sobre *todos* los datos (sin preprocesarlos) y siguen una filosofía de *pattern matching* —cuestión que se acentúa al ponderar la distancia del coseno por el número de palabras coincidentes entre el texto a clasificar y el modelo de cada dominio. No obstante, la estrategia de CT basada en RRA R es la que presenta una mayor robustez frente a este problema, como se ha demostrado a lo largo de los experimentos.

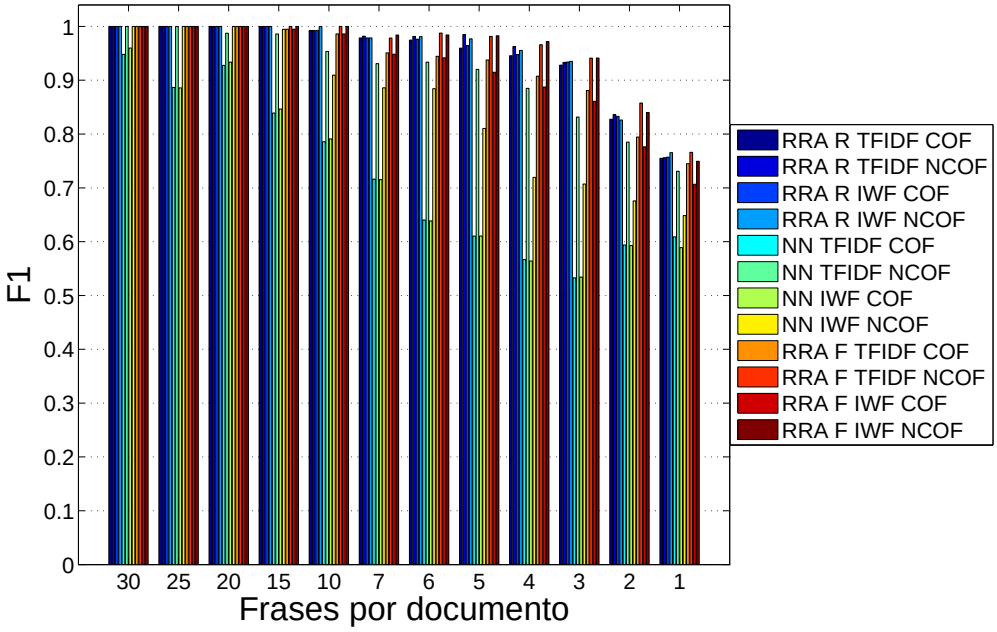


(a) Distancia del coseno sobre el corpus reducido.

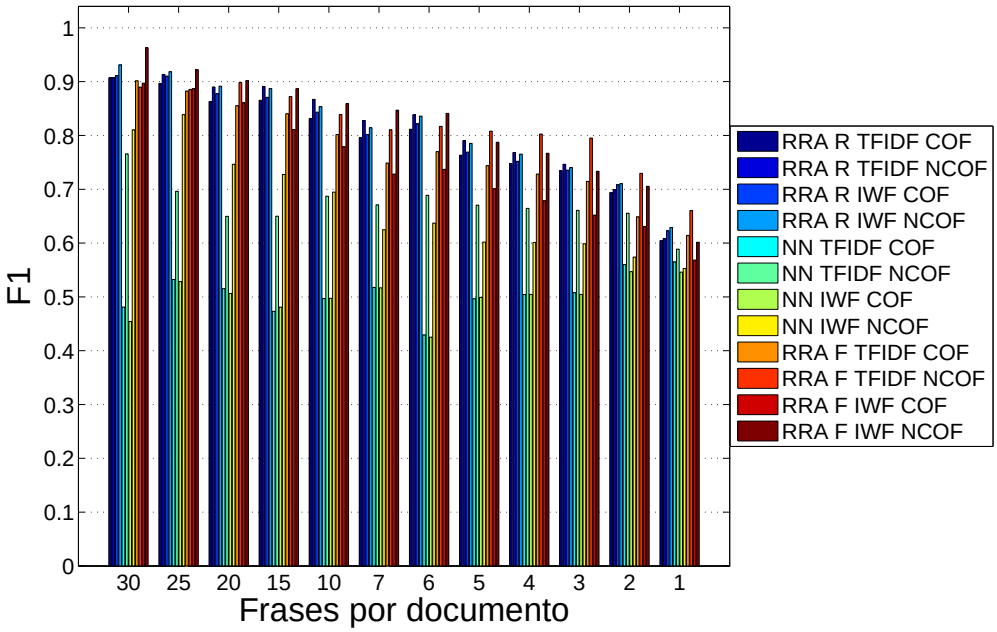


(b) Distancia del coseno sobre el corpus completo.

Figura 3.29: Eficiencia de clasificación obtenida por los tres métodos de CT estudiados sobre el corpus completo y su versión reducida, para distintas parametrizaciones del texto utilizando la distancia del coseno.



(a) Distancia del coseno ponderada por cPL sobre el corpus reducido.



(b) Distancia del coseno ponderada por cPL sobre el corpus completo.

Figura 3.30: Eficiencia de clasificación obtenida por los tres métodos de CT estudiados sobre el corpus completo y su versión reducida, para distintas parametrizaciones del texto utilizando la distancia del coseno ponderada por cPL.

Por otro lado, del análisis de las figuras 3.29 y 3.30 se puede observar cómo los métodos de CT basados en RRA presentan una mejor eficiencia de clasificación que NN. En este sentido, la CT basada en RRA R es el método que muestra una respuesta global más robusta respecto a las distintas parametrizaciones y medidas de similitud consideradas, en contraposición a métodos globales (RRA F y NN) que tienen un rendimiento más dependiente de la parametrización y la distancia de similitud escogidas. Concretamente, la inclusión de COF, trabajando simplemente con la distancia del coseno como medida de similitud, provoca un impacto negativo en los CT globales, con un mayor impacto en la CT basada en RRA F que en la basada en NN por el hecho de trabajar con un único vector patrón por dominio, en lugar de usar un vector por documento. Sin embargo, la CT basada en RRA F consigue una mejor respuesta que NN con las parametrizaciones sin coocurrencias (NCOF), tanto para la ponderación IWF como TFIDF. Por otro lado, incluir información estructural mediante la ponderación cPL sobre la distancia de similitud provoca un efecto positivo en el rendimiento del CT basado en RRA F, al equiparar las configuraciones con menor tasa de clasificación (COF) con las que presentan mejores prestaciones (NCOF) utilizando la distancia del coseno —con una respuesta no tan positiva para el CT basado en RRA R—, superando o igualando para cualquier configuración los mejores resultados obtenidos por el método de referencia. Finalmente, destacar que todos los métodos de CT comparados tienden a obtener resultados similares a medida que se reduce el número de frases por documento, fundamentalmente, debido al volumen reducido de datos de que dispone el clasificador, sea cual sea la distancia de similitud o parametrización del texto utilizadas.

Estos resultados demuestran como la CT basada en RRA, presenta un comportamiento mejor al obtenido por la CT basada en NN (método de referencia) en términos de eficiencia (F_1) y de robustez de la clasificación (versión completa o reducida del corpus, parametrización del texto y medida de similitud utilizadas), siempre que se escoja la configuración adecuada al modelo de representación de los datos escogido (RRA F o RRA R). Asimismo, la RRA R, aunque definida como aproximación de la CT basada en RRA F, muestra una respuesta muy buena en los experimentos, llegando a mejorar las configuraciones óptimas de la RRA F en algunos casos. Asimismo, la CT basada en NN presenta, en general, sólo un buen comportamiento para la parametrización TFIDF NCOF con un gran número de frases/documento sobre la versión reducida del corpus C_{Pub}, empeorando de forma evidente su rendimiento al trabajar sobre el corpus completo.

Experimento 6 - Estudio del coste computacional

Una vez estudiada la eficiencia de los métodos de clasificación de textos propuestos, se pasa a analizar el tiempo de ejecución de los mismos, cuestión también clave en el contexto de la CTH-MD. En este sentido, se realiza un estudio del coste computacional de cada uno de los métodos de CT estudiados hasta el momento, basados en: RRA R, RRA F y NN. En este estudio, se analiza tanto el coste de clasificación (test) como el de construcción del clasificador (entrenamiento) —se estudian ambos costes[30] para disponer de una visión global del comportamiento computacional de los métodos analizados, aunque el tiempo

[30]Para todos los costes computacionales se incluye el tiempo de lectura de datos y de acceso a disco.

de entrenamiento no es un elemento crítico para la CTH-MD, ya que los clasificadores se construyen de antemano (proceso *off-line*).

En el caso del CT basado en NN, la fase de entrenamiento simplemente hace referencia a la parametrización del texto y a su representación en el MEV extendido (términos, coocurrencias, etc.) (ver ecuación (3.2)). Sin embargo, en el caso del CT basado en RRA F, el tiempo de entrenamiento, además de la parametrización del texto, incluye la representación global de los datos mediante la RRA *global* junto a la generación de los vectores patrón ($\vec{p_n}$) de las RRA F de los dominios. En lo que se refiere al tiempo de clasificación (elemento clave para no ralentizar la síntesis), para el CT basado en NN se realizan tantas comparaciones (cálculos de distancias de similitud) como documentos contiene el MEV definido, en cambio, para el CT basado en RRA, sólo se realizarán $|\mathcal{C}|$ comparaciones —una por vector patrón—, donde $|\mathcal{C}|$ corresponde al número de dominios (categorías) considerados en la clasificación. La diferencia entre RRA F y RRA R residirá en el tamaño de los vectores comparados, cuestión que afecta al tiempo de cálculo de las distancias de similitud utilizadas (basadas en la distancia del coseno). Asimismo, el tiempo necesario para representar los datos sobre el espacio común de comparación es distinto. Para el CT basado en RRA F, el texto a clasificar debe representarse (orden de los términos y parámetros) según la RRA global, mientras que para el CT basado en RRA R, los dominios se representan mediante los vectores patrón aproximados ($\hat{\vec{p}_n}$) sobre el espacio vectorial definido por la RRA del texto a clasificar.

El análisis del coste computacional se obtiene como el tiempo medio (de clasificación o entrenamiento) sobre 10 ejecuciones de 1-*fold* del *random subsampling* utilizado en las pruebas de análisis de eficiencia de clasificación —estrategia utilizada para aumentar la consistencia estadística de los resultados respecto a los datos utilizados en las pruebas (ver introducción de la sección 3.4.2). El estudio del coste computacional se ha realizado sobre un PC (PIV 1.79GHz - 1 GB RAM) con sistema operativo Linux y compilador gcc 3.3.5. Las pruebas se han llevado a cabo sobre la versión reducida del corpus. No obstante, dada la distribución de las palabras por documento presentada en la figura 3.19, estos resultados son también extrapolables al corpus completo.

Análisis de los resultados: Para el método de clasificación de textos basado en NN, el estudio muestra una clara diferencia de comportamiento cuando se comparan las configuraciones que utilizan las coocurrencias (COF) con las que no las utilizan (NCOF) (ver figura 3.31). Debido al aumento de la dimensión del espacio vectorial cuando se trabaja con vectores que incluyen COF, el coste computacional del clasificador se dispara respecto a trabajar sin coocurrencias. Este efecto se hace evidente tanto en la fase de entrenamiento como en la de test. Por otro lado, el coste computacional de la NN crece de forma más o menos exponencial a lo largo del barrido de frases por documento sobre el que se realiza el estudio. En la fase de explotación, este comportamiento se argumenta por el aumento del número de comparaciones necesarias para determinar la categoría del texto de entrada a

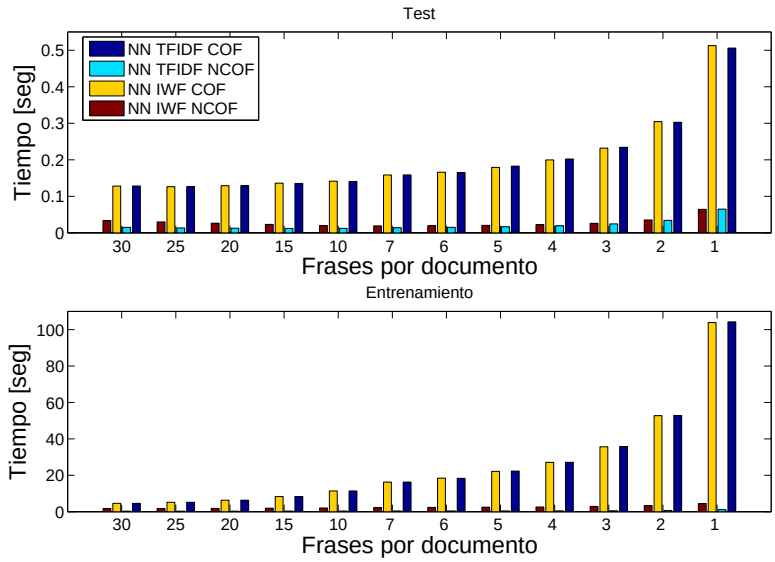


Figura 3.31: Coste computacional a lo largo del barrido de frases por documento realizado sobre el corpus reducido para el clasificador de textos basado en NN, tanto en test como en entrenamiento.

medida que se reduce el número de frases/doc[31], para un número constante de documentos (ver tabla 3.14). Sin embargo, si el número de documentos de comparación se mantuviera constante a lo largo del barrido, el coste computacional sería más estable —variación debida al coste computacional correspondiente al cálculo de la distancia del coseno entre los vectores, evitando la dependencia respecto al número de datos disponibles. Esta cuestión queda minimizada en el caso de los métodos RRA, donde el número de comparaciones es constante (una por vector patrón de dominio), independientemente del volumen de datos de entrenamiento.

Así pues, parece lógico que el rendimiento del método de clasificación de textos basado en RRA F sea bastante distinto al NN, en términos de coste computacional, presentando un comportamiento más estable a lo largo del barrido de frases/doc realizado (ver figura 3.32). La evolución del tiempo de clasificación en la fase de test utilizando RRA F es inversa a la que se obtiene con NN. En este caso, tamaños importantes de textos en la fase de test implican tiempos algo mayores para su representación en la RRA global, ya que los vectores serán de mayor tamaño (mayor número de parámetros), a diferencia de NN, donde esto significa reducir el número de comparaciones en clasificación (según la estrategia utilizada en los experimentos). Asimismo, la diferencia evidente entre los costes computacionales al incluir COF respecto NCOF que se observa en el CT basado en NN también está presente en RRA F, donde las versiones que incluyen coocurrencias en la representación de los textos

[31] El número de documentos crece siguiendo un tendencia exponencial al decrementar linealmente el número de frases por documento.

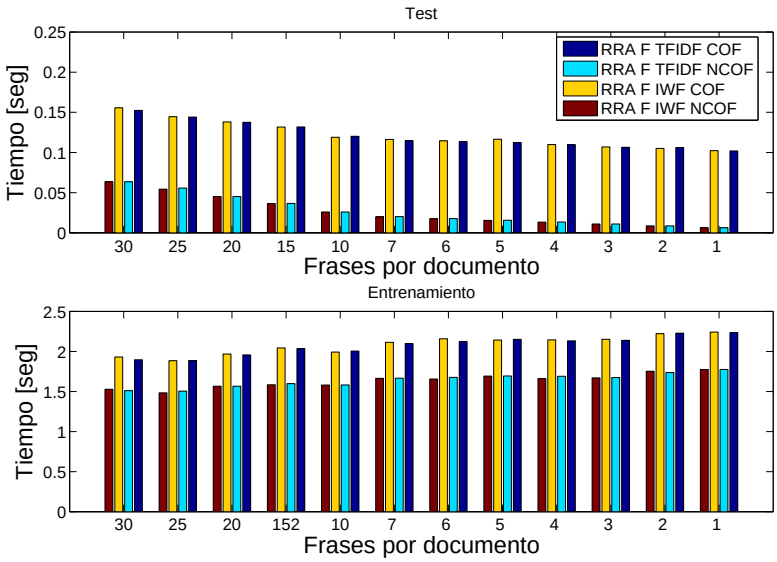


Figura 3.32: Coste computacional a lo largo del barrido de frases por documento realizado sobre el corpus reducido para el clasificador de textos basado en RRA F, tanto en test como en entrenamiento.

continúan presentando mayor coste computacional debido al aumento del tamaño de los vectores analizados.

Finalmente, el método de CT basado en RRA R es el que presenta los mejores resultados en términos del coste computacional, independientemente de la parametrización utilizada, con un comportamiento *casi* constante para cada uno de los puntos del barrido en la fase de entrenamiento (algo más que en RRA F). Además, el comportamiento global a lo largo del barrido del método durante la fase de test es distinto a los otros dos métodos analizados. Concretamente, a medida que disminuye el tamaño de los textos a clasificar, el coste computacional también disminuye de forma bastante proporcional (en dos tramos lineales con cambio de pendiente alrededor de 10 frases/doc). Este comportamiento decreciente se debe a que el espacio de comparación entre vectores lo define el texto a clasificar, por lo que, cuanto menor sea su tamaño, menor será su dimensión, reduciéndose el tiempo de representación y comparación de los datos. Comparándolo con el CT basado en RRA F, este método presenta unos costes computacionales menores, con reducciones relativas que van desde un 12 % para 30 frases/doc, pasando por un 47 % para 4 frases/doc, hasta un 120 % para 1 frase/doc, utilizando sus configuraciones óptimas en términos de tiempo de ejecución (TFIDF NCOF para RRA F y IWF NCOF para RRA R, aunque las variaciones en RRA R son mínimas). Por lo tanto, parece claro que el método de CT basado en RRA R consigue los mejores resultados en lo que se refiere al coste computacional en tiempo de clasificación, con una variación *proporcional* al tamaño de los datos a clasificar.

Así pues, del análisis del coste computacional de test se deduce que el CT basado RRA

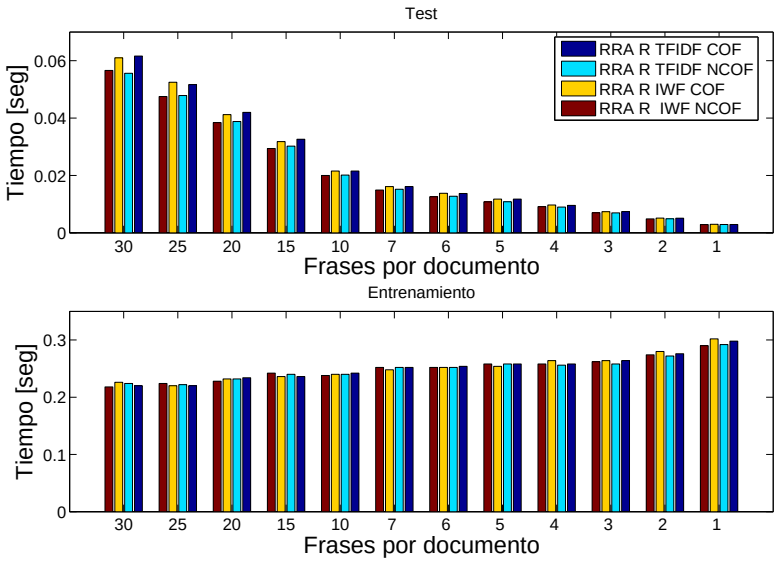


Figura 3.33: Coste computacional a lo largo del barrido de frases por documento realizado sobre el corpus reducido para el clasificador de textos basado en RRA R, tanto en test como en entrenamiento.

R es el más óptimo en términos de coste computacional, sobretodo en la zona del barrido que presenta un número medio/bajo de frases por documento —zona potencialmente más utilizada en el ámbito de la síntesis del habla— independientemente de la parametrización escogida. En segundo lugar se encuentra el CT basado en RRA F, que aunque con mayor coste computacional que el basado en RRA R, continua presentando mejores resultados para un número reducido de frases/doc que la NN. Finalmente, comentar que a medida que aumenta el número de frases por documento, los tiempos de clasificación convergen, llegando a presentar la NN los mejores resultados para grandes tamaños de texto —zona de menor interés para CTH. Esto es debido, fundamentalmente, al efecto de la reducción del número de vectores a comparar sobre el que se ha basado la presente comparativa. Bajo este mismo enfoque, en la zona de 10 a 1 frases/doc, los costes computacionales del CT basado en NN son mucho mayores que los de los clasificadores RRA, para sus configuraciones óptimas en términos de coste computacional. Finalmente añadir que, aunque la mayoría de métodos de CT estudiados presentan costes de computación de clasificación menores al segundo (< 0.07), cuanto menor sea este tiempo de clasificación, menos se sobrecargará el proceso de conversión de texto en habla multidominio debido a la incorporación del módulo de clasificación de textos.

Coste computacional *vs.* eficiencia de clasificación: Como estudio final de la comparación entre los métodos de clasificación estudiados, se presenta un pequeño análisis que permite comparar el rendimiento de los CT basados en RRA F, RRA R y NN en térmi-

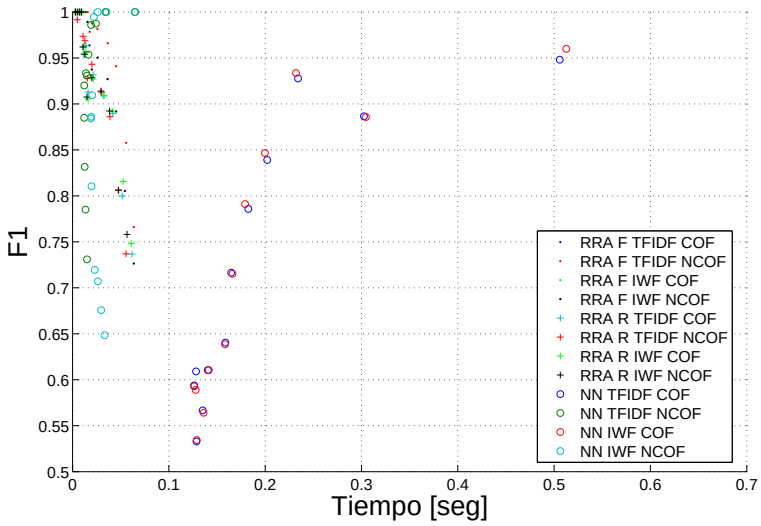
nos de eficiencia de clasificación (F_1) respecto al coste computacional que presentan. En este contexto, el mejor método será aquel que presente un compromiso *óptimo* entre ambos parámetros, es decir, que maximice su rendimiento de clasificación y minimice el coste computacional de la clasificación. La figura 3.34 presenta los resultados obtenidos para todos los métodos y sus configuraciones, en este caso, sobre el corpus reducido y utilizando S_3 (distancia del coseno ponderada por cPL, ecuación (3.37)). Cada método está representado por los 12 valores del barrido de frases/documento ($\{30, 25, 20, 15, 10, 7, 6, 5, 4, 3, 2, 1\}$) sobre el que se ha venido realizando los experimentos previos. Para cada configuración, los puntos con menor F_1 corresponden a los resultados obtenidos para 1 frase/doc, mientras que los de mayor F_1 corresponden a 25 o 30 frases/doc (como muestra la figura 3.30(a)).

De la figura 3.34(a) se puede observar claramente que las configuraciones COF del CT basado en NN son las más costosas en términos de coste computacional. Para poder comparar con más claridad el resto de resultados, la figura 3.34(b) presenta los mismos resultados que la figura anterior, pero eliminando estas dos configuraciones (NN TFIDF COF y NN IWF COF). En esta segunda figura se puede observar cómo el método de clasificación de textos basado en RRA R es el que presenta el mejor compromiso $F_1 - \textit{coste computacional}$ a lo largo del barrido, seguido por el basado en RRA F. Concretamente, para los mismos puntos del barrido de frases/doc, el CT basado en RRA R presenta menores costes computacionales, con F_1 muy parecidas, respecto a RRA F. Finalmente, añadir que las dos configuraciones de CT basadas en NN NCOF presentan peores resultados (compromiso) que las obtenidos con los clasificadores de textos basados en RRA.

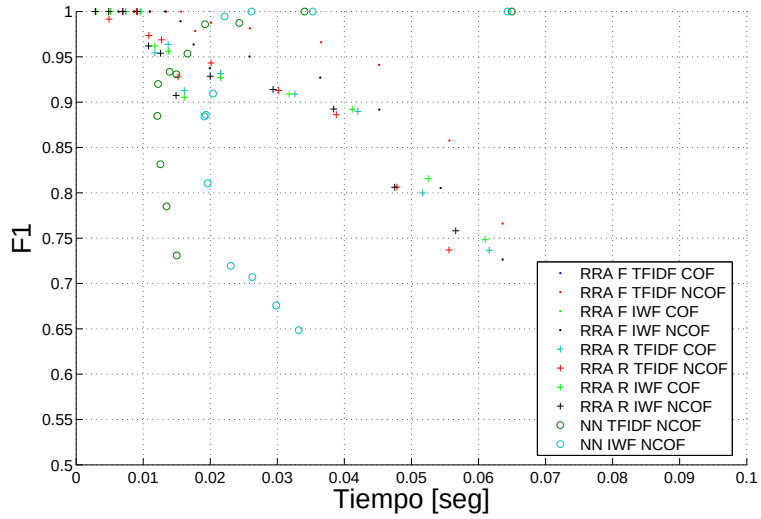
Así pues, se confirma que la propuesta de CT basada en RRA, y más concretamente la basada RRA R, es la que permite conseguir el mejor comportamiento en términos del binomio coste computacional *vs.* eficiencia de clasificación, cuestiones claves para la elección del método de asignación automática de dominio necesario para implementar la presente propuesta de sistema de CTH-MD. No obstante, el CT basado en RRA F también presenta buenos resultados, mejorando las prestaciones obtenidas con el método de clasificación de textos de referencia, basado en NN.

Experimento 7 - Evaluación subjetiva de la CTH-MD, considerando el impacto del funcionamiento del CT desarrollado

En este experimento se pretende evaluar de forma subjetiva los resultados sintéticos obtenidos del sistema de CTH-MD implementado, utilizando como clasificador de textos el sistema basado en la red relacional asociativa reducida (RRA R) con la parametrización IWF e incluyendo las coocurrencias (COF) de las palabras (configuración óptima del CT en términos de eficiencia de clasificación, utilizando la distancia del coseno ponderada por cPL) (Alías et al., 2006). Las pruebas se han realizado a nivel de 1 frase/doc (tamaño de documento mínimo del barrido estudiado, pero bastante habitual en el contexto de la CTH) sobre el corpus publicitario (C_{Pub}), el cual está dividido en tres categorías, etiquetadas como: educación, tecnología y cosmética, cuyos contenidos han sido grabados utilizando el estilo alegre (ALE), neutro (NEU) y sensual (SEN), respectivamente, tal y como se ha indicado en la introducción del presente apartado. El conjunto de frases de test (20 % del total de



(a) Comparativa global.



(b) Comparativa excluyendo las configuraciones COF del CT basado en NN.

Figura 3.34: Compromiso entre el coste computacional y la tasa de clasificación obtenidos para los tres métodos de CT estudiados sobre el corpus reducido, para distintas parametrizaciones del texto y utilizando la distancia del coseno ponderada por cPL.

Tabla 3.15: Listado de frases correctamente clasificadas que se utilizan para evaluar la calidad sintética de la CTH-MD sobre el corpus publicitario estudiado.

1	Aprenda idiomas con CCC.
2	El curso de inglés para toda la familia.
3	El libro más divertido.
4	En teoría una escuela de negocios.
5	Enciclopedia de la naturaleza de España.
6	Formación ambiental a distancia.
7	Ha llegado el momento de aprender idiomas con la tecnología del nuevo milenio.
8	La enciclopedia definitiva.
9	La obra de consulta más actual y completa.
10	Las mejores obras del mundo.
11	Las obras más bellas del mundo.
12	Libro de la competición 93.

(a) Frases alegres bien clasificadas.

1	Aire de mujer.
2	Antibolsas y antiojeras.
3	Cada mujer tiene un tipo de piel.
4	Colección de maquillaje para la primavera del 2000.
5	La nueva firma de moda para mujer del nuevo milenio.
6	Máxima protección demostrada clínicamente.
7	Quédese con lo mejor del sol.
8	Nuevas sensaciones.
9	Presentando la nueva fragancia para hombre.
10	Los salvavidas de su piel.
11	Pura relajación para sus sentidos, puro tratamiento para tu cuerpo.
12	Quédese con lo mejor del sol.
13	Color intenso, mayor duración.
14	Si elige el mejor tratamiento para su rostro, le obsequiamos con los mejores desmaquillantes.
15	Vaqueros Armani.

(b) Frases sensuales bien clasificadas.

frases de cada categoría) se extrae del corpus para evitar la elección de sus unidades durante

la conversión de texto en habla.

En cuanto a la síntesis, el módulo de selección de unidades se ajusta para encontrar la secuencia de unidades más larga posible (siguiendo la misma filosofía descrita en el "*Experimento 2*" de la presente sección —ver ecuación (3.42)) dentro del dominio indicado por el clasificador de textos. A nivel prosódico, cuando el CT acierta el dominio del texto, se utiliza la prosodia de la frase grabada (estrategia *copy prosody*) como prosodia objetivo del texto a sintetizar. Sin embargo, cuando el CT asigna el texto a un dominio distinto al que está grabado, resulta necesario estimar la prosodia del texto sobre ese dominio a partir del modelo prosódico correspondiente (ver (Miralles, 2005) para más detalles).

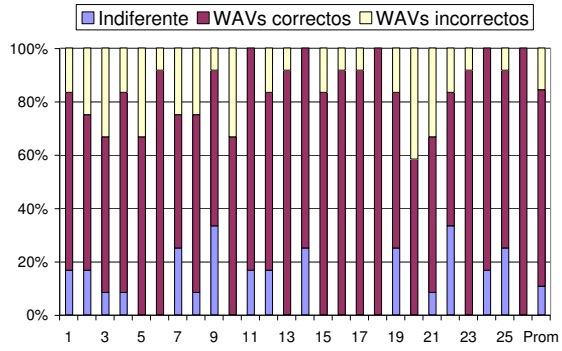
El experimento subjetivo se divide en dos partes para poder analizar el impacto del funcionamiento del clasificador de textos (basado en RRA R) en el rendimiento del sistema de síntesis multidominio. En ambas pruebas se presenta a los evaluadores dos versiones sintéticas de la misma frase para que éste escoja la que le parezca más *adecuada* (calidad conseguida y estilo de locución utilizado). Además, los evaluadores pueden escuchar tantas veces como quieran las frases, escogiendo entre las dos versiones sintéticas o indicando que el resultado le es indiferente. Las pruebas han sido realizadas por 26 miembros del Departamento de Comunicaciones y Teoría de la Señal de Enginyeria i Arquitectura La Salle (Universidad Ramón Llull).

Evaluación subjetiva de los aciertos del clasificador: La primera prueba se enmarca en el contexto del correcto funcionamiento del CT basado en RRA R desarrollado, es decir, se toman en consideración los aciertos de clasificación. Para ello, se comparan los resultados obtenidos al sintetizar el texto de entrada con el corpus neutro (utilizado como referencia de la calidad que se conseguiría a partir de un corpus genérico) y con el corpus indicado por el CT, trabajando en ambos casos como prosodia objetivo, la extraída de la frase a clasificar. Por lo tanto, debido a que cuando el sistema de CTH-MD asigna el texto de entrada al dominio correcto, éste funciona esencialmente como un CTH-DR construido para ese dominio, esta prueba es equivalente a comparar la CTH-DR con la CTH generada a partir de un corpus genérico utilizando una misma prosodia objetivo.

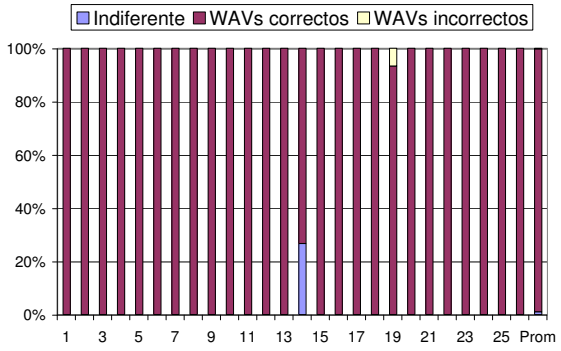
En la interfaz de test se indica al evaluador el estilo de locución deseado para cada frase, y se le pide que escoja, de las dos síntesis candidatas, la que mejor le transmita ese estilo (alegre o sensual, en este caso), obviando los pequeños errores de síntesis que puedan contener[32]. Para ello se presentan al evaluador 12 frases etiquetadas correctamente como alegres (ver tabla 3.15(a)) y 15 como sensuales (ver tabla 3.15(b)), que han sido seleccionadas de un conjunto más amplio de frases obtenido mediante un análisis del corpus utilizando un algoritmo *greedy* para asegurar una distribución equilibrada de las unidades. De este conjunto inicial, se escogió el subconjunto que presentaba mejor calidad sintética global (tanto a partir del dominio indicado por el CT como la obtenida del dominio neutro,

[32] Al utilizar una función de coste simple, aparecen problemas en puntos de concatenación o en unidades que están prosódicamente alejadas de las deseadas, por lo que la síntesis utilizada —basada en TD-PSOLA— tiene serios problemas para alcanzar determinados objetivos. En un futuro cercano, se pretende incorporar la investigación realizada en el ámbito del ajuste de pesos (capítulo 2) para incorporarla a la estrategia de CTH-MD y, así, minimizar estos problemas.

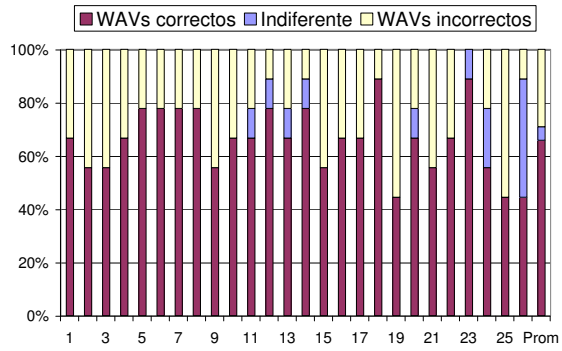
evitando sesgar los resultados en uno u otro sentido). Los resultados de este experimento se presentan en la figura 3.35(a) para el dominio alegre (ALE) y en la figura 3.35(b) para el dominio sensual (SEN).



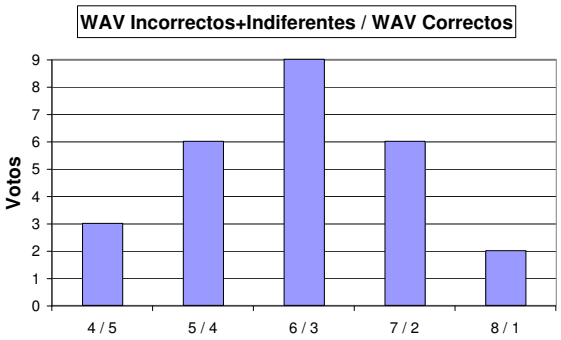
(a) Prosodia ALE + corpus NEU *vs.* prosodia ALE + corpus ALE.



(b) Prosodia SEN + corpus NEU *vs.* prosodia SEN + corpus SEN.



(c) Errores de clasificación.



(d) Histograma de la distribución de la preferencia de los usuarios sobre la síntesis debida a los errores del CT.

Figura 3.35: Preferencias de los evaluadores sobre las parejas sintéticas. *Wav correctos* indica el porcentaje de elecciones del evaluador a favor del resultado obtenido según lo indicado por el clasificador, *Wav incorrectos* indica porcentaje de elecciones del evaluador contrarias a lo indicado por el clasificador (corpus neutro en los aciertos del CT o corpus indicado por el CT cuando se equivoca), e *Indiferente* indica el porcentaje de casos donde el evaluador fue incapaz de escoger (síntesis igual de buenas, igual de inadecuadas, etc.).

Del análisis de los resultados obtenidos, se puede concluir que, tanto para el estilo alegre como el sensual, existe una clara tendencia de los evaluadores a escoger la síntesis obtenida de utilizar el corpus correspondiente al estilo indicado por el CT respecto a la versión sintética obtenida a partir del corpus neutro. Concretamente, para alegría, el promedio

de los resultados indica una preferencia media del 74 % de la síntesis multidominio a lo largo de las comparativas, que sube hasta el 85 % al incluir las comparativas indiferentes, con la particularidad que ningún evaluador presenta una predilección superior al 40 % para la síntesis obtenida a partir del dominio neutro. De igual forma, los resultados obtenidos para la prueba sobre el dominio sensual muestran una preferencia abrumadora hacia el uso del dominio indicado por el CT ante la síntesis de ese estilo a partir del corpus neutro. Sólo existe un error en todas las pruebas junto a un único usuario que presenta cuatro indiferencias. De estos resultados se deduce la dificultad de simular determinados estilos de locución (prosodia) a partir de un corpus neutro mediante modificación de la señal en el contexto de la síntesis concatenativa, por lo que resulta necesario disponer, además de un buen modelo prosódico, de los corpus correspondientes de forma explícita, como en (Iida et al., 2000; Campbell, 2005). Fundamentalmente, esto es debido a que estos estilos presentan ciertas particularidades vocales que, por el momento, no ha sido posible resolver a partir del modelado prosódico y el posterior procesado digital de la señal (Turk et al., 2005) (ver apartado 3.5 para una discusión más detallada).

Tabla 3.16: Listado de frases utilizadas para evaluar el impacto de los errores de clasificación en la estrategia de CTH-MD. La columna izquierda indica la etiqueta original respecto a la indicada por el clasificador de textos.

Confusión	**Frase**
ALE→SEN	Para realizar esta obra, he dedicado 30 años de investigación y amor por la naturaleza.
SEN→NEU	Resístete a envejecer.
SEN→ALE	Superbrillo.
ALE→NEU	Autoras con sentido y sensibilidad.
NEU→SEN	Cógelo y habla.
ALE→NEU	El juego que levanta pasiones.
NEU→SEN	Soluciones a medida.
ALE→SEN	Este libro le hará ser de otro modo a la mujer, al hombre y a Dios.
SEN→NEU	Pero no se pueden sustraer al perfume.

Evaluación subjetiva de los errores del clasificador: La segunda parte del experimento pretende analizar subjetivamente el impacto en la síntesis de los errores de clasificación, es decir, estudiar qué sucede cuando el clasificador ofrece una etiqueta distinta a la indicada por el experto. Por lo tanto, este experimento es equivalente a comparar el *peor* caso en el funcionamiento del CTH-MD respecto al CTH-DR correspondiente al dominio del texto. En este caso, el formato de prueba es el mismo que el que se ha descrito para la primera parte del experimento, pero incorporando ciertas modificaciones. Por un lado, se compara el resultado de la síntesis sobre el dominio etiquetado de antemano respecto al resultado que se obtiene sobre el dominio indicado por el clasificador (ver tabla 3.16). Para

ello, además de utilizar la prosodia propia de la frase escogida para la prueba (*copy prosody*), resulta necesario conocer la prosodia asignada al texto dentro del dominio indicado por el CT (distinto al que pertenece la frase de test). Para ello, se utiliza el modelo prosódico correspondiente entrenado sobre cada uno de los dominios (ver (Miralles, 2005) para más detalles). Por otro lado, el evaluador deberá escoger de la pareja de frases sintéticas aquella que tenga mayor naturalidad en términos del estilo de locución sintético según el contenido del mensaje generado, tomando en consideración su propio criterio, ya que el estilo de la síntesis dependerá del dominio utilizado.

Del análisis de los resultados obtenidos (ver figura 3.35(c)), se deduce que esta prueba es más complicada que la anterior, fundamentalmente, porque en este caso entran en juego factores de decisión más complejos (p.ej. el criterio de cada evaluador para juzgar el grado de naturalidad de las síntesis). De todos modos, existe una cierta tendencia a escoger el resultado de la síntesis teórica, es decir, el indicado por la etiqueta del texto (manual), que corresponde al % de *Wav correctos* en la figura. Sin embargo, la tendencia no es tan clara como en la primera prueba, presentando patrones de usuario bastante distintos entre sí (mayor desviación entre las elecciones de los usuarios). Si se analiza la distribución del histograma de los votos de la figura 3.35(d), se pueden observar los distintos criterios seguidos por los evaluadores comparando las preferencias entre el binomio *Wav Incorrecto + Indiferente* (elección de la síntesis sobre el dominio indicado por el CT o indiferencia) respecto a *Wav Correcto*. La media del histograma se encuentra en 6/3, pero existen patrones tan dispares como 8/1 o 4/5.

Preguntados los evaluadores, se constató la mayor dificultad de esta prueba respecto a las dos anteriores (p.ej. era necesario un mayor número de iteraciones antes de elegir), ya que a diferencia de aquellas, donde se primaba la calidad de la síntesis (respuesta más homogénea de los usuarios), en esta prueba se tenía que considerar además lo *adecuado* que resultaba uno u otro estilos de síntesis para el mensaje en cuestión. Por ejemplo, la primera frase del experimento "*He dedicado 30 años de investigación y amor por la naturaleza*", que pertenece al dominio de educación (según lo etiquetado) y, por lo tanto, está grabada con estilo alegre, para algunos de los evaluadores no era aceptable como tal. Es decir, esperaban una locución neutra (p.ej. tipo documental) o triste, entre otras, pero no sensual, que era la otra opción propuesta. Por ello, su votación era indiferente. Sin embargo, para otros, el estilo alegre les resultaba totalmente creíble.

Esta respuesta heterogénea de los evaluadores —en cierto modo ambigua y algo más indefinida que en los primeros experimentos—, correla, de algún modo, con los errores de clasificación del método de CT propuesto (basado en RRA), que son debidos, fundamentalmente, a la presencia de frases sin una clara pertenencia a ninguna de las categorías del corpus (p.ej. "*Soluciones a medida*"). De algún modo, esta situación es debida a la ambigüedad inherente entre el estilo de locución y el texto a sintetizar, donde relaciones más complejas de las consideradas en este trabajo de investigación —correspondencia entre el estilo de locución y el dominio —quedan fuera del alcance del mismo. De todos modos, todavía existe trabajo para mejorar el funcionamiento del CT en el contexto de la clasificación de textos cortos (p.ej. a nivel de frase), con el objetivo de evitar errores en frases del tipo "*Pero no se pueden sustraer al perfume*", que es clasificada como neutra, cuando parece

que debería ser etiquetada como sensual al incluir el término *perfume*.

3.5. Discusión

En este capítulo se ha definido una propuesta de sistema de CTH que pretende presentar una alternativa hacia la consecución de una síntesis genérica perfecta, trabajando sobre múltiples dominios elegibles en tiempo de conversión de texto en habla. Esta propuesta, que ha sido denominada como CTH multidominio o CTH-MD, nace con el objetivo de aumentar la flexibilidad de los sistemas de CTH, sin perder la naturalidad del mensaje sintético obtenido por el CTH-DR equivalente. Como primer paso hacia el desarrollo de sistemas de CTH-MD, se ha seguido la filosofía de la CTH basada en corpus o selección de unidades, trabajando sobre un corpus multidominio formado por dominios independientes (debido a las características de los estilos de locución considerados). Asimismo, en el presente capítulo del trabajo se ha descrito la aproximación diseñada para el bloque de clasificación de textos que se incorpora a la arquitectura del CTH-MD con el objetivo de determinar de forma automática el dominio más adecuado para llevar a cabo la síntesis del texto de entrada. Como se ha comentado, el algoritmo propuesto parte de los criterios clásicos para la clasificación temática y estilística (no temática) de documentos, y los adapta, mediante la parametrización del texto y las medidas de similitud estudiadas, a las necesidades planteadas por la CTH-MD. A continuación, para cerrar el presente capítulo, se presentan algunas reflexiones sobre la propuesta realizada, los módulos desarrollados y los resultados obtenidos hasta el momento.

Coexistencia de estrategias de síntesis y modelos de corpus

El concepto de la CTH multidominio, que en este trabajo ha sido abordada a partir de la síntesis basada en corpus, es extrapolable a otras estrategias de síntesis (p.ej. basada en modelos de Markov, articulatoria, etc.), ya que el objetivo último de la CTH-MD es mejorar la flexibilidad de la síntesis, tomando en consideración las estrategias de síntesis (modelos, corpus, procesado de la señal, etc.) más adecuadas —que consigan una mayor naturalidad— para cada estilo de locución (arquitectura flexible). Bajo el enfoque de la síntesis basada en corpus, esta problemática se puede abordar mediante el diseño y la grabación de los corpus correspondientes a cada uno de los dominios considerados (p.ej. (Iida et al., 2000; Campbell, 2005)). Por lo tanto, el dominio no se modela, sino que *simplemente* se reproduce a partir de la información disponible en el corpus (Hofer, Richmond y Clark, 2005), consiguiendo una calidad óptima cuando se reproduzca el estilo almacenado. Sin embargo, si se pretende obtener una señal de voz con unas características distintas a las de la señal grabada (estilo de locución, emoción o calidad vocal), la modificación de la misma producirá una reducción de la calidad sintética (Yamagishi et al., 2003; Yamagishi et al., 2005). De las dos estrategias de diseño de corpus para CTH-MD basada en corpus, la estrategia *tiering* es la más costosa en términos de aumento del número de dominios, ya que implica diseñar y crear un nuevo corpus de voz completo (asegurando la cobertura y balanceo del dominio de trabajo, así como

sucede en CTH-DR) para el nuevo dominio, a diferencia de la estrategia *blending* que sólo implica añadir el subcorpus correspondiente al dominio deseado (no necesita asegurar la cobertura del dominio) acompañando a un corpus genérico (p.ej. si se quiere disponer de un corpus con distintas intensidades en las emociones (Hofer, Richmond y Clark, 2005)). Por otro lado, las estrategias de síntesis basadas en modelos ocultos de Markov (HMM) han permitido aumentar de forma decisiva la flexibilidad y la compactación de los sistemas de CTH actuales (gracias a la parametrización de la señal de voz), a cambio de obtener calidades sintéticas menores, tipo *vocoder* (Black, 2002; Toda y Tokuda, 2005; Zen y Toda, 2005). Sin embargo, estos sistemas son capaces de generar distintas emociones o estilos de locución a partir del modelado estadístico de los datos del corpus, utilizando: datos específicos de dominio, modelados mixtos, adaptaciones a dominio a partir de pequeñas colecciones de frases del dominio objetivo o estilos intermedios, entre otros (Yamagishi et al., 2003; Yamagishi et al., 2004; Tachibana et al., 2004; Miyanaga, Masuko y Kobayashi, 2004; Yamagishi et al., 2005).

La propuesta de CTH-MD que se ha presentado en este trabajo de investigación no sólo permite escoger una u otra estrategia de síntesis, sino que también permite utilizar soluciones híbridas que, aunque no presenten una calidad de síntesis homogénea, consiguen aumentar la flexibilidad del sistema de síntesis —por ejemplo, combinando la síntesis basada en corpus más alguna estrategia de transformación de voz, con la síntesis basada en HMM. Esta cuestión queda abierta para futuras investigaciones a partir de los resultados obtenidos hasta el momento, con el objetivo de poder 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 disponibles en cada momento, en lugar de tener que crear un nuevo sistema de CTH para cada nueva aplicación.

Por otro lado, la propuesta de CTH-MD permite también la coexistencia de distintos dominios en un mismo corpus de voz. De este modo, se puede escoger la tipología de corpus más adecuada según las necesidades de la aplicación (*tiering*, *blending*, opciones mixtas, ... ver sección 3.2.2), considerando que, hasta el momento, no todos los dominios (emociones, estilos de locución, etc.) han podido ser modelados de forma precisa (Bulut, Narayanan y Syrdal, 2002; Hamza et al., 2004; Jiang et al., 2005). Por un lado, ciertos dominios deben estar explícitamente representados en el corpus para ser reproducidos fidedignamente (Black, 2002) (p.ej. determinadas emociones como la alegría, según (Iriondo et al., 2000; Iida et al., 2000; Iriondo et al., 2004) o estilos de locución con características vocales propias -calidad vocal-, según (Schröder, 2004; Campbell y Mokhtari, 2003; Turk et al., 2005; Campbell, 2005)), mientras que, por otro lado, existen otros dominios que pueden ser generados sintéticamente de forma *bastante* realista (p.ej. tristeza desde un corpus neutro mediante modificaciones prosódicas (Montero et al., 1999; Iriondo et al., 2000; Iriondo et al., 2004), o incorporando pequeños subcorpus a un corpus de propósito general de tamaño mayor, p.ej. buenas o malas noticias (Hamza et al., 2004)).

Finalmente, comentar que, en la implementación del CTH-MD basado en corpus que se ha llevado a cabo, 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 (Hamza et al., 2004; Hofer, Richmond y Clark, 2005)—, flexibilizando la preselección de las unidades que implica la selección de dominio según la estrategia *tiering* considerada hasta el momento. En un futuro, se pretende estudiar el funcionamiento de la CTH-MD basada en corpus utilizando esta filosofía.

Valor añadido de la propuesta de CTH-MD

La propuesta de CTH-MD permite abordar el problema de la optimización de los sistemas de CTH basados en selección de unidades desde un punto de vista diferente al convencional. Según la implementación de la CTH-MD llevada a cabo en este trabajo, al mismo tiempo que se escoge el dominio más adecuado para realizar la síntesis, se consigue reducir el coste computacional del proceso de selección. Esto es debido a que, una vez escogida una de las secciones de la base de datos (cualquiera de los dominios del corpus, p.ej. $Dominio_{nm}$ en la figura 3.3), el conjunto de unidades sobre el que se realiza la búsqueda se ve reducido de forma considerable respecto al total de unidades presentes en el corpus multidominio (ver tabla 3.12 o la figura D.7 como ejemplos). Así pues, se introduce un nuevo enfoque para la reducción del número de unidades consideradas en el proceso de búsqueda, en este caso, basado en la similitud entre el texto de entrada y los dominios del corpus, gracias a la arquitectura multidominio propuesta.

Aunque las pruebas se han llevado a cabo, por el momento, sobre el nivel más bajo de la estructura jerárquica de la propuesta de arquitectura multidominio descrita, el CTH-MD está preparado para trabajar con distintos niveles de agrupación de sus datos. Bajo esta estructura, si el CT es incapaz de asignar con cierta fiabilidad el texto de entrada a uno de los subdominios considerados, éste puede utilizar el nivel jerárquico inmediatamente superior (realizando una nueva búsqueda, o utilizando el macrodominio que incluya los dominios de mayor pertenencia). En el peor de los casos, cuando el clasificador sea incapaz de seleccionar un dominio o macrodominio concreto dentro de la estructura jerárquica del corpus (por ejemplo la de la figura 3.18), el CTH puede pasar a utilizar, o bien todo el corpus multidominio, o bien el subcorpus neutro para generar la señal sintética, ralentizándose la síntesis respecto a haberlo hecho directamente debido al coste computacional adicional del proceso de clasificación. No obstante, por un lado, esta situación no será la más habitual, y por otro, el coste computacional de los métodos de CT propuestos es bajo, como se ha mostrado en los experimentos que se han realizado. Esta situación descrita para CTH-MD, también ha sido justificada en el ámbito de los sistemas de reconocimiento que utilizan múltiples modelos de lenguaje en paralelo, mejorando la precisión y la eficiencia del sistema de reconocimiento (Lane et al., 2005) y reduciendo el tamaño del espacio de búsqueda al conocer el dominio actual del discurso (Rüggenmann y Gurevych, 2004).

Igualmente, la propuesta de CTH-MD permite incluir todas las estrategias de reducción de coste computacional del proceso de búsqueda que se han descrito en la sección 2.1.2 (preselección, *clustering*, *pruning*...), añadiendo un nivel más de agrupación de los datos, en este caso, en función del dominio del texto a sintetizar.

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 PLN) (ver p.ej. (Dutoit, 1997)), 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 (p.ej. (Sagisaka, Yamashita y Kokenawa, 2004; Sagisaka, Yamashita y Kokenawa, 2005)) o la emoción subyacente en el mensaje (p.ej. (Tao y Tan, 2004; Sugimoto et al., 2004; Ovesdotter, Roth y Sproat, 2005; Hofer, Richmond y Clark, 2005), con el objetivo de mejorar la naturalidad de la síntesis. En el trabajo de (Sagisaka, Yamashita y Kokenawa, 2004; Sagisaka, Yamashita y Kokenawa, 2005) se demuestra la correlación entre las variaciones prosódicas (concretamente, de la curva de F0) 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. Siguiendo una filosofía similar, la detección de emociones a partir del texto se basa en considerar que las frases están formadas por algunas palabras con funcionalidad emotiva acompañadas de otras que no la tienen (Hofer, Richmond y Clark, 2005). En (Sugimoto et al., 2004), este análisis se centra sólo en considerar el número y tipo de adjetivos de cada frase, dejando para investigaciones futuras el análisis de las relaciones entre nombres, verbos, adjetivos y adverbios del texto para estimar mejor la emoción de cada texto. En (Hofer, Richmond y Clark, 2005) se diseña un corpus con tres emociones (neutra, alegre y enfado) a partir de 400 frases extraídas de textos periodísticos (con buena cobertura de los difonemas del idioma de trabajo), 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, condicionando el módulo de selección para escoger unidades de la misma emoción dentro de cada palabra. Además, el número total de palabras emotivas de la frase a sintetizar puede ser controlado mediante unos pesos que permiten dar mayor o menor nivel de emotividad al mensaje sintético. 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 (Tao y Tan, 2004) se presenta un trabajo similar, en este caso para el Chino, mediante la creación del diccionario a partir de palabras clave (adjetivos, nombres y verbos), modificadores (p.ej. adverbios o expresiones típicas) y palabras metafóricas (indicando sentimientos positivos o negativos). A partir de estos parámetros, conjuntamente con un análisis y desambiguación morfosintáctica del texto (en inglés, *Part-of-Speech* (POS) *tagging*), el sistema es capaz de

estimar la emoción subyacente en el texto.

En el ámbito de los CTH dependientes de aplicación, destaca el trabajo de (Oversdotter y Sproat, 2005; Ovesdotter, Roth y Sproat, 2005), orientado a la lectura de cuentos infantiles. En este trabajo 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, POS *tagging*, puntuación de la frase, distribución de palabras, etc.), incorporando al análisis del texto conocimiento externo al mismo —en este caso, conociendo la temática de la historia y utilizando la información contenida en WordNet (Fellbaum, 1998). Siguiendo un enfoque similar, se encuentran el trabajo de (Chuang y Wu, 2002), que incorpora una red semántica al análisis emotivo del texto, o el trabajo de (Liu, Lieberman y Selker, 2003) que incorpora una red de conocimiento genérico o sentido común (*common sense*) al problema.

En este mismo contexto, cabe añadir que en el ámbito de la investigación sobre clasificación de textos también existe una incipiente línea de trabajo con el objetivo de determinar el grado de afecto del texto o la actitud que tenía el autor al redactarlo (p.ej. determinar si la opinión del usuario de un producto es positiva o negativa) (ver la conferencia *Exploring Attitude and Affect in Text: Theories and Applications*[33] de la *American Association for Artificial Intelligence* (AAAI) *http://www.aaai.org*).

En otro plano, pero persiguiendo el mismo objetivo de mejorar la naturalidad de la síntesis, se encuentran trabajos como los de (Sundaraman y Narayanan, 2002; Sundaram y Narayanan, 2003) en los que se estudia cómo generar voz espontánea (incluyendo suspiros, pausados, onomatopeias—'*mmm*', '*uh*', '*aha*', etc. en la síntesis) a partir del análisis y el modelado de conversaciones espontáneas —aunque en este caso, como comentan los autores, la mejora de la naturalidad a través del aumento de la espontaneidad de la síntesis conlleva perder cierta inteligibilidad del mensaje sintético.

Todos estos trabajos demuestran la incipiente investigación que busca mejorar la naturalidad de la síntesis a partir de la extracción de mayor información del texto de entrada, a través de detectar su estilo, género literario, emotividad, intención, etc., cuestión que, simplemente ha empezado a ser tratada en el ámbito de la conversión de texto en habla y que abre la puerta a futuros trabajos de investigación.

Estrategias de clasificación de textos

La idea de indicar al CTH-MD el dominio del texto a sintetizar de forma automática es abordable, de entrada, utilizando cualquier tipo de estrategia de clasificación de textos. En este contexto es necesario argumentar la necesidad de adaptar el CT al entorno de la CTH-MD, debido a que, por un lado, se dispone, en general, de un corpus con un número de documentos y frases por documento muy reducido y, por otro, la clasificación de textos se realiza, habitualmente, sobre textos de tamaño extremadamente reducido (p.ej. 1 frase). Por ello, se pasa a discutir algunos de los elementos relacionados con la definición y el funcionamiento de la propuesta, así como la justificación del método utilizado como referencia para validar sus resultados.

[33] *http://www.clairvoyancecorp.com/Research/Workshops/AAAI-EAAT-2004/home.html*

Método de referencia: ¿SVM, ICA, NN o *bigramas*?

Como se ha descrito en el apartado 3.4.2 de los experimentos, el método de clasificación de textos basado en *Nearest Neighbour* (NN) ha sido el escogido como referencia para contrastar los resultados obtenidos por las propuestas basadas en RRA. Seguidamente se presenta el estudio que ha permitido escoger este método como algoritmo de referencia, comparando su comportamiento respecto a otros métodos aplicables a la CT en el contexto de la CTH-MD descrito en este trabajo de investigación.

Como se ha comentado, la aplicación clásica de la CT es la clasificación temática de documentos. Para ello, en la investigación en este ámbito se suele trabajar con importantes colecciones de documentos (y de frases por documento), como, por ejemplo, las colecciones *Reuters-21758* y *OHSUMED* (Sebastiani, 2002). En este contexto es donde el algoritmo basado en *Support Vector Machines* (SVM) ha demostrado un óptimo funcionamiento para el problema de la clasificación temática de textos (Joachims, 1998; Sebastiani, 2002). Por ello, las primeras pruebas que se realizaron en este trabajo para diseñar un CT en el contexto de la CTH-MD se basaron en el algoritmo SVM^{light} de Joachims (2000). Sin embargo, como se indica en (Sassano, 2003), a medida que el volumen de datos de entrenamiento se reduce respecto a la dimensionalidad del MEV empleado para representarlos, el método SVM pierde eficiencia, llegando a dejar de funcionar correctamente en el caso de trabajar con un MEV de dimensión mayor que el número de ejemplos. Este comportamiento de SVM se debe a la relación existente entre el número de datos de entrenamiento y el tamaño del MEV a parametrizar, que para *kernels* lineales, necesita disponer de un orden de $O(N)$ ejemplos para modelar correctamente un espacio $\mathbb{R}^N$ (Shawe-Taylor y Cristianini, 2004) —mientras que serán necesarios aún más ejemplos para modelar *kernels* no lineales.

Debido a la estrategia de clasificación utilizada, donde no se trata con un CT únicamente temático y se entrena al CT sólo con los textos del corpus de voz (ver el apartado "*Volumen de datos de entrenamiento*"descrito a continuación), la relación entre la dimensión del espacio vectorial generado por los datos de entrenamiento y el número de ejemplos se acentúa al no eliminar la variabilidad de los términos (extracción del lema o *stemming*) ni las palabras vacías del texto (lista de parada o *stop list*) —es más, esta relación empeora al considerar las coocurrencias de las palabras en el modelado de los textos, es decir, aumenta el número de parámetros a modelar para el mismo volumen de datos de entrenamiento. En este caso, el corpus C_{Pub} dispone de 2590 frases (ejemplos) (ver tabla 3.8) sobre un espacio de 15134 dimensiones (contando únicamente las palabras), pasando a ser de 1367 frases con 10452 palabras para la versión reducida del corpus (ver figura 3.19). Por todo ello, en este caso, SVM no presenta resultados satisfactorios para el problema de clasificación planteado (es incapaz de clasificar correctamente los textos). No obstante, la propuesta de sistema de clasificación de textos basada en RRA ha sido aplicada sobre las cinco categorías más pobladas de la colección *Reuters-21758* (ver "*Experimento 1*"), validando también la viabilidad de la misma sobre un corpus de dimensiones considerables.

A la vista de los pobres resultados ofrecidos por SVM, se estudian otras alternativas para obtener un CT de referencia, que tengan en cuenta todas las palabras de la colección de documentos (no se aplica ningún tipo de filtrado típico de los CT temáticos). Entre

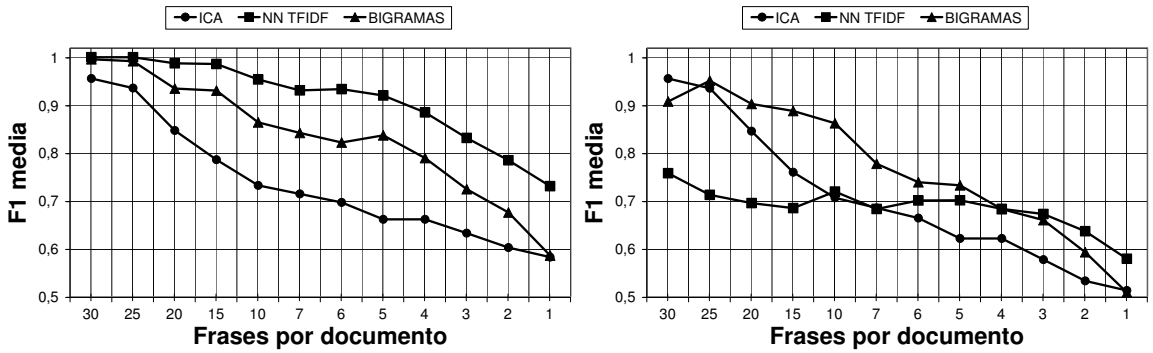
las técnicas existentes se escogen tres estrategias con enfoques de clasificación distintos. En primer lugar, se analiza el funcionamiento del sistema de clasificación de textos basado en el análisis en componentes independientes (ICA) de los textos a partir de la reducción de dimensionalidad conseguida mediante la indexación de la semántica latente (LSI) de los textos (Alías et al., 2004b; Sevillano, Alías y Socoró, 2004), que ha sido utilizado, por un lado, para la búsqueda de dominios de forma semisupervisada, y por otro, para la jerarquización temática del contenido del corpus, como se ha descrito en este capítulo. Este método, aunque es capaz de obtener una jerarquía temática coherente con los contenidos del corpus (ver apartado 3.4.1), no consigue obtener una buena eficiencia de clasificación (F_1) para pocas frases por documento (ver figura 3.36). Sin embargo, presenta buenos resultados cuando los documentos contienen un número no muy reducido de frases, con una buena robustez al ruido de los datos de entrenamiento en esta zona del barrido (ver figura 3.36(b)), pero con prestaciones muy pobres cuando se reduce el número de frases por documento, debido a su carácter totalmente temático (a menor número de frases, menor número de palabras significativas temáticamente).

Finalmente, también se aborda el problema de la CT en el ámbito de la CTH-MD mediante un método probabilístico de clasificación de textos. Para ello se diseña y desarrolla un CT basado en *n-gramas*, concretamente *bigramas* —donde el *grama* en este caso corresponde a un carácter, a diferencia de las aproximaciones basadas en la palabra[34]. Es decir, se modela cada una de las categorías (dominios) utilizando un modelo de lenguaje probabilístico obtenido a partir de la distribución de las parejas de caracteres presentes en el texto de cada categoría (Cavnar y Trenkle, 1994). Durante el proceso de entrenamiento se halla la probabilidad condicionada de aparición de cada bigrama en cada una de las categorías (la probabilidad de aparición de un carácter condicionada por la aparición de un carácter anterior). Se ha escogido trabajar con bigramas por su buen compromiso entre la información de contexto modelada y el volumen de datos disponible para modelar de forma robusta cada categoría (a mayor n, se obtiene mayor información contextual, pero por el contrario, a menor n, se logra una mayor robustez estadística para un mismo volumen de datos de entrenamiento). De su aplicación sobre el corpus C_{Pub} se deduce que este método presenta un comportamiento intermedio entre ICA y NN sobre el corpus reducido (ver figura 3.36(a)). Sin embargo, cuando se trabaja sobre el corpus completo, el método basado en bigramas presenta mejores resultados que ICA y NN en la zona alta —más de 4 frases por documento— del barrido de frases por documento (ver figura 3.36(b)), demostrando su mayor robustez ante la presencia de frases temáticamente ambiguas. Este comportamiento es lógico, ya que el método trabaja a un nivel inferior a la palabra (carácter), por lo que su eficiencia no depende tanto de las palabras que forman el texto sino de los caracteres que lo constituyen.

Como resultado de este estudio, se decidió utilizar el método de CT basado en NN (en la figura 3.36 representado por su configuración TFIDF) como método de referencia de los experimentos[35] por: *i)* ser el que presenta el mejor comportamiento sobre el corpus reducido

[34]Por el tamaño reducido del conjunto de entrenamiento, se ha descartado abordar el modelado de bigramas de palabras, al no disponer de suficiente robustez estadística.

[35]En un futuro se pretende estudiar la versión k-NN, para $k > 1$, que asigna al texto de entrada la categoría



(a) Corpus reducido.

(b) Corpus completo.

Figura 3.36: Eficiencia de clasificación media obtenida por los métodos de CT basados en ICA, NN y bigramas sobre los tres dominios del corpus publicitario. En este caso, NN utiliza el MEV definido a partir de las ponderación TFIDF de los términos (sin coocurrencias), mientras que ICA se aplica sobre un espacio de LSI con factor $k = 3$ (el número de componentes independientes es igual al número de categorías considerado).

(ver figura 3.36(a)), *ii)* conseguir los mejores resultados en la zona del barrido con menos frases por documento —la más habitual en el contexto de la CT— (ver figura 3.36(b)). Además, se da la situación que se trata de un método del ámbito del *pattern matching*, así como los basados en RRA (RRA R y RRA F).

Método propuesto: RRA + MEV

Con la propuesta inicial de representar los textos mediante redes relacionales asociativas (RRA) se buscaba disponer de una tipología de modelado del texto capaz de tratar con elementos más allá de los habituales en el ámbito de la clasificación temática o estilística de documentos (denominados parámetros estructurales). Gracias a la RRA, se han podido definir parámetros y distancias de clasificación que tienen en cuenta la secuencialidad de los términos del texto (longitud del patrón -PL y cPL-) y sus relaciones (coocurrencias -COF-), además de los parámetros típicos de la CT temática (frecuencia del término -TF- y frecuencia inversa de las palabras en los documentos -IDF-), incluyendo un nuevo parámetro temático denominado frecuencia inversa de la palabra (IWF), centrado en la importancia de la palabra dentro del texto en lugar de ponderarlo según su frecuencia de aparición y discriminabilidad a lo largo de la colección de documentos (TFIDF). Esta representación del texto ha sido explotada utilizando un modelo de espacio vectorial (MEV) extendido (para incluir las COF) sobre el que se ha llevado a cabo la tarea de la clasificación de los textos de entrada del CTH-MD, siguiendo un enfoque típico del ámbito de la CT. No obstante, cabe puntualizar que la información contenida en la RRA puede ser explotada, además

más votada de los k documentos más cercanos al documento dentro del conjunto de entrenamiento.

de utilizando un MEV, por otras estrategias de clasificación. Por ejemplo, la tipología de la RRA permite una explotación probabilística de los datos mediante su adaptación a un modelo de estados finitos o modelo de Markov visible (Rabiner y Juang, 1993) para llevar a cabo la tarea de clasificación —p.ej. modelo probabilístico de palabras[36] (López-Cózar, 2003; Peng y Schuurmans, 2003), pudiendo así, incorporar el concepto de *gramática* o *modelo de lenguaje* típicamente utilizados en el ámbito del reconocimiento de palabras (Rabiner y Juang, 1993). No obstante, será necesario analizar con detenimiento el comportamiento de este enfoque probabilístico respecto al que se ha llevado a cabo en el presente trabajo de investigación, en futuros estudios.

Por otro lado, la RRA permite abordar tanto la tarea de la clasificación temática (en los experimentos con distancia del coseno etiquetados mediante NCOF, ya que no se utilizaban las coocurrencias para la clasificación de los textos), como incorporar información de la relación entre las palabras del texto en métodos de CT clásicos (p.ej. se ha incorporado información de las coocurrencias de las palabras en el CT basado en NN, en los experimentos etiquetados mediante COF). Asimismo, esta tipología de red podría utilizarse para otras tareas de clasificación de textos, como la detección de género o la determinación de la autoría de un texto, siempre que se incorpore la información necesaria para la tarea (p.ej. se puede añadir en cada nodo de palabra la función morfosintáctica que esta desempeña en la frase). Estas cuestiones quedan abiertas para futuras investigaciones, donde se estudiará la aplicación de la RRA en distintas tareas de clasificación así como su posible explotación por otras estrategias de clasificación.

Otra de las ventajas de entrenar el método de clasificación de textos a partir de la información representada según la RRA global, es que permite incorporar nuevos textos de entrenamiento fácilmente, es decir, es un método de fácil actualización. Es decir, una vez construida la RRA de un determinado dominio D_n, ésta puede enriquecerse con nuevos textos sin tener que reentrenar todo el sistema de clasificación, ya que, para el caso RRA F, sólo es necesario recorrer de nuevo la red global para incorporar el nuevo texto, actualizando las RRA F de cada dominio. Para el caso de la estrategia RRA R, el proceso aún es más sencillo, ya que sólo es necesario incorporar el texto a la RRA del dominio en cuestión, sin tener que actualizar el resto de RRAs ya que el espacio de representación lo define el texto a clasificar. Además esta flexibilidad de actualización, también es aplicable para extraer textos que no sean muy apropiados para el dominio modelado (p.ej. textos temáticamente ambiguos), permitiendo, asimismo, incorporar conocimiento experto ponderando, por ejemplo, aquellas palabras o estructuras de palabras más significativas para cada dominio. Así pues, los sistemas de CT basados en el modelado y representación de los textos mediante una RRA son modelos versátiles y fácilmente actualizables, para cualquier tarea de clasificación de textos para la que se apliquen.

[36]En este caso, el peso de cada palabra se puede interpretar como la probabilidad de palabra (unigrama), mientras que el peso de las coocurrencias de palabras puede formularse como la probabilidad del bigrama de palabras o probabilidad de transición entre estados. Asimismo, los parámetros PL y cPL, dan información de la probabilidad de pertenencia del texto de entrada sobre cada dominio, a partir de la máxima secuencia de palabras coincidentes entre ambos (de algún modo, interpretable como n-gramas de longitud variable).

Volumen de datos de entrenamiento y clasificación

Existen distintas posibilidades a la hora de entrenar el clasificador de textos. En el presente trabajo de investigación, se ha optado por entrenar los métodos estudiados a partir de los textos que dispone el corpus de voz publicitario utilizado también para síntesis (C_{Pub} en tabla 3.8). Otra posible aproximación consistiría en entrenar el CT con grandes cantidades de documentos correspondientes a los dominios de interés, escogiendo a posteriori el subconjunto de textos utilizados para grabar el corpus de voz correspondiente. No obstante, esta segunda opción presenta dos inconvenientes fundamentales. Por un lado, no siempre resulta fácil recopilar grandes cantidades de textos adecuados a cada dominio[37], y por otro, como se deduce de las pruebas realizadas sobre el corpus completo respecto al reducido, si los datos escogidos para el entrenamiento no son lo suficientemente adecuados (p.ej. frases temáticamente ambiguas o que pueden pertenecer a distintos dominios), el ruido de los datos tiene un impacto decisivo en la eficiencia de los clasificadores desarrollados. El impacto será menor para aquellas estrategias de clasificación que modifican la representación de los datos mediante la reducción de la dimensionalidad, la extracción de parámetros significativos, etc., como se ha comprobado al utilizar un clasificador basado en ICA (ver figura 3.36(b)). De todos modos, aunque el número de datos de entrenamiento no es ni mucho menos comparable con el de las grandes colecciones de documentos con las que suelen trabajar los CT temáticos (ver (Sebastiani, 2002) para más información), los resultados obtenidos son bastante satisfactorios (p.ej. $F_1 \approx 0.78$ para 1 frase por documento) para el grado de complejidad del problema —clasificación para textos de tamaño muy reducido—, aunque se presume que todavía queda espacio para la mejora.

El volumen reducido de datos en la fase de clasificación (además del volumen de datos de entrenamiento) ha sido uno de los elementos fundamentales para el estudio y la definición de nuevas propuestas para la CT distintas a las clásicas, debido a las necesidades particulares que presenta la conversión de texto en habla para la clasificación de textos. El problema principal a superar radica en el hecho de disponer de pocos datos tanto para el entrenamiento como para la clasificación. En este contexto y como se demuestra en (Alías et al., 2004b), resulta necesario realizar un barrido del conjunto de datos de test para evaluar el comportamiento del clasificador, aumentando así la consistencia estadística de los resultados. Para ello, en los experimentos se han utilizado las estrategias *10-fold crossvalidation* o *10-fold random-subsampling*, según el conjunto de datos disponible en cada caso —los experimentos con los corpus de textos se realizaron utilizando *crossvalidation*, gracias a su tamaño, sin embargo, los experimentos sobre el corpus publicitario se realizaron con *random-subsampling*, por su menor tamaño y para disponer de un mejor muestreo del corpus teniendo en cuenta que las frases estaban ordenadas alfabéticamente.

[37]Las grandes colecciones de documentos, p.ej. Reuters-21758 (Lewis, 1994), han sido desarrolladas en inglés. En este contexto, la extensión (tamaño del corpus) y la calidad (cobertura de los documentos para cada dominio) de un corpus equivalente en castellano o catalán difícilmente podrían alcanzar, en el contexto del presente trabajo de investigación, la de los corpus en inglés que han sido etiquetados y revisados durante años.

Análisis de los resultados de las pruebas subjetivas del CTH-MD

Las pruebas realizadas para estudiar la presente propuesta de CTH-MD se han sustentado en un elemento clave: la correlación entre el dominio del texto y el estilo de locución utilizado, asignación definida por los expertos del CAP-UAB (Alías et al., 2004b). 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, cabe destacar 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 (siguiendo la filosofía *tiering*). Estos resultados se encuentran en la línea de otros trabajos, como los de (Meng et al., 2002; Chu et al., 2002; Black, 2002; Fischer, Botella y Kunzmann, 2004; Hamza et al., 2004), 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 conversión de texto en habla genéricos (CTH-PG). Así pues, en este contexto, el sistema de CTH-MD es capaz de determinar de forma automática el corpus y el modelo prosódicos más adecuados para el texto de entrada, evitando la necesidad de disponer de un texto previamente etiquetado (siempre que se cumplan las restricciones anteriormente comentadas). Por otro lado, cuando el clasificador de textos incorporado en el sistema de CTH-MD asigna el texto de entrada a un dominio distinto al que fue grabado (error de clasificación), los resultados obtenidos muestran como los evaluadores utilizan un criterio de selección mucho más vago que en el primer caso, ya que, en este caso, se está incorporando información cognitiva (relación entre el mensaje y el estilo de la síntesis) en la comparación del resultado de dos síntesis de dominio restringido (CTH-DR$_1$ *vs.* CTH-DR$_2$).

Por otro lado, 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. No obstante, a partir de la propuesta descrita en el capítulo 2, se pretende extrapolar los resultados obtenidos en el diseño de una función de coste basada en la percepción humana, para disponer de una selección de unidades *completa* en CTH-MD. Gracias a ello, será posible sustituir la selección de unidades basada en la recuperación de las mayores secuencias de unidades del corpus por una selección que tome en consideración criterios prosódicos y lingüísticos a la hora de escoger las unidades para la síntesis. Una vez incorporada la nueva función de coste para el sistema de CTH-MD, 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 según su dominio y la función de coste considerada. Estas pruebas sólo se podrán llevar a cabo bajo esta configuración, ya que trabajar con la función de coste utilizada en los experimentos aquí expuestos (sólo considera la secuencialidad de las unidades seleccionadas) puede conducir a la selección de unidades procedentes de zonas del corpus con características acústicas muy distintas (prosodia, calidades vocales, etc.), provocando el denominado efecto *patchwork* (Breen y Jackson, 1998), como se ha descrito en la introducción del presente capítulo.

Capítulo 4

Ajuste robusto de marcas de *pitch*

El proceso de generación de voz en sistemas de síntesis concatenativa parte de fragmentos o unidades de voz real —generalmente pronunciadas por un locutor profesional— almacenadas en un corpus. Este corpus, además de contener las muestras de voz, suele disponer de un conjunto de etiquetas que caracterizan la señal grabada (Black y Taylor, 1997b; Lenzo y Black, 2000; Saito y Sakamoto, 2005). En el contexto de la síntesis basada en selección de unidades, esta información es necesaria para elegir y procesar adecuadamente las unidades de voz empleadas durante la generación del habla sintética. Por lo tanto, el etiquetado *robusto* (con el mínimo número de errores posible) del corpus de voz es fundamental para conseguir una señal sintética de alta calidad (Taylor, Black y Caley, 2000-2003; Clark, Richmond y King, 2005).

Uno de los elementos principales de este etiquetado son las *marcas de pitch*[1], que indican la ubicación temporal de los *pseudo*periodos de la señal de voz. Estrictamente hablando, deberían denominarse *marcas del periodo fundamental* (T_0), ya que este es el parámetro físico que etiquetan —designan el mínimo desplazamiento temporal positivo, de entre el conjunto infinito de desplazamientos posibles, que consigue un ajuste óptimo de la señal que presenta un determinado patrón de repetición o periodicidad (para una señal totalmente periódica) (de Cheveigné y Kawahara, 2002). Sin embargo, debido a que generalmente existe una relación directa entre *pitch* y T_0 (de Cheveigné y Kawahara, 2002), se suele hablar de *marcas de pitch*. Las marcas de *pitch* son utilizadas en procesos que necesitan conocer la posición temporal de los periodos de la señal de voz (Veldhuis, 2000). En el contexto de la conversión de texto en habla destacan, entre otros: la modificación de la duración o el tono de la señal, por ejemplo mediante PSOLA (Moulines y Charpentier, 1990), donde las marcas de *pitch* juegan un papel fundamental (Laprie y Colotte, 1998; Lenzo y Black, 2000; Colotte y Laprie, 2002), el etiquetado prosódico síncrono respecto a la periodicidad de la señal de voz (Black y Font Llitjós, 2002), la definición de los puntos de concatenación en

[1]El *pitch*, o tono, es un atributo perceptivo de los sonidos y el habla no sordos. Según algunos autores, se puede definir como la frecuencia fundamental (F_0) de la onda senoidal teórica que produce la misma percepción subjetiva que el sonido estudiado, aunque existen otras teorías —ver (Gerhard, 2003) para más información.

tiempo de síntesis (Balestri et al., 1999), etc.

Aunque el etiquetado del corpus de voz puede realizarse manualmente (Lo, Lee y Ching, 1998; van Son, Binnenpoorte y Pols, 2001), lo más habitual es aplicar un sistema automático de etiquetado, que suele acompañarse de una revisión y verificación manual de sus resultados (Black y Taylor, 1997b; Stylianou, 1999; Montero et al., 2000; Teixeira et al., 2001; Colotte y Laprie, 2002; Campillo y Rodríguez Banga, 2002; Saito y Sakamoto, 2005). No obstante, cuando se trata de un corpus para síntesis basada en selección de unidades, la fase de corrección manual exhaustiva resulta extremadamente costosa, debido al gran volumen de datos con el que se debe tratar (Taylor, Black y Caley, 2000-2003). Por ello, la incorporación de sistemas más *robustos* para el etiquetado automático del corpus puede ayudar de forma decisiva a minimizar el trabajo manual (o los errores de síntesis si los resultados automáticos no son revisados).

Generalmente, el análisis automático de la periodicidad de la señal de voz pretende, por un lado, la detección o extracción de la periodicidad de la señal, mediante un algoritmo denominado en inglés *Pitch Detection Algorithm*[2] (PDA), y por otro, posicionar temporalmente las marcas de *pitch* en las posiciones donde se identifican los diferentes pseudo-periodos de la señal de voz, utilizando un algoritmo denominado en inglés como *Pitch Marking Algorithm* (PMA). Ambos análisis son tareas difíciles de resolver con precisión, debido a la propia naturaleza de la señal de voz (Hess, 1983; Barner, 1996; Chen y Kao, 2001; de Cheveigné y Kawahara, 2002; Yu y Wang, 2004; Ferencz et al., 2004). A continuación se detallan algunos de de los motivos que hacen de estos procesos una tarea compleja y provocan la mayoría de errores de los PDA y PMA actuales:

- El habla se genera a partir de un proceso físico, por lo que la voz no es una señal perfectamente periódica, sino que es *cuasi*estacionaria (Harbeck et al., 1995).
- Existen multitud de posibles estructuras o patrones en la señal de voz según el locutor, el sonido que se esté articulando, la pronunciación, el estado emocional del hablante, etc. (Barner, 1996).
- La señal de excitación (pulso glotal) no es una señal siempre regular debido a la presencia de ciertas inestabilidades en el sistema excitador (Sun, 2000) o a irregularidades del proceso de vibración glotal (de Cheveigné y Kawahara, 2002).

En el presente trabajo de investigación se ha desarrollado un algoritmo de filtrado de marcas de *pitch* con el objetivo de disponer de un proceso automático que permita obtener, a partir de una secuencia de marcas de *pitch* iniciales, unas marcas mejor ajustadas a la periodicidad de la señal, optimizando la consistencia del etiquetado del corpus en lo que se refiere a la información relacionada con el tono de la señal de voz. A continuación se describen los elementos fundamentales relacionados con la propuesta, que está basada en la

[2]También denominados *Pitch Tracking Algorithm*, en algunos trabajos (Bagshaw, Hiller y Jack, 1993; Talkin, 1995; Droppo y Acero, 1998; Quast, Schreiner y Schroeder, 2002; Kasi y Zahorian, 2002; Li, Malkin y Bilmes, 2004; Liu et al., 2005), entre otros.

programación dinámica, discutiendo también sobre las distintas señales normalmente utilizadas, los criterios utilizados para ubicar temporalmente las marcas, los métodos existentes de marcado de *pitch*, así como los resultados obtenidos al aplicar el algoritmo propuesto, sobre distintos algoritmos de referencia y utilizando dos corpus de voz distintos —uno de referencia en la literatura relacionada y otro desarrollado para síntesis del habla basada en selección de unidades. Asimismo, se introduce una nueva medida de evaluación para los algoritmos de marcado de *pitch* que permite valorar su fiabilidad de marcado independientemente del criterio local de ubicación de marcas escogido.

4.1. Señal de entrada utilizada

Tanto los PDA como PMA pueden trabajar con distintos datos de entrada. Se puede trabajar directamente sobre la señal de voz, o bien, partir de la señal electroglotal (EGG), e incluso utilizar la señal obtenida de la aplicación de sensores electromagnéticos —aunque estos últimos pueden ser dañinos para la salud y requieren de equipos muy sofisticados (Ferencz et al., 2004). Debido a que la señal de voz presenta una gran variabilidad (por los motivos que se acaban de enumerar), según sean la señal y el método de análisis escogidos, los resultados obtenidos pueden diferir entre sí considerablemente (Ferencz et al., 2004).

La señal EGG se almacena simultáneamente con la señal de voz haciendo uso de un aparato que se denomina *electroglotógrafo* o *laringógrafo*. Este aparato mide la resistencia eléctrica entre dos electrodos colocados uno a cada lado de la garganta, detectando las vibraciones de los pliegues vocales a partir de los cambios de impedancia medidos mediante los electrodos (Barner, 1996; Sakamoto y Saito, 2000; de Cheveigné y Kawahara, 2001). Esta señal, pues, contiene información de la respuesta de la glotis dentro del proceso de generación de la señal de voz en forma de tren de pulsos glotales. A partir de esta señal, se puede estimar la frecuencia fundamental de la señal de forma más *sencilla* que utilizando directamente la señal de voz[3] (de Cheveigné y Kawahara, 2001; Bánhalmi et al., 2005). El motivo fundamental que dificulta la estimación de la periodicidad directamente sobre la señal de voz es que ésta es el resultado de filtrar la excitación mediante las resonancias del tracto vocal. Esta modificación puede provocar que la señal resultante pierda el carácter claramente periódico que suele presentar la señal excitadora (de Cheveigné y Kawahara, 2002). En cambio, la señal EGG no contiene el efecto de filtrado del tracto vocal, presentando máximos de energía en los puntos de apertura de los pliegues vocales, por lo que se simplifica de forma significativa el proceso de detección o marcado de *pitch* (Barner, 1996). De todos modos, esto no significa que la señal EGG sea siempre una señal *perfecta* para esta tarea (Plante, Meyer y Ainsworth, 1995), ya que se trata de una señal que suele presentar bajos niveles energéticos (dependiendo del individuo y de su sexo) (Bagshaw, 1994), y, consecuentemente, es muy vulnerable a la presencia de ruidos (Ferencz et al., 2004). Además, ciertas partes de la señal EGG pueden perderse *(i)* por malos contactos entre los electrodos y la piel del locutor o *(ii)* en modos de pronunciación con poca variación en la superficie de

[3]También existen aproximaciones mixtas que intentan trabajar con ambas señales: voz y EGG, ver p.ej. (Ferencz et al., 2004).

contacto del pliegue vocal (de Cheveigné y Kawahara, 2001).

Por otro lado, no siempre se dispone de la señal EGG para ser utilizada como información de referencia para el marcado de *pitch* de un corpus de voz, por el coste económico del electroglotógrafo o debido a las propias particularidades del corpus, p.ej. datos previos a la adquisición del electroglotógrafo, corpus de habla espontánea, etc. Asimismo, el uso del EGG implica aumentar la incomodidad del locutor y el tiempo de grabación, con las consecuencias negativas que esto puede provocar en lo que se refiere a la calidad de la señal de voz. Por estos motivos, en el presente trabajo de investigación se ha optado por trabajar directamente sobre la señal de voz para disponer de un método que permita su aplicación a cualquier señal de voz disponible, sin las restricciones que el hecho de incorporar la señal EGG implican.

4.2. Ubicación de las marcas de *pitch*

Comparado con la vasta investigación que trata sobre PDAs —iniciada a finales de los años 60—, existe un menor número de trabajos que traten explícitamente el problema de la ubicación de las marcas de *pitch* o marcado de *pitch* (ver (Chen y Kao, 2001)). No obstante, este problema ha sido investigado implícitamente por los métodos de estimación de la posición del instante de cierre glotal[4], en los que, si bien no se habla explícitamente de marcas de *pitch*, el objetivo perseguido es el mismo: determinar la posición temporal de los periodos en las tramas de la señal de voz sonora (Kounoudes, Naylor y Brookes, 2002).

La cuestión fundamental que diferencia los PMA de los PDA es la necesidad de decidir la ubicación temporal de la marca de *pitch* dentro del periodo de la señal sonora. Normalmente, las marcas se colocan siguiendo alguna propiedad característica de la señal de voz —criterio local— (Veldhuis, 2000). Los criterios más habituales pasan por colocar la marca de *pitch* siguiendo: los máximos o mínimos de la señal (Goncharoff y Gries, 1998; Veldhuis, 2000; Chen y Kao, 2001; Colotte y Laprie, 2002; Lin y Jang, 2004; Dikshit, Zahorian y Nagulapati, S., 2005), el primer paso por cero anterior al máximo de la señal (Balestri et al., 1999), el instante de cierre glotal —a partir de la señal de voz (Ananthapadmanabha y Yegnanarayana, 1975; Cheng y O'Shaughnessy, 1989; Moulines et al., 1990; Barner, 1996; Kounoudes, Naylor y Brookes, 2002), de su transformada *wavelet* (Kobayashi et al., 1998; Ngoc y d'Alessandro, 1999; Sakamoto y Saito, 2000) o de la señal EGG (Krishnamurthy y Childers, 1986; Ferencz et al., 2004; Dikshit, Zahorian y Nagulapati, S., 2005)—, o el centro de gravedad de la trama (Stylianou, 1998; Stylianou, 1999), entre otros. La elección del criterio puede depender del contexto de aplicación del marcado, pero, sea cual sea, las marcas de *pitch* siempre deberán ajustarse temporalmente siguiendo el periodo fundamental de la señal de voz (Veldhuis, 2000).

[4]El instante de *cierre glotal*, como su nombre indica, es el momento en que la glotis del hablante se cierra durante el proceso de generación del habla, siendo este concepto una simplificación del complejo proceso físico relacionado con el cierre glotal (Barner, 1996). Este instante es una de las referencias más utilizadas para detectar y marcar la periodicidad instantánea de la señal de voz. Sin embargo, el instante de cierre glotal no siempre corresponde con un pico claro de la señal de voz. Para voz susurrante, por ejemplo, el instante de cierre glotal corresponde a un mínimo (Ngoc y d'Alessandro, 1999).

Sin embargo, decidir la ubicación de las marcas de *pitch* en cada trama considerando únicamente el criterio local, suele conllevar la aparición de errores de marcado (Veldhuis, 2000; Colotte y Laprie, 2002). Esto sucede, principalmente, en las zonas de cambio de periodicidad (transición del tono de la señal) (Ferencz et al., 2004), en puntos de cambio de sonoridad (Harbeck et al., 1995; Dikshit, Zahorian y Nagulapati, S., 2005; Bánhalmi et al., 2005) o en unidades con sonoridad mixta (p.ej. fricativas sonoras) (Mann, 1999). Para minimizar la aparición de este tipo de errores de marcado, se suele incorporar información complementaria al criterio local, tal y como se describe a continuación.

La mayoría de los PDA y PMA[5] actuales siguen la metodología descrita en (Talkin, 1995), que consta de tres fases: preprocesado de la señal de entrada, generación de candidatos (valores de F_0 o marcas) y postprocesamiento (generalmente mediante programación dinámica) (Li, Malkin y Bilmes, 2004) —aunque también existen trabajos que se basan en estrategias totalmente distintas, p.ej. ver (Mann, 1999; Sha, Burgoyne y Saul, 2004; Sha, Burgoyne y Saul, 2004; Sha y Saul, 2005; Hagmüller y Kubin, 2005). Siguiendo esta metodología, los PMA toman en consideración: (*i*) las m mejores posiciones candidatas que cumplen el criterio local considerado dentro de la trama de análisis[6] (p.ej. $m = 2$ (Chen y Kao, 2001), $m = 3$ (Kounoudes, Naylor y Brookes, 2002; Lin y Jang, 2004) o $m = 5$ (Laprie y Colotte, 1998)), (*ii*) la distancia entre marcas de tramas consecutivas respecto a la frecuencia fundamental instantánea que les corresponde (indicada por el PDA) (Laprie y Colotte, 1998; Veldhuis, 2000; Chen y Kao, 2001; Colotte y Laprie, 2002; Dikshit, Zahorian y Nagulapati, S., 2005) y, a veces, (*iii*) se incorpora el grado de correlación entre tramas consecutivas (Talkin, 1995; Veldhuis, 2000; Colotte y Laprie, 2002; Kounoudes, Naylor y Brookes, 2002). Toda esta información se suele agrupar mediante una función de coste —*coste local* y *coste de transición*—, que permite evaluar la *bondad* de las m marcas candidatas. La secuencia *óptima* de marcas de *pitch* será aquella que consiga minimizar la función de coste considerada a lo largo de la señal de voz analizada (Talkin, 1995; Goncharoff y Gries, 1998; Laprie y Colotte, 1998; Veldhuis, 2000; Chen y Kao, 2001; Colotte y Laprie, 2002; Kounoudes, Naylor y Brookes, 2002; Lin y Jang, 2004; Dikshit, Zahorian y Nagulapati, S., 2005).

Así pues, una vez elegido el criterio local al que ajustar las marcas de *pitch*, los PMA suelen incorporar la información global mediante la función de coste mencionada, aplicando un algoritmo de programación dinámica para escoger la mejor solución global de marcado (conjunto de posiciones temporales) a lo largo de la señal de voz, minimizando así los errores de ubicación debidos a considerar sólo el criterio local como elemento de decisión.

[5] Los PMA pueden ser interpretados como un refinado de los PDA, ya que éstos dan información de la F_0 a nivel de trama y los primeros determinan temporalmente la periodicidad (T_0) de la señal de voz periodo a periodo.

[6] Si el criterio local consiste en ajustar las marcas de *pitch* al máximo de la señal dentro del período, las posiciones candidatas serán las m muestras correspondientes a los máximos locales dentro del periodo, ordenadas según su nivel de amplitud.

4.3. Análisis de las propuestas de PMAs existentes

Como se acaba de describir, la gran mayoría de las propuestas de PMAs existentes presentan dos elementos en común: *(i)* tratan de ajustar, con mayor o menor intensidad (según una determinada ponderación), la posición de las marcas de *pitch* al valor instantáneo de F_0, a nivel de trama, indicado por un PDA (o proceso equivalente); (*ii*) se define una función de coste, calculada a partir de los costes local y de transición, que requiere de un ajuste adecuado del conjunto de parámetros que toma en consideración.

Aunque los PDA suelen incluir un postprocesamiento global para reforzar la continuidad de la F_0 estimada (de Cheveigné y Kawahara, 2002) —p.ej. mediante filtrados de mediana (Rabiner y Schafer, 1978; Bagshaw, Hiller y Jack, 1993), interpolación de los valores (Ngoc y d'Alessandro, 1999), programación dinámica (Ney, 1982; Hess, 1983; Harbeck et al., 1995; Kasi y Zahorian, 2002), *pattern matching* (Secrest y Doddington, 1982), o correcciones globales probabilísticas (Ying, Jamieson y Michell, 1996), entre otros—, todavía existen problemas para determinar con exactitud la frecuencia fundamental de la señal analizada (Yu y Wang, 2004), por lo que los PDA continúan siendo fuente de investigación (de Cheveigné y Kawahara, 2002) —ver trabajos recientes como (Li, Malkin y Bilmes, 2004; Hosom, 2005; Bánhalmi et al., 2005; Achan et al., 2005), entre otros. Así pues, incorporar información del PDA para determinar la posición de las marcas de *pitch* conlleva arrastrar los errores que el PDA haya podido causar, tanto en la fase de decisión de la sonoridad[7] de la trama como en la fase encargada de determinar su frecuencia fundamental. Esto provoca que los PMA sean sensibles a los errores del PDA, sobretodo cuando éstos son importantes (Veldhuis, 2000; Colotte y Laprie, 2002). Por otro lado, el ajuste de los parámetros que incorpora la función de coste se realiza, habitualmente, de forma empírica ($\{\alpha, \beta, \gamma\}$ en (Lin y Jang, 2004), $\{\alpha, \gamma, \tau_{min}\}$ en (Veldhuis, 2000), $\{\delta, \delta', \gamma\}$ (Colotte y Laprie, 2002),$\{w_i, A_d, ZCB\}$ en (Kounoudes, Naylor y Brookes, 2002), *local/transition cost ratio* en (Dikshit, Zahorian y Nagulapati, S., 2005), entre otros). Este ajuste heurístico generalmente es un proceso bastante costoso, difícil de sistematizar, y que no garantiza siempre un funcionamiento óptimo del algoritmo (Li, Malkin y Bilmes, 2004). Por lo tanto, resulta interesante buscar una estrategia que minimice el impacto de ambos problemas.

A partir de estas consideraciones, en (Alías y Iriondo, 2001a) se presentó una primera propuesta de algoritmo de filtrado de marcas, adaptando parte del algoritmo descrito en (Goncharoff y Gries, 1998), como postprocesamiento del resultado obtenido por el algoritmo RAPT (acrónimo de *Robust Algorithm for Pitch Tracking*)(Talkin, 1995). En ese trabajo, se consiguió mejorar la robustez de las marcas de *pitch*, consiguiendo una tasa de acierto respecto a las marcas de referencia —en términos de F_0 media por fonema sonoro— mejor o igual que la obtenida con cada uno de los métodos de partida (RAPT y el descrito en (Goncharoff y Gries, 1998)) sobre un pequeño corpus de unos 5 minutos de voz (ver (Alías y Iriondo, 2001a)). Siguiendo la filosofía de ese primer trabajo, a continuación se

[7]Habitualmente, los PDA disponen de un primer bloque que determina la sonoridad de las tramas, para, seguidamente, determinar la frecuencia fundamental sólo sobre las tramas sonoras (Cheng y O'Shaughnessy, 1989; Bagshaw, Hiller y Jack, 1993; Droppo y Acero, 1998; de Cheveigné y Kawahara, 2001; Kasi y Zahorian, 2002; de Cheveigné y Kawahara, 2002; Lin y Jang, 2004; Hagmüller y Kubin, 2005; Liu et al., 2005).

formula, de forma genérica, un algoritmo que permite: *(i)* ajustar la posición de las marcas de *pitch* a partir de la F_0 estimada por un PDA, o *(ii)* postprocesar las marcas de *pitch* designadas por un PMA, con el objetivo de evitar errores locales de marcado y aumentar la robustez de las marcas de *pitch* obtenidas. Por lo tanto, este algoritmo persigue conseguir una evolución (dinámica) suave de la curva de F_0 obtenida a partir de una secuencia de marcas de *pitch* iniciales, sin forzar un seguimiento estricto de la F_0 instantánea mediante complejas funciones de coste, sino simplemente evitando variaciones bruscas en la posición de las marcas. De este modo, se pretende absorber los errores de los procesos anteriores (PDA o PMA), dando una mejor consistencia a las marcas de *pitch* obtenidas. Recientemente, Dikshit, Zahorian y Nagulapati, S. (2005) han trabajado en esta misma línea, incorporando un margen de *seguridad* alrededor de la F_0 estimada por el PDA, en lugar de considerar su valor estricto para determinar la posición de las marcas de *pitch*.

Como se describe a continuación, el ajuste del presente algoritmo es bastante sencillo y permite mejorar los resultados obtenidos por un PMA y/o ajustar consistentemente (robustamente) las marcas de *pitch* a partir de la F_0 indicada por un PDA cualquiera. Para ello, el proceso se basa en la aplicación de un algoritmo de programación dinámica para el ajuste local de las marcas de *pitch*, considerando la evolución global del período fundamental de la señal de voz, tal y como se describe a continuación.

4.4. Algoritmo de filtrado de marcas de *pitch*

Seguidamente, se describe el algoritmo desarrollado para mejorar la robustez de las marcas de *pitch* y, así, optimizar el etiquetado de los corpus de voz utilizados en el contexto de la conversión de texto en habla basada en selección de unidades (CTH-SU) en el que se enmarca el presente trabajo de investigación. Como se ha comentado, se trata de un algoritmo que fue concebido inicialmente como módulo de postprocesamiento para cualquier PMA (p.ej. (Alías y Iriondo, 2001a)), pero que posteriormente ha sido adaptado para ser utilizado también como algoritmo de posicionamiento temporal de las marcas de *pitch*, dado un PDA cualquiera (ver figura 4.1) —en la línea de las propuestas de PMA aplicables a cualquier PDA descritas en (Laprie y Colotte, 1998; Veldhuis, 2000; Colotte y Laprie, 2002; Dikshit, Zahorian y Nagulapati, S., 2005). Este algoritmo se ha denominado *Pitch Marks Filtering Algorithm* (PMFA) (Alías, Monzo y Socoró, 2006), ya que pretende mejorar la robustez del marcado de la señal de voz, a partir de las marcas que recibe como dato de entrada, mediante el *filtrado* de las marcas espurias (debidas a inserciones o exceso de marcas, u omisiones o falta de marcas). Así pues, PMFA persigue ubicar con precisión las marcas de *pitch* sobre la señal de voz a partir de una secuencia de marcas iniciales ($m^f(n)$ y $m^i(n)$, respectivamente, en la figura 4.1) —nótese que el diagrama de bloques incluye un PMA simple (sPMA) para la ubicación inicial de las marcas a partir de los valores de F_0 subministrados por el PDA. Así pues, la secuencia de marcas de *pitch* iniciales $m^i(n)$ en el contexto del algoritmo desarrollado puede corresponder a:

- Las marcas obtenidas por un PMA de entrada, $m^i(n) = \text{PMA}[x(n)]$.

- Las marcas generadas mediante un algoritmo simple de marcado de *pitch* (sPMA) a partir del valor de la frecuencia fundamental por trama ($F_0(t)$) obtenida por el PDA de entrada, $m^i(n) = \text{sPMA}[\text{PDA}[x(n)]$.

  donde $x(n)$ corresponde al vector de $1 \leq \ n \leq M$ muestras de la señal de entrada utilizada (voz, EGG, etc.) y $m^i(n)$ representa el vector de marcas de *pitch* iniciales (también de M muestras), con posiciones activas (de valor la unidad) en las muestras que corresponden a las posiciones de las marcas de *pitch* —definición también aplicable a $m^f(n)$.

  Añadir que si el PDA suministra los valores de cada trama $F_0(t)$ en [Hz], para $1 \leq t \leq T$ tramas de análisis (p.ej. ventanas de $25.6ms$ con paso de $10ms$ en el corpus de referencia *Keele database* (Plante, Meyer y Ainsworth, 1995)), estos se convierten a información de periodicidad —en muestras— mediante $T_0(t) = round\big(\frac{f_s}{F_0(t)}\big)$, siendo t el índice de trama y f_s la frecuencia de muestreo utilizada[8]. Además, en este cálculo, resulta necesario controlar los valores de $F_0(t) = 0$, como se comenta más adelante.

Por otro lado, como la mayoría de los PMA referenciados en el apartado anterior, PMFA está basado en la programación dinámica. El algoritmo desarrollado está inspirado en los trabajos presentados en (Harbeck et al., 1995) y (Goncharoff y Gries, 1998), ya que la optimización dinámica está restringida según la variación máxima de la periodicidad entre tramas consecutivas (S_{max} en adelante) —dentro del margen de frecuencias fundamentales considerado $[F_{0\,min}, F_{0\,max}]$. No obstante, a diferencia de estos trabajos que se aplican directamente sobre la señal de voz, la presente propuesta utiliza como dato de entrada la secuencia de marcas de *pitch* obtenida de un PMA o a partir de los valores de F_0 entregados por un PDA, cuestión que provocará ciertos reajustes del algoritmo de programación dinámica, como se describe a continuación. En este contexto, cabe destacar que, a diferencia de otros métodos, PMFA no utiliza una función de coste compleja para decidir la ubicación de las marcas, ni busca ajustar *exactamente* las marcas de *pitch* según los valores instantáneos de $F_0(t)$ estimados por el PDA para cada trama, por lo que se consigue minimizar el impacto de los errores del PDA en la ubicación temporal de las marcas de *pitch*. Se trata de un algoritmo de fácil ajuste (sólo es necesario determinar la variación máxima trama a trama y fijar el criterio local escogido, según el diagrama de bloques de la figura 4.1) y aplicable como postprocesamiento de cualquier algoritmo de marcado o extracción de *pitch*. Asimismo, PMFA no evalúa la sonoridad de la señal, cuestión típica en los PDA o PMA, evitando incorporar los errores de este módulo durante la fase de posicionamiento temporal de las marcas.

En cuanto al criterio local escogido para ubicar las marcas de *pitch* en este trabajo, se ha escogido situar las marcas de *pitch* siguiendo los máximos en valor absoluto de la señal de voz, dentro de las tramas periódicas, así como en (Goncharoff y Gries, 1998; Veldhuis, 2000). No obstante, PMFA permite el ajuste dentro del periodo de las marcas de *pitch* a partir de cualquiera de los criterios locales descritos anteriormente: cierre glotal, máximos o mínimos de la señal, etc. Simplemente resultaría necesario incorporar el cálculo y/o proceso

[8] La función $round(x)$ realiza el redondeo al entero más cercano al valor de x.

pertinentes durante la fase de posicionamiento de las marcas. En lo que se refiere a las tramas aperiódicas (silencios o tramos sordos), las marcas se distribuyen de forma que transicionen suavemente entre los valores de T_0 de los tramos sonoros anterior y posterior —a diferencia de otras propuestas, donde se colocan las marcas siguiendo un ritmo regular predefinido, p.ej. cada $10ms$ (Black y Taylor, 1997b; Cosi et al., 2001; Campillo, Alba y Rodríguez Banga, 2005).

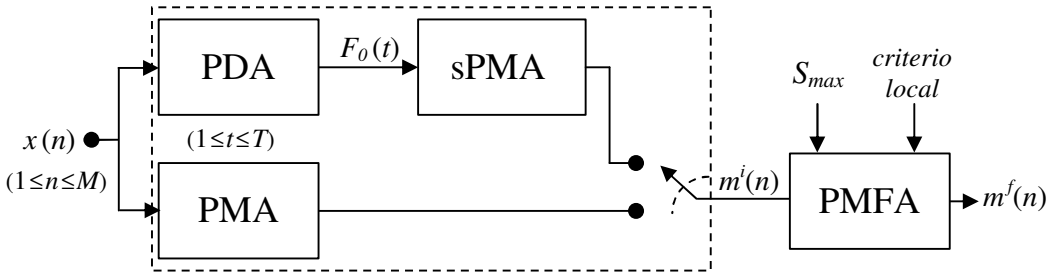


Figura 4.1: Diagrama de bloques de un sistema de marcado automático de *pitch* a partir de las M muestras de la señal $x(n)$ analizada (voz, EGG, etc.), que incorpora el filtrado automático de las marcas utilizando el algoritmo propuesto (J indica el número de tramas para el filtrado utilizadas). $F_0(t)$ representa la periodicidad a nivel de trama para las T tramas de análisis utilizadas por el PDA (y el PMA).

A partir del conjunto de marcas iniciales $m^i(n)$ obtenidas a partir de las señal $x(n)$ utilizada, el PMFA obtiene las marcas de *pitch* finales $m^f(n)$ mediante un proceso dividido en dos fases:

1. **Filtrado de errores:** se eliminan los errores de $m^i(n)$ debidos a posiciones de marcas o valores de T_0 espurios a partir de una primera aplicación del algoritmo de programación dinámica, restringido según la variación máxima de T_0 intertrama permitida (S_{max}).

2. **Ubicación temporal de las marcas:**

   - Primera estimación de la posición temporal de las marcas a partir de los valores de T_0 filtrados.
   - Refinamiento de la posición temporal de las marcas según el criterio local escogido —en las pruebas realizadas, el máximo en valor absoluto de la señal dentro del periodo— mediante una segunda pasada del algoritmo de programación dinámica restringido, en este caso, con S'_{max}.

A continuación, se pasan a describir los procesos involucrados en cada una de estas fases.

4.4.1. Obtención de las marcas de *pitch* iniciales

Como se acaba de comentar, el PMFA se puede aplicar a la salida de un PMA o de un PDA. En el primer caso, la marcas de *pitch* iniciales $m^i(n)$ serán directamente las obtenidas por el propio PMA. En cambio, en el segundo caso, será necesario generar estas marcas siguiendo la secuencia de $F_0(t)$ (o $T_0(t)$) por trama entregada por el PDA. Para ello, se aplica un algoritmo simple de marcado de *pitch* (sPMA en la figura 4.1) como paso previo a la aplicación del PMFA. En este trabajo, se utiliza el sPMA propuesto en (Goncharoff y Gries, 1998), aunque también se podría haber optado por otras alternativas, entre ellas la descrita por Lin y Jang (2004). En este caso, el sPMA genera la secuencia de marcas de *pitch* a partir de una marca inicial, fijada arbitrariamente sobre la muestra $n = 1$. Esta marca se toma como referencia para la distribución temporal del resto de marcas siguiendo los valores de periodicidad entregados por el PDA, una vez extendidos —interpolados[9]— para toda la señal de voz a partir de los valores de $T_0(t)$ de trama, obteniendo la secuencia $T_0(n)$ para todas las muestras de señal ($1 \leq n \leq M$). A continuación se describe el algoritmo base del sPMA utilizado en este trabajo —adaptado de (Goncharoff y Gries, 1998):

$$
\begin{aligned}
&m^i(n) = 0,\ \ 1 \leq n \leq M;\\
&n = 1;\\
&mientras\ (n \leq M),\\
&\qquad m^i(n) = 1;\\
&\qquad n = n + T_0(n);\\
&fin
\end{aligned}
\tag{4.1}
$$

donde M es el número de muestras de la señal de voz analizada.

Como resultado del sPMA, el vector $m^i(n)$ contendrá un conjunto de posiciones activas (no nulas) que corresponderán a la estimación inicial de la ubicación temporal —en este caso, en muestras— de las marcas de *pitch*. Este algoritmo tiene como prerrequisito que no existan zonas de $T_0(n)$ con periodicidad nula, para conseguir que el proceso recursivo de (4.1) no quede cerrado en un bucle infinito ($n+ = 0$). Para ello, resulta necesario sustituir los valores de $F_0(t) = 0$, que corresponden habitualmente a tramas sordas, por valores no nulos. En este caso, se ha escogido incorporar el valor de F_0 medio de la señal[10](sin considerar las tramas con $F_0(t) = 0$) para permitir que el proceso recursivo avance con normalidad. Una vez finalizado este proceso, las marcas $m^i(n)$ correspondientes a estas zonas, son de nuevo eliminadas antes de aplicar el PMFA, para no interferir en su comportamiento.

[9]En este trabajo, se utiliza interpolación de orden uno o *zero-hold* para extender los valores de periodicidad de cada trama para todas las muestras que la forman, generando así la curva de periodicidad para toda la señal.

[10]En un experimento preliminar se optó por mantener el valor de $F_0(t)$ de la trama sonora anterior más cercana. Esta opción fue descartada, ya que normalmente los valores extremos corresponden a zonas complicadas de etiquetar (transiciones de sonoridad), por lo que se arrastraba el error a toda la zona interpolada.

4.4.2. Filtrado de errores

A continuación, tanto si las $m^i(n)$ proceden del sPMA como de un PMA, se aplica el primer proceso del PMFA encargado de corregir los posibles errores gruesos de marcado o estimación de la F_0 —provocarán la presencia de omisiones o inserciones en $m^i(n)$. Este proceso está basado en el análisis de la secuencia de marcas iniciales mediante ventaneo junto a la aplicación, en primera instancia, de un algoritmo de programación dinámica con restricciones, descrito a continuación. Durante este proceso de filtrado de errores, se limita la periodicidad a un determinado margen de valores ($T_{0\,min} \leq T_0 \leq T_{0\,max}$) que corresponde al inverso —en muestras temporales— del margen de frecuencias fundamentales considerado en el análisis:

$$F_{0\,min} \leq F_0 \leq F_{0\,max} \tag{4.2}$$

Cualquier valor de periodicidad fuera de este rango, será eliminado (se considera valor espurio o *outlier*).

El margen de valores donde buscar la F_0 de la señal analizada es uno de los parámetros típicos de los algoritmos de PDA (Hosom, 2005). Existen muchos y diversos trabajos que, partiendo de enfoques distintos para abordar el problema de la detección de la frecuencia fundamental de la señal, acotan la F_0 a un margen de valores predefinido. Algunos de los rangos de valores —en Hz— típicamente utilizados son: [50, 500] en (Li, Malkin y Bilmes, 2004; Hosom, 2005; Chen y Kao, 2001), [50, 800] en (Sakamoto y Saito, 2000), [50, 400] en (Sha y Saul, 2005), [60,400] en (Goncharoff y Gries, 1998; Bánhalmi et al., 2005), {[80, 400], [67, 500] o [40, 800]} en las comparativas de de Cheveigné y Kawahara (2001), [40, 500] en (Liu et al., 2005) o [50, 550] en (Sun, 2002), entre otros[11]. Añadir que, obviamente, cuanto más se ajuste el margen a los valores de reales de F_0 del corpus, mejores resultados podrá ofrecer el algoritmo de detección o marcado de *pitch* aplicado (de Cheveigné y Kawahara, 2001).

A continuación se describen los procesos involucrados en la fase de filtrado de errores, que contemplan la obtención de la matriz de valores candidatos de periodicidad por trama analizada y el algoritmo de programación dinámica con restricciones (o restringido).

Obtención de la matriz de periodicidad: Para obtener una estimación fiable de la periodicidad de la señal analizada, el algoritmo propuesto primero ventanea el vector de marcas de *pitch* iniciales ($m^i(n)$, $1 \leq n \leq M$) a ritmo constante, obteniendo J tramas de análisis de la periodicidad de la señal (se utiliza una ventana rectangular, ya que no se pretende ponderar la secuencia de marcas). El número de marcas que contendrá cada ventana dependerá del tamaño y el paso de las ventanas (L y R, respectivamente, donde $L \geq R$[12]) y de la F_0 de la señal analizada (p.ej. si se utiliza una ventada de $5ms$ para

[11] De todos modos, cabe comentar que existen métodos que no necesitan restringir completamente la búsqueda, por ejemplo en (de Cheveigné y Kawahara, 2002) se designa un margen superior entorno a un cuarto de la frecuencia de muestreo (p.ej. 4000Hz para un muestreo con f_s = 16KHz), que equivale a no tener límite superior a la práctica, dado que son frecuencias muy agudas para el habla humana.

[12] Si $L > R$ se aumenta la redundancia del análisis al considerar las varias marcas para la estimación de la periodicidad de más de una ventana.

una señal de voz con un rango de F_0 de $[50, 550]$ Hz, cada trama contendrá entre 0 y 3 marcas). Una vez obtenida la secuencia de J ventanas de análisis con $J = floor\big(\frac{M-L}{R}+1\big)$,[13] se calculará la matriz de periodicidades $\mathbf{P}$ para todas la trama analizadas mediante el algoritmo (4.3).

$$
\begin{aligned}
&\mathbf{P} = (p_{ij}) = 0, \;\; 1 \leq i \leq (T_{0\,max} - T_{0\,min} + 1), \;\; 1 \leq j \leq J;\\
&j, k = 1;\\
&\;mientras\,(j \leq J),\\
&\quad mientras\,\{(1 + (j-1)\cdot R \leq I_{m^i}(k+1) \leq (L + (j-1)\cdot R)\},\\
&\qquad si(T_{0\,min} \leq (I_{m^i}(k+1) - I_{m^i}(k)) \leq T_{0\,max}),\; entonces\\
&\qquad\quad i = \big(I_{m^i}(k+1) - I_{m^i}(k)\big) - T_{0\,min} + 1;\\
&\qquad\quad p_{ij} = 1;\\
&\qquad fin\\
&\qquad k = k + 1;\\
&\quad fin\\
&\quad j = j + 1;\\
&\;fin
\end{aligned}
\tag{4.3}
$$

donde $I_{m^i}(k)$ representa la secuencia de índices de muestra donde se hallan las marcas del vector $m^i(n)$ obtenida según el proceso indicado en (4.4), en este caso con $1 \leq k \leq K^i$, siendo K^i el número total de marcas de *pitch* que $m^i(n)$ contiene.

$$
\begin{aligned}
&n, k = 1;\\
&mientras\,(n \leq M),\\
&\quad si\big(m(n) = 1\big),\; entonces\\
&\qquad\quad I_m(k) = n;\\
&\qquad\quad k = k + 1;\\
&\quad fin\\
&\quad n = n + 1;\\
&fin\\
&K = k - 1;
\end{aligned}
\tag{4.4}
$$

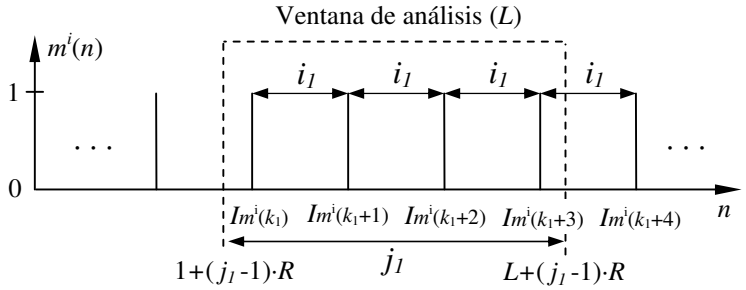
Como resultado de la aplicación del algoritmo (4.3), se obtiene una matriz binaria $\mathbf{P}$ de tamaño $(T_{0\,max} - T_{0\,min} + 1) \times J$ (representada en la expresión (4.5)), cuyas componentes no nulas ($p_{ij} = 1$) corresponderán a las filas i indicadas por la diferencia entre la posición de dos marcas consecutivas $(I_{m^i}(k+1) - I_{m^i}(k))$ y la columna j de la trama a la que pertenece el índice de la marca $I_{m^i}(k+1)$ estudiada, siempre que esta diferencia esté dentro del margen

[13] La función $floor(x)$ realiza el truncamiento al entero inferior más cercano al valor de x.

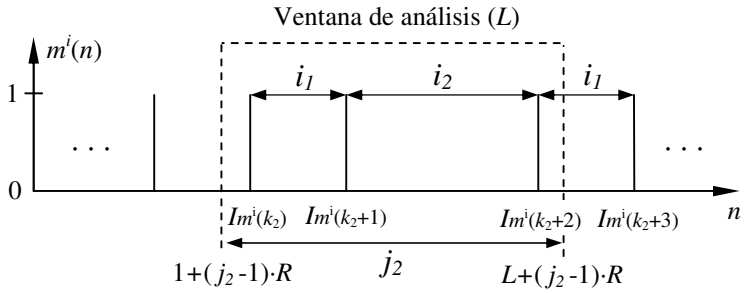
de periodicidad considerado (ver ejemplos de la figura 4.2). Así pues, la periodicidad de la fila i será $T_{0\,min} + i - 1$.

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1J} \\ p_{21} & p_{22} & \cdots & p_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ p_{F1} & p_{F2} & \cdots & p_{FJ} \end{pmatrix} \tag{4.5}$$

donde el número de filas $F = (T_{0\,max} - T_{0\,min} + 1)$, para simplificar la representación matricial.



(a) Análisis de la trama j_1 totalmente periódica (p_{ij_1} en la matriz (4.6)).



(b) Análisis de la trama j_2 con distintos valores de peridicidad (p_{ij_2} en la matriz (4.6)).

Figura 4.2: Ejemplo del análisis de la secuencia de marcas de *pitch* $m^i(n)$, cuyos índices de marca recoge el vector $I_{m^i}(k)$, para dos ventanas de análisis distintas.

$$\mathbf{P} = \left(\begin{array}{c|ccccccc} p_{ij} & 1 & \cdots & j_1 & \cdots & j_2 & \cdots & J \\ \hline 1 & p_{11} & \cdots & 0 & \cdots & 0 & \cdots & p_{1J} \\ \vdots & \vdots & \ddots & 0 & \ddots & 0 & \ddots & \vdots \\ i_1 & \vdots & \ddots & \mathbf{1} & \ddots & \mathbf{1} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 & \ddots & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 & \ddots & 0 & \ddots & \vdots \\ i_2 & \vdots & \ddots & 0 & \ddots & \mathbf{1} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 & \ddots & 0 & \ddots & \vdots \\ F & p_{F1} & \cdots & 0 & \cdots & 0 & \cdots & p_{FJ} \end{array}\right) \tag{4.6}$$

Gracias a este proceso de análisis, cada trama j dispondrá de un número distinto de estimaciones (candidatos) de periodicidad: desde ninguna hasta más de una, según la periodicidad de la señal (los valores de F_0 o $m^i(n)$ entregados por el PDA y el PMA, respectivamente) y la configuración de análisis utilizada, como se ha indicado anteriormente (ver ejemplos de la figura 4.2). De este modo, el algoritmo de programación dinámica que se aplica a continuación podrá disponer de distintos candidatos a lo largo de la matriz de periodicidad $\mathbf{P}$, pudiendo filtrar así los valores de periodicidad erróneos debidos al mal funcionamiento del PMA o del PDA. En este contexto, si la trama analizada es *completamente* periódica, su columna correspondiente de la matriz $\mathbf{P}$ sólo podrá contener una fila no nula, a diferencia de las tramas con una periodicidad menos clara, que contendrán distintas filas candidatas (debidas a inserciones, p.ej. en zonas de transición de periodicidad), o serán totalmente nulas, debidas a omisiones o zonas sordas (ver ejemplo de la expresión (4.6) generada a partir de las dos tramas de la figura 4.2). Por lo tanto, las primeras (zonas bien marcadas) serán útiles para *desambiguar* las segundas (sobremarcadas o inframarcadas), mediante la información contextual que este análisis incorpora y sobre la que se aplica el algoritmo de programación dinámica restringido diseñado. Además, como se ha comentado anteriormente, los valores de periodicidad que quedan fuera del rango de frecuencias estudiado (denominados como valores espurios o *outliers* en inglés) serán eliminados gracias a la restricción del margen de periodicidad considerado.

Por otro lado, una de las características fundamentales de la propuesta se basa en trabajar con valores binarios, cuestión que permite independizar el funcionamiento del PMFA del criterio local utilizado por el PMA o el PDA+sPMA a la hora de reubicar las marcas de *pitch* sobre la señal de voz, ya que en la primera fase de análisis (filtrado de errores) simplemente se contempla la diferencia relativa entre la posición de las marcas ($I_{m^i}(k+1)-I_{m^i}(k)$) para poder filtrar las marcas espurias por contexto. De este modo, se consigue que el PMFA se pueda aplicar de forma genérica sobre cualquier PMA o PDA.

Algoritmo de programación dinámica restringido: Una vez obtenida la matriz de periodicidad $\mathbf{P}$ con los valores candidatos de periodicidad por trama, se procede a aplicar un

algoritmo de programación dinámica con restricciones —similar al descrito en (Harbeck et al., 1995) y (Goncharoff y Gries, 1998)— con el objetivo de determinar la mejor distribución de valores de $T_0(j)$ a lo largo de las tramas de análisis ($1 \leq j \leq J$) sobre la matriz de periodicidad $\mathbf{P} = (p_{ij})$. El algoritmo se divide en las dos fases típicas de cualquier algoritmo de programación dinámica (p.ej. Viterbi (1967)). Durante la primera fase (proceso hacia delante o *forward*, en inglés) se construye una estructura *trellis* con todos los caminos de periodicidad de trama posibles (ver ejemplo de las figuras 4.3 y 4.4), pero con la restricción de una máxima variación de pendiente entre tramas consecutivas según el parámetro S_{max} (ver ecuación (4.7)). De este modo, se limita la variación máxima de $T_0(j) - T_0(j-1)$ —expresada en muestras temporales — entre tramas consecutivas (Goncharoff y Gries, 1998; Hosom, 2005), evitando así la presencia de variaciones *bruscas* intertrama. Esta restricción persigue eliminar los valores erróneos de periodicidad presentes en $\mathbf{P}$ a provocados por el PMA o el PDA de entrada, pero sin excluir la posibilidad que la señal de voz contenga fluctuaciones *rápidas* de periodicidad (controladas según los valores de S_{max} y el tamaño de ventana considerados).

$$\Big|c(j) - c(j-1)\Big| \leq S_{max},\ 2 \leq j \leq J \tag{4.7}$$

donde $c(j)$, $1 \leq j \leq J$, representa los caminos (o secuencias de periodicidades) contemplados por el algoritmo de programación dinámica (en el anexo B.1 se discute el número de transiciones entre tramas y el número total de caminos contemplados).

Durante la construcción hacia adelante de la estructura *trellis*, se calculan las métricas acumuladas (de izquierda a derecha) para cada una de las casillas de la matriz $\mathbf{P}$ según el número de marcas que han sido asignadas a cada trama —la métrica computa el número de casillas activas acumuladas hasta llegar a la casilla actual. Este proceso se implementa, en este caso, replicando en cada posición de la matriz $2S_{max} + 1$ casillas según la procedencia de los caminos que llegan a ella (ver figuras 4.3 y 4.4).

Una vez finalizado el proceso hacia delante, se aplica el proceso hacia atrás (*backward*, en inglés), mediante el que se escoge la secuencia de casillas de la estructura *trellis* que conforman el camino óptimo $c^s(j)^*$ (donde s corresponde al valor de S_{max} utilizado, ya que el camino resultante dependerá del valor de la restricción de pendiente considerado) mediante un algoritmo de *backtracking* (Rabiner y Juang, 1993). El objetivo de este proceso es encontrar la secuencia de valores de periodicidad óptima por trama a lo largo de la matriz $\mathbf{P}$ de valores candidatos. Para ello se parte de la casilla o las C casillas p_{iJ} que presenten una *métrica* acumulada máxima (ecuación (4.8)), es decir, que siguen mejor la periodicidad de las tramas indicada por el PMA o PDA de entrada (pasan por el mayor número de casillas activas) según la restricción de pendiente máxima considerada (ver ejemplos en las figuras 4.3 y 4.4). Gracias a este proceso, las tramas *claramente* periódicas (con una o pocas estimaciones candidatas de periodicidad local) guiarán el camino óptimo, constituyendo el medio para desambiguar las tramas de periodicidad poco clara (p.ej. transiciones de sonoridad, unidades fricativas sonoras, etc.) con un mayor número de valores de periodicidad

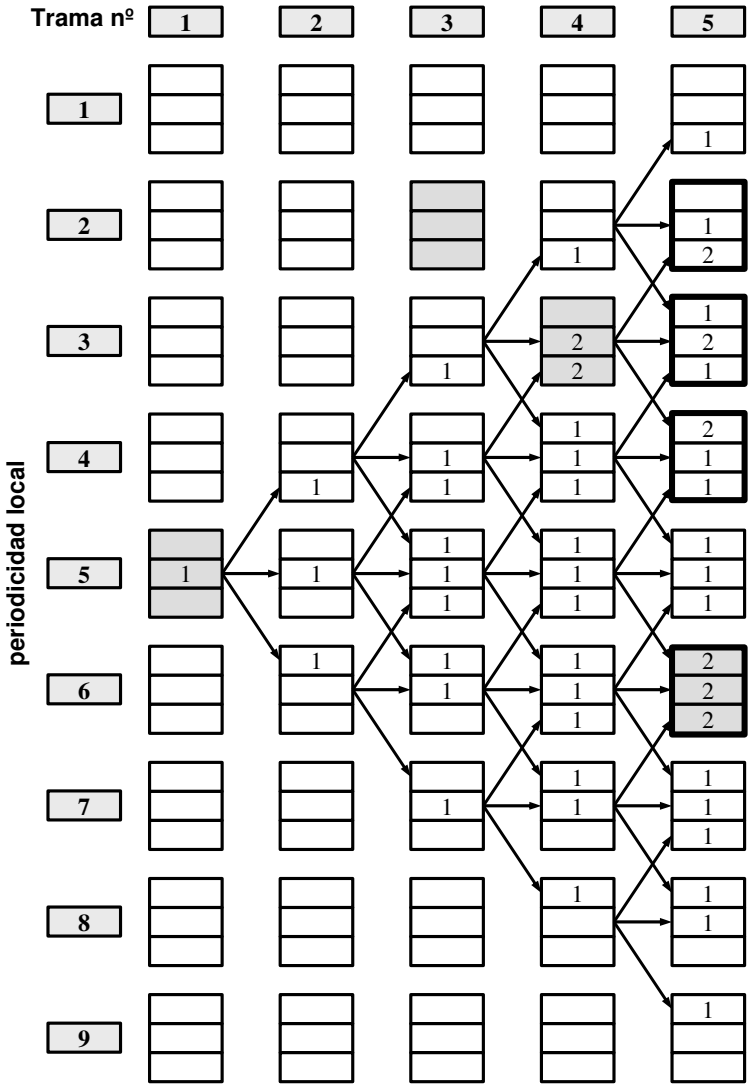


Figura 4.3: Fragmento del resultado del proceso *forward* del algoritmo de programación dinámica (partiendo sólo de la casilla p_{51} para simplificar la representación), restringido por una $S_{max} = 1$, sobre una matriz binaria $\mathbf{P}$ de 9×5 (las casillas no nulas están sombreadas).

candidatos.

$$c_n^s(j) = \underset{c^s(j)}{^{argmax}} \left(\sum_{j=1}^{J} p_{c^s(j)\, j} \right), \; 1 \leq n \leq N \tag{4.8}$$

donde $c_n^s(j)$ debe cumplir con la restricción de pendiente máximo indicada en la ecuación (4.7) y N indica el número total de caminos que consiguen maximizar la métrica considerada.

No obstante, hay que tener en cuenta que, al trabajar con una métrica entera (número

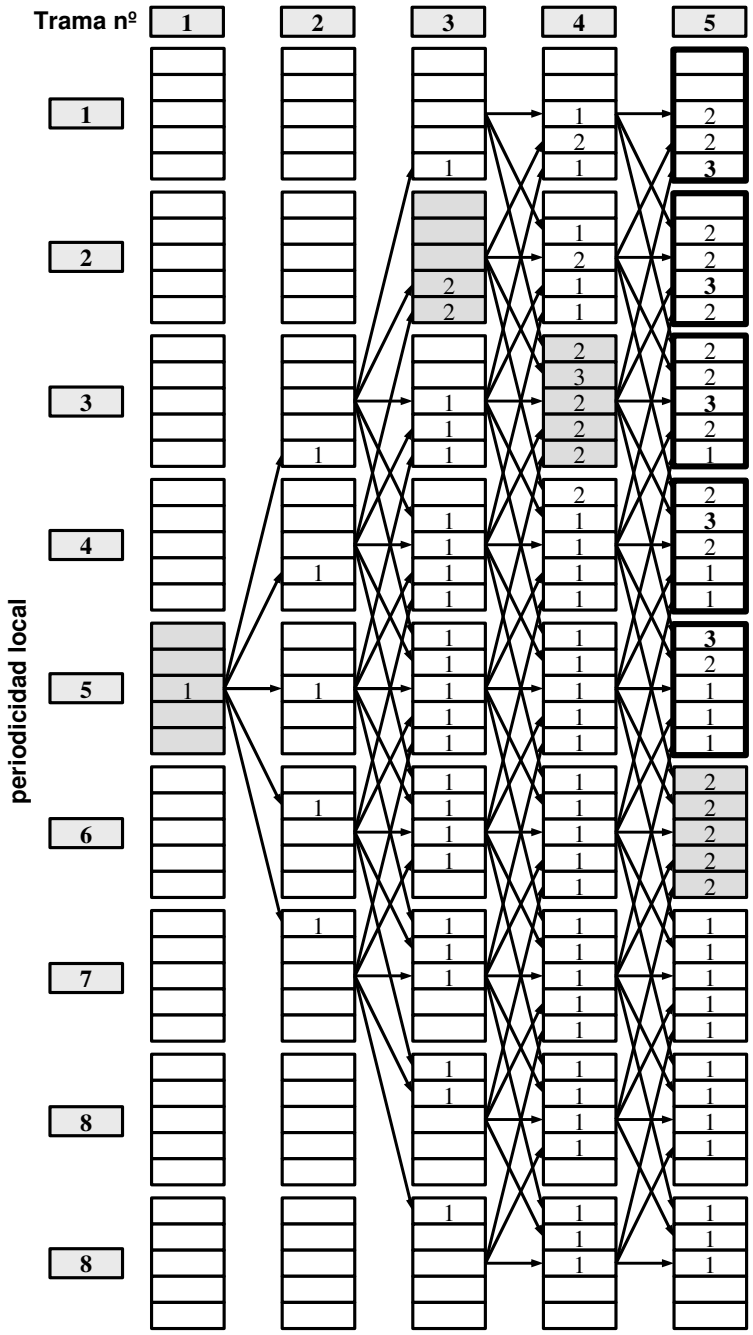


Figura 4.4: Fragmento del resultado del proceso *forward* del algoritmo de programación dinámica (partiendo sólo de la casilla p_{51} para simplificar la representación), restringido por una $S_{max} = 2$, sobre la misma matriz binaria de la figura 4.3.

de casillas activas acumuladas), es frecuente la presencia de diversos caminos con métrica acumulada máxima (las N soluciones posibles de $c_n^s(j)$ en la ecuación (4.8)). Por ello, es necesario incluir un segundo criterio para escoger el camino óptimo entre los N caminos candidatos que llegan a las C casillas de la última trama con métrica idéntica. En este caso, durante el proceso *backward* —también denominado *N-backtrackin*— se escogerá como camino óptimo aquél que presente una menor variabilidad global —dando lugar a la curva de *pitch* más suave de entra las soluciones de la ecuación (4.8)— a lo largo de la estructura *trellis* construida sobre la matriz de periodicidad $\mathbf{P}$, según la ecuación (4.9).

$$c^s(j)^* = \underset{n}{argmin}\left(\sum_{j=1}^{J}\left|c_n^s(j) - c_n^s(j-1)\right|\right) \tag{4.9}$$

por lo que el valor de periodicidad estimado por trama, una vez finalizada la primera fase del proceso de optimización (filtrado los errores), será $\hat{T}_0(j) = T_{0\,min} + c^s(j)^* - 1$.

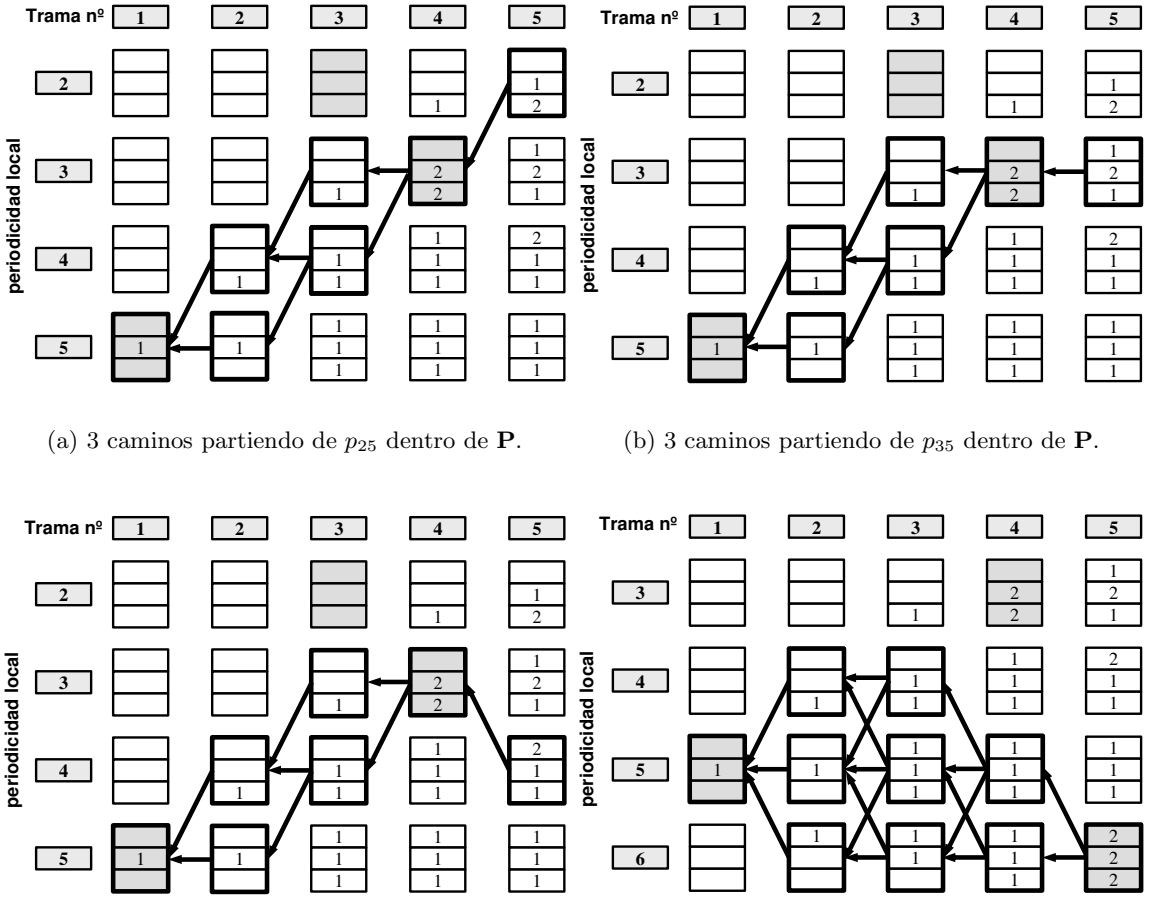
En el ejemplo de la figura 4.3 se puede observar como en la última columna de la estructura (p_{i5}) existen $C = 4$ casillas con la misma métrica acumulada (según la ecuación (4.8)): p_{25}, p_{35}, p_{45} y p_{65} (resaltadas mediante el correspondiente recuadro en negrilla). Gracias al proceso *N-backtracking*, primero se determinará el conjunto de caminos posibles $c_n^s(j)$ que llegan a las 4 casillas con una métrica acumulada equivalente, para, a continuación, escoger de entre ellos el camino que presente una menor variación global, minimizando el criterio de variación de la periodicidad (maximizando la suavidad) de la curva de F_0, dando lugar al camino óptimo de los N posibles $c_n^s(j)$ (ver ecuación (4.9)). En la figura 4.5 se presenta todos los caminos resultantes de la primera fase del *N-backtracking*, que parten de la casilla p_{51}, como resultado de la aplicación del algoritmo descrito sobre la estructura *trellis* definida en la figura 4.3 con $S_{max} = 1$ —todos ellos presentan la misma métrica acumulada (según la ecuación (4.8)) ya que pasan por 2 casillas activas de la matriz de periodicidad del ejemplo presentado en la figura.

En la tabla 4.1 se presentan todos los caminos obtenidos junto a su variación global calculada según la ecuación (4.9). De entre los caminos posibles con menor variabilidad (en este caso parten de la casilla p_{65}), finalmente se escoge el camino $c^1(j)^* = c_{19}^1(j) = (5, 6, 6, 6, 6)$, representado en la figura 4.5(d), tomando como criterio final de selección escoger el camino, de entre los que presentan menor variabilidad global, que presenta una menor variabilidad del camino óptimo trama a trama de derecha a izquierda (es decir, la que mantenga una menor variabilidad de final a inicio[14]). Por otro lado, para el caso de la estructura *trellis* obtenida según $S_{max} = 2$ en la figura 4.4, el camino óptimo será distinto al obtenido con $S_{max} = 1$, ya que gracias a aumentar el valor de S_{max} el proceso *forward* puede observar la casilla $p_{23} = 1$ —consiguiendo, en este caso, $C = 5$ casillas finales con una métrica acumulada de 3— a diferencia de trabajar con $S_{max} = 1$. En este caso, y aplicando el mismo proceso de *N-backtracking* que se acaba de describir, se obtiene el camino $c^2(j)^* = (5, 3, 2, 3, 3)$, ya que es capaz de albergar tres casillas activas de la matriz y posee la mínima variabilidad

[14]En un futuro, se estudiarán otros criterios, además de minimizar la pendiente (primera derivada) trama a trama, como por ejemplo, considerar también la segunda derivada para escoger el camino óptimo final.

Tabla 4.1: Caminos obtenidos del N-*backtracking* ($N = 19$) aplicado al ejemplo de la figura 4.3 con $S_{max} = 1$, partiendo de las $C = 4$ casillas con métrica acumulada idéntica.

Casilla	**Camino**	**Varición**
p_{25}	$c_1^1(j) = (5, 4, 3, 3, 2)$	3
p_{25}	$c_2^1(j) = (5, 4, 4, 3, 2)$	3
p_{25}	$c_3^1(j) = (5, 5, 4, 3, 2)$	3
p_{35}	$c_4^1(j) = (5, 4, 3, 3, 3)$	2
p_{35}	$c_5^1(j) = (5, 4, 4, 3, 3)$	2
p_{35}	$c_6^1(j) = (5, 5, 3, 3, 3)$	2
p_{45}	$c_7^1(j) = (5, 4, 3, 3, 4)$	3
p_{45}	$c_8^1(j) = (5, 4, 4, 3, 4)$	3
p_{45}	$c_9^1(j) = (5, 5, 4, 3, 4)$	3
p_{65}	$c_{10}^1(j) = (5, 4, 4, 5, 6)$	3
p_{65}	$c_{11}^1(j) = (5, 4, 5, 5, 6)$	3
p_{65}	$c_{12}^1(j) = (5, 5, 4, 5, 6)$	3
p_{65}	$c_{13}^1(j) = (5, 5, 6, 5, 6)$	3
p_{65}	$c_{14}^1(j) = (5, 6, 5, 6, 6)$	3
p_{65}	$c_{15}^1(j) = (5, 6, 6, 5, 6)$	3
p_{65}	$c_{16}^1(j) = (5, 5, 5, 5, 6)$	1
p_{65}	$c_{17}^1(j) = (5, 5, 5, 6, 6)$	1
p_{65}	$c_{18}^1(j) = (5, 5, 6, 6, 6)$	1
p_{65}	$c_{19}^1(j) = (5, 6, 6, 6, 6)$	1



(a) 3 caminos partiendo de p_{25} dentro de $\mathbf{P}$.

(b) 3 caminos partiendo de p_{35} dentro de $\mathbf{P}$.

(c) 3 caminos partiendo de p_{45} dentro de $\mathbf{P}$.

(d) 10 Caminos partiendo de p_{65} dentro de $\mathbf{P}$.

Figura 4.5: Representación de los $N = 19$ posibles caminos óptimos que se obtienen del ejemplo presentado en la figura 4.3 con $S_{max} = 1$ durante de la fase de *backtracking*.

de entre las soluciones halladas. Por lo tanto, a través de este pequeño ejemplo, se puede observar claramente la dependencia del resultado con el valor de S_{max} escogido, parámetro fundamental del algoritmo de filtrado de marcas de *pitch* propuesto (por lo que será un elemento clave dentro del apartado 4.6 de experimentos que se presenta más adelante).

Finalmente, en la figura 4.6 se presenta un ejemplo sobre señal real del resultado de este proceso, mostrando el camino óptimo obtenido sobre la matriz de valores candidatos de periodicidad por trama siguiendo todos los criterios para seleccionar la secuencia óptima de valores que se acaban de describir, en este caso, con $S_{max} = 3$. Como resultado de este proceso se obtiene la secuencia de valores estimados de $\hat{T}_0(j)$ por trama a lo largo de las J tramas de análisis consideradas, filtrando tanto los valores espurios que pudieran contener los datos de entrada ($F_0(t)$ del PDA o $m^i(n)$ del PMA) como los que no cumplen con la

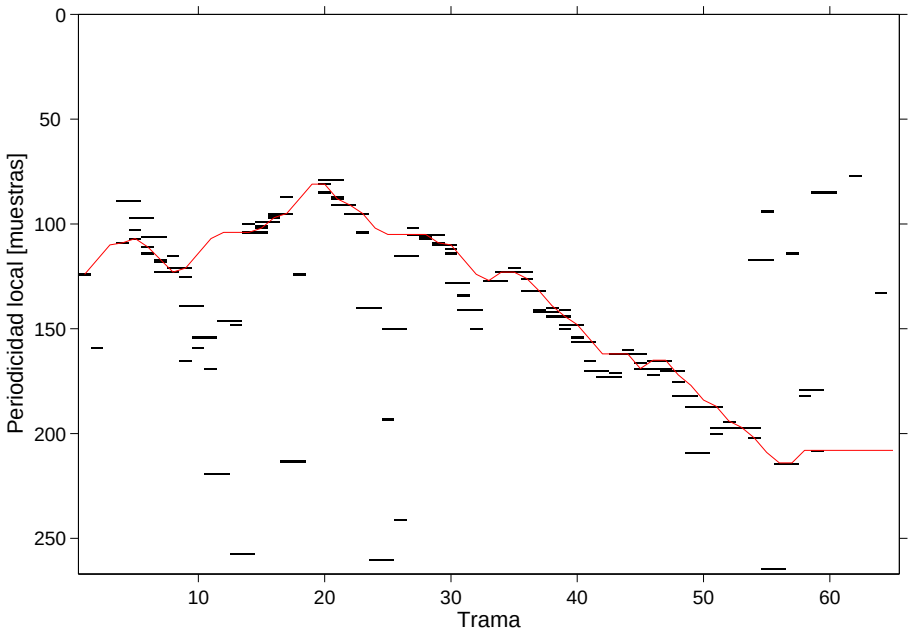


Figura 4.6: Ejemplo real de la matriz de periodicidad **P**, sobre la que se representa el camino óptimo obtenido después de la primera fase de aplicación del algoritmo de programación dinámica restringido (filtrado de errores) con $S_{max} = 3$.

restricción de variabilidad trama a trama fijados por el valor de S_{max}.

4.4.3. Ajuste local de las marcas de *pitch*

Una vez obtenida la secuencia de valores de periodicidad por trama $\hat{T}_0(j)$ $(1 \leq j \leq J)$, se procede a posicionar las marcas de *pitch* dentro de cada uno de los periodos de la señal de voz. Para ello, de los distintos criterios locales que existen para ubicar la marca dentro del periodo de la señal sonora, en el presente trabajo de investigación, se ha optado por ajustar las marcas de *pitch* al máximo de amplitud de la señal en valor absoluto mediante el proceso equivalente al descrito en (Goncharoff y Gries, 1998). No obstante, como ya se ha comentado, el PMFA admite cualquier otro criterio local, como por ejemplo, el descrito en (Lin y Jang, 2004) que escoge colocar si es mejor colocar las marcas sobre los máximos o los mínimos de la señal mediante un proceso previo.

Con el objetivo de posicionar temporalmente las marcas de *pitch*, primero se aplica el sPMA descrito en el algoritmo (4.1) para disponer de una primera estimación de las marcas finales $(m_1^f(n))$ (ver figura 4.8(a)) —siguiendo un proceso idéntico al descrito para determinar las marcas de *pitch* a partir del PDA, ya que en este estadio del proceso el PMFA vuelve a disponer de una estimación de la periodicidad a nivel de trama, pero que

en este caso es más robusta. A continuación, una vez obtenidos los K^f índices de marca junto a sus posiciones $I_{m_1^f}(k)$ según el algoritmo (4.4), se procede a ajustar las marcas de *pitch* dentro de cada uno de los periodos de señal según el criterio local y el proceso de maximización de este criterio considerados. Para ello, se genera una matriz $\mathbf{S}$ de tramas de señal $x(n)$ *pitch*-síncronas a partir de ventanas *hanning* de tamaño $2 \cdot T_{0\,max}$ centradas en las posiciones indicadas por el vector $I_{m_1^f}(k)$ tal y como se indica en la ecuación (4.10) (Goncharoff y Gries, 1998) (ver el ejemplo presentado en la figura 4.7).

$$\begin{aligned} \mathbf{S} &= (s_{ij}),\ \forall(1 \leq i < 2 \cdot T_{0max},\ 1 \leq j \leq K^f), \\ s_{ij} &= \Theta\big(x(n)\big) \cdot w_h(i)\ ,\ n = i + I_{m_1^f}(j) - T_{0\,max} + 1 \end{aligned} \tag{4.10}$$

donde $w_h(i)$ representa una ventana *hanning* de tamaño $2 \cdot T_{0max}$ y centrada en la marca $I_{m_1^f}(j)$ que pondera los $2 \cdot T_{0max}$ valores del criterio local $\Theta\big(x(n)\big)$ considerado.

De este modo, en la línea de lo expuesto en el apartado 4.2, se contempla un máximo de $m = 2 \cdot T_{0max}$ posiciones candidatas para ubicar la marca de *pitch* dentro de cada periodo —todas aquellas muestras de señal dentro de la ventana de análisis que se ajusten al criterio local considerado ($\Theta\big(x(n)\big)$ en la ecuación (4.10)). No obstante, la ventana *hanning* va penalizando gradualmente las muestras candidatas a medida que estas se alejan del centro de la trama (posición inicial estimada de la marca de *pitch*). La información a extraer de la señal de voz $x(n)$ dependerá del criterio local utilizado. En este caso, la métrica utilizada para seleccionar la posición escogida es proporcional a la amplitud de la señal en valor absoluto, por lo que $\Theta\big(x(n)\big) = |x(n)|$ en la ecuación (4.10) y se busca ajustar las marcas al máximo de amplitud del periodo. Para ello, es necesario disponer de un proceso encargado de maximizar el ajuste de las marcas de *pitch* al criterio local considerado a lo largo de la señal de voz[15]. En este caso, una vez obtenida la matriz $\mathbf{S}$ de tramas *pitch*-síncronas de dimensión $2 \cdot T_{0\,max} \times K^f$, se aplica de nuevo el algoritmo de programación dinámica descrito anteriormente, pero ahora, con el objetivo de determinar la desviación temporal (en muestras) que se debe aplicar a cada marca de *pitch* (*offset*, en inglés) para que ésta se ajuste a la posición óptima del periodo según el criterio local considerado, en este caso, la posición de máxima amplitud de la señal en valor absoluto[16] (ver (Goncharoff y Gries, 1998) para más detalles). La continuidad de la desviación aplicada en tramas *pitch*-síncronas consecutivas se asegura, de nuevo, mediante una nueva restricción de variabilidad máxima S'_{max}, siendo $S'_{max} > S_{max}$ (Goncharoff y Gries, 1998). De este modo, el algoritmo de programación dinámica tiene un margen mayor de variabilidad intertrama (menor dependencia trama a trama) para determinar el *offset* particular de cada marca (variación intratrama) respecto a la posición inicial estimada durante la fase de filtrado de errores, por lo que S'_{max} debe aumentar su valor para permitir mayores diferencias en la variación relativa de la posición

[15] Como se ha comentado en la introducción del capítulo, sólo maximizar localmente el criterio local, sin ninguna restricción de variación global, puede provocar discontinuidades trama a trama de la frecuencia fundamental estimada.

[16] Si por ejemplo, el criterio local fuera colocar la marca en el punto de máxima pendiente, sería necesario incorporar esta información en la ecuación (4.10), p.ej. mediante $\Theta\big(x(n)\big) = x(n+1) - x(n)$, y el algoritmo de programación dinámica debería definir el ajuste local de marcas a partir del camino que consiguiera, trama a trama, determinar la posición de pendiente máxima más cercana al centro de cada trama.

de la marca de *pitch* estimada en periodos de la señal de voz consecutivos[17].

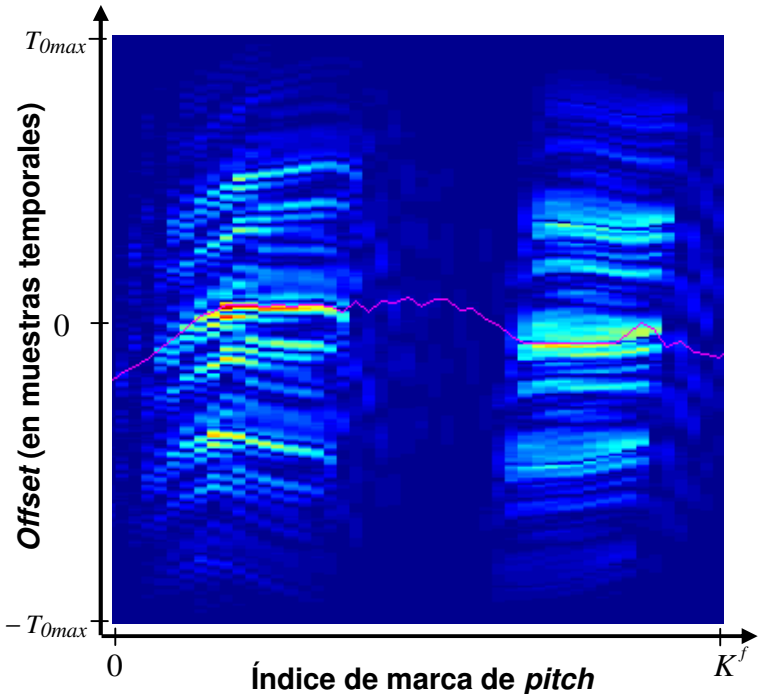


Figura 4.7: Ejemplo de la matriz de señal **S**, representada mediante su sonograma, sobre la que se representa el camino óptimo obtenido después de la segunda fase de aplicación del algoritmo de programación dinámica restringido (ajuste local de marcas) del PMFA.

Siguiendo lo descrito para la primera fase del método (filtrado de errores), el camino óptimo será aquel que consiga una métrica (amplitud en valor absoluto, en este caso) acumulada máxima sobre la matriz **S** (ver ecuación (4.8) y figura 4.7). Es decir, será aquella secuencia de posiciones de la matriz que se ajuste mejor a los máximos de la señal en valor absoluto dentro de cada trama *pitch*-síncrona de análisis, cumpliendo con la restricción de variación máxima considerada. A continuación se aplica el algoritmo de *backtracking* descrito, con la particularidad que, en este caso, la probabilidad de obtener más de un camino óptimo con la misma métrica acumulada es mínima, ya que la métrica utilizada es ahora $\mathbb{R}$. No obstante, el algoritmo escogería el camino óptimo de los N posibles caminos utilizando los mismos criterios de suavidad que se han descrito para la estimación de la periodicidad de trama sobre la matriz **P**. Finalmente, el camino óptimo obtenido se puede interpretar como un ajuste fino (*offset* de la posición inicial de las marcas referenciado al centro de la ventana de análisis de tamaño $2 \cdot T_{0max}$) utilizada (ver figura 4.8(b)) de la estimación

[17]Como se ha comentado, la señal de voz no es una señal perfectamente periódica y esto provoca que la desviación entre la posición inicial estimada de la marca y el punto que maximiza el criterio local considerado en periodos consecutivos no sean idénticos (es decir, necesiten ajustes particulares).

incial de la posición de las marcas ($m_1^f(n)$) (ver figura 4.8(a)), ajustándolas a sus posiciones finales $m^f(n)$, resultado final del proceso descrito en este apartado (ver figura 4.8(c)).

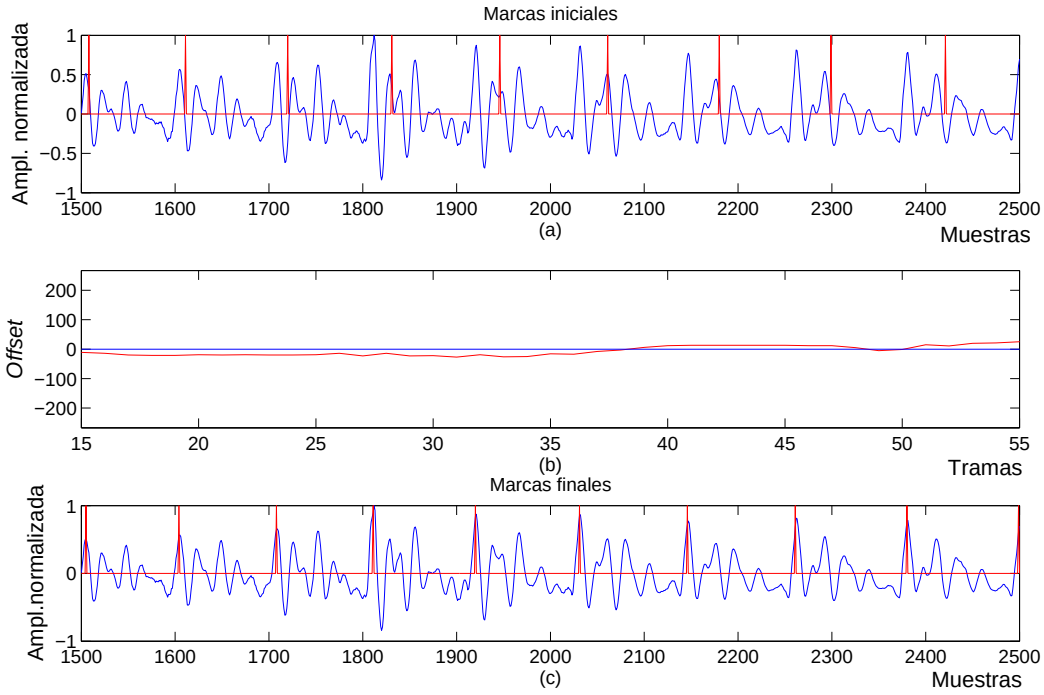


Figura 4.8: Ejemplo del resultado de (a) el posicionamiento inicial de las marcas, (b) el desplazamiento temporal a aplicar a la posición inicial estimada de las marcas *pitch* (*offset*) en la zona de interés y (c) el alineamiento final conseguido, a partir de la matriz de señal **S** de la figura 4.7 (en este caso, las marcas se han ajustado al máximo de amplitud de la señal en valor absoluto).

4.5. Evaluación

Para evaluar el funcionamiento del PMFA en términos de PDA y de PMA es necesario disponer de: *(i)* unas medidas objetivas y/o subjetivas que permitan validar el buen funcionamiento del algoritmo propuesto, *(ii)* unos algoritmos de referencia para comparar los resultados obtenidos, y *(iii)* un corpus de voz con valores de referencia (F_0 -o T_0- y/o marcas de *pitch*)[18] donde contrastar las mejoras obtenidas. A continuación se describen estos

[18]Generalmente, los valores de referencia se obtienen a partir de la señal EGG grabada conjuntamente con la señal de voz, por ejemplo (Krishnamurthy y Childers, 1986; Bagshaw, Hiller y Jack, 1993; Sakamoto y Saito, 2000; de Cheveigné y Kawahara, 2002; Sun, 2002; Kounoudes, Naylor y Brookes, 2002; Bánhalmi et al., 2005; Hosom, 2005; Dikshit, Zahorian y Nagulapati, S., 2005), aunque en el corpus desarrollado en el

elementos para el presente trabajo de investigación, haciendo énfasis en la presentación de una nueva medida de evaluación objetiva de PMAs que permite comparar de forma genérica y fiable su funcionamiento, independientemente del criterio local de posicionamiento de marcas utilizado (máximos de amplitud, paso por cero, etc.).

4.5.1. Medidas de evaluación

Para comprobar el buen funcionamiento de un PDA o un PMA, es necesario disponer de medidas de evaluación que permitan validar su bondad respecto a los valores de F_0 (o T_0) o las marcas de *pitch* de referencia. En este trabajo, el funcionamiento del PMFA presentado se evalúa doblemente. Por un lado, éste se valida analizando su funcionamiento mediante medidas de evaluación y corpus de referencia utilizados habitualmente en el ámbito de los PDAs; y, por otro lado, el PMFA se evalúa como marcador de *pitch* (PMA) utilizando un corpus de voz propio (con un mayor tamaño respecto a los típicamente utilizados como referencia en otros trabajos) junto a una nueva medida de evaluación que se presenta en este trabajo.

Evaluación de los PDAs: A continuación se describen las medidas de evaluación más usadas para analizar objetivamente el funcionamiento de los algoritmos de extracción o detección de *pitch* (Rabiner et al., 1976; Bagshaw, 1994). Habitualmente se evalúa por separado el funcionamiento de los dos bloques que típicamente constituyen el PDA por separado: el bloque de detección de sonoridad y el bloque encargado de determinar la periodicidad. Sin embargo, existen trabajos en los que la comparativa entre PDAs se realiza desactivando el módulo de sonoridad para realizar una evaluación global del algoritmo (de Cheveigné y Kawahara, 2001; Sun, 2002; Gerhard, 2003; Hosom, 2005).

La tasa de error en la estimación de la sonoridad de las tramas, es decir, el porcentaje de tramas sonoras o sordas etiquetadas erróneamente, se calcula mediante el *Unvoiced Error Rate* (UER) y el *Voiced error Rate* (VER) (Bagshaw, Hiller y Jack, 1993; Droppo y Acero, 1998; Sun, 2002; Kasi y Zahorian, 2002; Li, Malkin y Bilmes, 2004; Sha, Burgoyne y Saul, 2004; Sha y Saul, 2005; Liu et al., 2005; Achan et al., 2005). Asimismo, la tasa de error en la estimación del valor de la frecuencia fundamental por trama suele computarse típicamente mediante el *Gross Error Rate* (GER). El GER contabiliza el porcentaje de tramas —sólo considerando las tramas correctamente designadas como sonoras— que se desvían de forma significativa del valor de F_0 de referencia, es decir, se trata de un cálculo relativo. Se contabilizará como error aquella trama cuya diferencia relativa respecto al valor de referencia supere un determinado umbral[19], generalmente del 20 % (Bagshaw, Hiller y Jack, 1993; Bagshaw, 1994; Ying, Jamieson y Michell, 1996; Sun, 2000; de Cheveigné y Kawahara, 2001; Sun, 2002; Kasi y Zahorian, 2002; Li, Malkin y Bilmes, 2004; Sha, Burgoyne y Saul, 2004;

marco de este trabajo no se incorpora esta información al no disponer del equipo necesario.

[19]Este umbral se define para absorber valores de F_0 claramente alejados del valor de referencia, de ahí el adjetivo *gross* que acompaña al nombre de la métrica. En la literatura, a estos errores se les suele denominar F_0 *halving* o *doubling*, sin que esto signifique estrictamente que el PDA haya cometido errores de octava (F_0 mitad o doble) en la estimación de la F_0 de la trama (Bagshaw, 1994; Liu et al., 2005).

Sha y Saul, 2005; Achan et al., 2005; Liu et al., 2005; Bánhalmi et al., 2005). Asimismo, las tramas con una desviación menor al umbral definido pueden ser contabilizadas como errores finos de estimación, calculados mediante el *Fine Pitch Error Rate* (FPER). En este caso, en lugar de calcular el porcentaje de tramas con este tipo de error, se puede calcular la media y la varianza de las desviaciones (Droppo y Acero, 1998), mediante el error cuadrático medio (o en inglés *Mean Square Error*, MSE) (Kasi y Zahorian, 2002) o el RMSE (*Root* MSE) (Sha y Saul, 2005).

Evaluación de los PMAs: Así como para los PDAs existen diversas medidas más o menos estandarizadas para evaluar el funcionamiento de cada uno de los módulos que los componen, el caso de los PMAs es distinto, ya que no existe una medida de evaluación tan *estandarizada*. Por un lado, Veldhuis (2000) indica que la bondad de los PMA sólo puede ser analizada de forma indirecta a través del resultado de su aplicación, por ejemplo, mediante modificaciones prosódicas de la señal síncronas con el *pitch*. Por otro lado, existen trabajos que proponen medidas objetivas de evaluación comparando las marcas obtenidas con las marcas de referencia (revisadas manualmente), mediante comparación directa marca a marca (Chen y Kao, 2001; Lin y Jang, 2004) —se computa como error la marca estimada que no coincide exactamente con la posición de referencia—, o bien, dando un cierto margen en la comparación para permitir ciertos desalineamientos no computables como error —en (Sakamoto y Saito, 2000) se permite una diferencia relativa del 5 % o una variación absoluta de $\pm 0.25ms$ (Kounoudes, Naylor y Brookes, 2002) o $2.5ms$ (Vincent, Rosec y Chonavel, 2006). Asimismo, si las marcas siguen el mismo criterio local de ubicación, se puede computar la *tasa de omisiones* e *inserciones* del algoritmo evaluado (tasas de falso rechazo y falsa alarma), contabilizando el número de marcas que faltan y el número de marcas que sobran respecto al total de marcas de referencia, respectivamente (Vincent, Rosec y Chonavel, 2006).

De cualquier modo, todas estas comparativas sólo son aplicables si las marcas de referencia y las evaluadas comparten el mismo criterio local, es decir, se ajustan según el mismo criterio de posicionamiento dentro del periodo de la señal, lo facilita la comparativa marca a marca. Así pues, cuando los PMAs no compartan el criterio (p.ej. máximo de la señal *vs.* primer paso por cero antes del máximo de la señal), será más complicado comparar su funcionamiento, debido a la desviación intrínseca provocada por el criterio local considerado. Para evitar este problema, existe la posibilidad de primero alinear temporalmente las marcas y luego compararlas (Harbeck et al., 1995; Kounoudes, Naylor y Brookes, 2002; Vincent, Rosec y Chonavel, 2006). Sin embargo, este proceso puede emmascarar la comparativa debido a los posibles errores de alineamiento.

En este trabajo se propone una nueva medida de evaluación para los PMAs, inspirada en el GER de los PDAs, que permite comparar marcadores con criterios de ajuste local distintos. Esta medida se ha denominado *Gross Pitch Marks Error Rate* (GPMER) (ver ecuación (4.14)), y consiste en evaluar las marcas de *pitch* a partir de sus diferencias relativas de periodicidad: p_r para la periodicidad de referencia (ecuación (4.11)) y p'_r para la periodicidad estimada por el PMA utilizado (ecuación (4.13)), en lugar de hacerlo según

su posición específica dentro de la señal de voz[20]. Si esta diferencia es mayor que un determinado umbral γ (en este trabajo, se utiliza $\gamma = 0.2$, por similitud con el umbral del 20 % del GER), esa marca será considerada como errónea. Por lo tanto, se trata de una medida inspirada en GER, siendo, de algún modo, un refinamiento de GER, ya que éste solo analiza la bondad de los PDA a nivel de trama, mientras que GPMER lo hace a nivel de marca, tomando el intervalo de comparación según las marcas de referencia. De este modo, a partir de la posición de las marcas de referencia se podrán detectar con cierta facilidad las inserciones (demasiadas marcas entre dos marcas de referencia) y las omisiones (inexistencia de marcas). Durante la comparativa de valores de periodicidades evaluadas dentro del rango de comparación indicado por las marcas de referencia, las omisiones serán detectadas como valores de periodicidad significativamente mayores a la periodicidad de referencia, mientras que las inserciones se detectarán por una secuencia de valores de periodicidad evaluada claramente menores a los de referencia (ver ejemplo de la figura 4.9). Además, para evitar sesgar el resultado de la medida de evaluación, tanto las inserciones como las omisiones serán computadas como un *único* error entre dos marcas consecutivas de referencia. Si este factor no se considerara, las inserciones provocarían un aumento ficticio del número de errores, al computar como error todas las diferencias entre marcas menores a la de referencia dentro de la zona de comparación (ver ejemplos de la figura 4.9).

Las periodicidades p_r de referencia se definen como:

$$p_r(k) = I_m(k+1) - I_m(k),\ 1 \leq k \leq K^r - 1 \tag{4.11}$$

donde K^r indica el número total de marcas y $I_m(k)$ es la secuencia de índices temporales de las marcas.

Los índices de periodicidad del PMA evaluado que entran en juego en la definición de la periodicidad $p'_r(k)$ a comparar con las periodicidades de referencia $p_r(k)$ en la medida GPMER, para un determinado índice k se definen como:

$$\mathcal{K}'_k = \left\{ k' \;\middle|\; I_m(k) \leq \frac{I_{m'}(k'+1) + I_{m'}(k')}{2} < I_m(k+1) \right\}, \tag{4.12}$$

donde $I'_{m'}(k')$ representa la secuencia de índices de las marcas generadas por el PMA evaluado y $\mathcal{K}'_k$ representa el conjunto de periodicidades estimadas asignadas a una misma periodicidad $p_r(k)$ de referencia dentro del intervalo definido por sus correspondientes marcas de referencia $I_m(k)$ y $I_m(k+1)$. Además, en la ecuación $\emptyset$ denota conjunto vacío. Se puede observar de la ecuación 4.12 que, para la asignación, se utiliza como referencia el punto medio de la posición de las marcas estimadas $\frac{I_{m'}(k'+1)+I_{m'}(k')}{2}$ (ver el ejemplo de la figura 4.9).

[20]En la línea de lo descrito por (Dikshit, Zahorian y Nagulapati, S., 2005) para validar el resultado del PMA respecto a las marcas de referencia obtenidas de la señal EGG, pero, utilizando ambos el mismo criterio local, en este caso.

$$p'_r(k) = \begin{cases} \underset{k' \in \mathcal{K}'_k}{Argmax} \Big| p_r(k) - \big(I_{m'}(k'+1) - I_{m'}(k')\big) \Big|, & \mathcal{K}'_k \neq \emptyset \\ I_{m'}\Big(\underset{r}{Argmin} \{ I_{m'}(r) \geq \frac{I_m(k+1)+I_m(k)}{2} \} \Big) - & \\ -I_{m'}\Big(\underset{r}{Argmax} \{ I_{m'}(r) \leq \frac{I_m(k+1)+I_m(k)}{2} \} \Big), & \mathcal{K}'_k = \emptyset \end{cases} \tag{4.13}$$

En la expresión (4.13) se define el cálculo de la periodicidad evaluada $p'_r(k)$ mediante una función definida por tramos aplicada sobre la secuencia de índices de marca $I'_m(k')$ y $I_m(k)$, una vez designado el alineamiento de periodicidades según el contenido de $\mathcal{K}'_k$ (4.12). En el primer caso, $\mathcal{K}'_k \neq \emptyset$, $p'_r(k)$ se define a partir de las distinas periodicidades estimadas que se han asignado a una misma periodicidad de referencia $p_r(k)$. Para la comparación, se escoge aquella $p'_r(k)$ que presente una mayor desviación —en valor absoluto— respecto a la de referencia[21], para realizar una única comparación —evitando sesgar el cálculo de GPMER debido a la presencia de inserciones. El segundo caso tiene en cuenta la situación donde $\mathcal{K}'_k = \emptyset$, es decir, ninguna periodicidad estimada $p'_r(k')$ ha sido asignada a la periodicidad $p_r(k)$ (p.ej. debido a una omisión). En este caso es necesario encontrar la primera pareja de marcas que se encuentren ubicadas antes y después del punto medio del intervalo de referencia que define esa periodicidad $p_r(k)$.

$$\text{GPMER}(\,\%) = \frac{\#\Big(\frac{|p'_r - p_r|}{p_r}\Big) > \gamma}{K^r - 1} \cdot 100 \tag{4.14}$$

donde # indica *número de* o cardinalidad, p'_r es la periodicidad local evaluada (estimada por el PMA) y p_r es el valor de referencia correspondiente, cuya posición global guía el proceso de comparación, para conseguir comparar vectores de marcas de *pitch* de longitudes distintas, como se ha descrito en las expresiones (4.12) y (4.13). Como se ha comentado, en este trabajo se utiliza $\gamma = 0.2$, como en el GER clásico.

En la figura 4.9 se presenta un pequeño ejemplo ilustrativo de distintas de las situaciones típicas en el cálculo del GPMER, a partir de lo descrito en las ecuaciones (4.12) y (4.13). En este ejemplo, se puede observar como el primer par de marcas evaluadas se asigna al primer intervalo de comparación $[I_m(1), I_m(2)]$, ya que su punto medio $I_m(1) \leq \frac{1}{2}(I_{m'}(1)+I_{m'}(2)) < I_m(2)$. No obstante, el valor de $p'_r(1)$ utilizado para la comparativa en GPMER (ecuación (4.14)) no es $a = I_{m'}(2) - I_{m'}(1)$, ya que en este caso $\mathcal{K}'_1 = 1, 2$, debido a que el punto medio de la siguiente periodicidad estimada también pertenece a $k = 1$. En este caso, y aplicando el primer caso de la ecuación (4.12), $p'r_(2) = b = I_{m'}(3) - I_{m'}(2)$, ya que $|p_r(1) - b| > |p_r(1) - a|$. Seguidamente, y siguiendo un proceso similar, se obtiene que $p'r_(2) = c$, en este caso debido a que al intervalo $[I_m(2), I_m(3)]$ sólo se le asigna un único valor de periodicidad estimada. Lo mismo sucede en $p'_r(3) = e$, con la única particularidad que las dos marcas generadas

[21] También se podría haber calculado $p'_r(k)$ para $k' \in \mathcal{K}'_k$ a partir del promedio de las distintas periodicidades, $\frac{1}{N_k} \sum_{k' \in \mathcal{K}'_k} \Big(I_{m'}(k'+1) - I_{m'}(k') \Big)$, siendo N_k el número de marcas del PMA evaluado dentro del intervalo de comaparción, pero por el momento, esta opción se ha descartado ya que puede provocar un enmascaramiento en la comparativa.

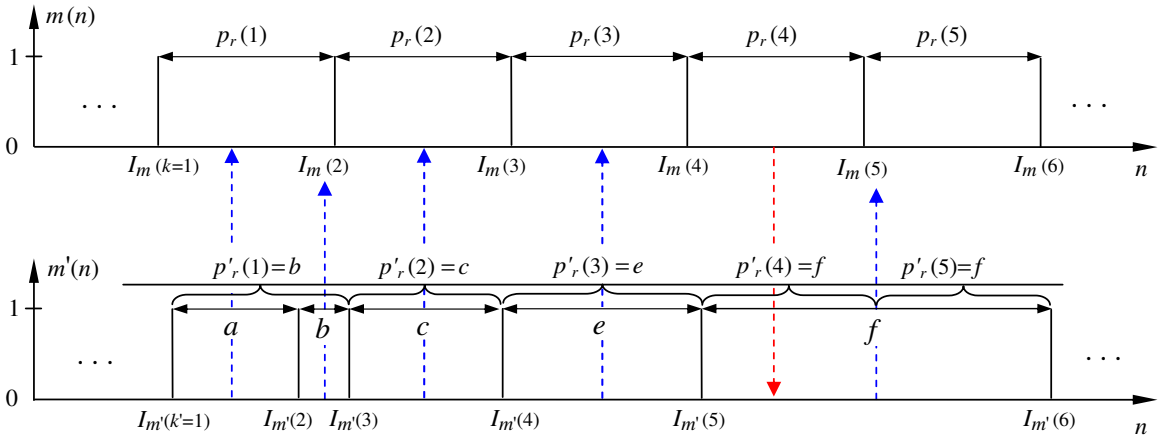


Figura 4.9: Ejemplo de la comparativa de marcas de *pitch* evaluadas $m'(n)$ respecto a la secuencia de marcas de referencia $m(n)$, según lo indicado en la ecuación (4.13).

por el PMA caen fuera del intervalo $[I_m(3), I_m(4)]$, a diferencia de $p'_r(3)$, cuestión que no afecta al cálculo según las expresión (4.12). Finalmente, el punto medio de las marcas $I_{m'}(5)$ y $I_{m'}(6)$ corresponde a $p_r(5)$, por lo que es necesario designar $p'_r(4)$ mediante el segundo caso contemplado por la expresión (4.13), ya que $\mathcal{K}'_4 = \emptyset$. Para ello, como se muestra en la figura mediante una flecha discontinua de color rojo, se busca dentro del conjunto de marcas estimadas $m'(n)$ el primer índice de marca situado por encima de este punto ($I_{m'}(6)$, en este caso) junto al primer índice de marca ubicado por debajo de este punto ($I_{m'}(5)$, en este caso). Por lo tanto, se obtiene que $p'_r(4) = f$, el mismo valor que se ha asignado a $p'_r(5)$. A partir de estos valores, se podrá calcular el GPMER correspondiente a la secuencia de marcas evaluadas $m'(n)$ según el umbral considerado en la ecuación (4.14).

Finalmente, añadir que, además de las pruebas objetivas descritas, el funcionamiento de los PMAs también se puede evaluar mediante pruebas subjetivas, pero tampoco existe un proceso estandarizado. En (Sakamoto y Saito, 2000) se realiza un test donde se evalúa la calidad de la señal de voz modificando la F_0 de la señal en un $\pm 30\,\%$ de forma *pitch*-síncrona (no se describe ni el número de frases ni de evaluadores). En (Lin y Jang, 2004) se realiza una evaluación subjetiva con 10 evaluadores, modificando el *pitch* de la señal en una octava —se intuye que modifican todo el corpus.

4.5.2. Algoritmos de referencia

En el presente trabajo, se evalúan las prestaciones del PMFA desarrollado a partir de tres algoritmos representativos del estado actual de la investigación en el ámbito de la detección y el marcado de *pitch* —que han sido utilizados como referencia en publicaciones recientes (Sun, 2002; Gerhard, 2003; Li, Malkin y Bilmes, 2004; Sha, Burgoyne y Saul, 2004; Sha y Saul, 2005; Achan et al., 2005; Bánhalmi et al., 2005; Vincent, Rosec y Chonavel, 2006). En este caso, se ha escogido trabajar con:

- RAPT implementado mediante la función *get_f0* del paquete ESPS (Talkin, 1995), como PMA representativo. RAPT utiliza la autocorrelación y la predicción lineal (LPC) como métodos de análisis de la señal y los acompaña de un postprocesamiento basado en programación dinámica, para estimar los valores de F_0 y ubicar las marcas de *pitch*.
- YIN (de Cheveigné y Kawahara, 2002) y SHRp (Sun, 2000; Sun, 2002) como PDAs, por la calidad demostrada respecto a otras propuestas y por partir de enfoques distintos para determinar la frecuencia fundamental de las tramas de voz. YIN analiza la periodicidad temporal de la señal a partir de la autocorrelación de las tramas de análisis, que se acompaña de varios postprocesos que le permiten minimizar la presencia de errores. En cambio, SHRp estudia la periodicidad de la señal en el dominio frecuencial incorporando información de la percepción del *pitch* mediante el denominado *Subharmonic-to-Harmonic Ratio* (SHR), intentando solucionar así los errores típicos de los PDA.

Los resultados obtenidos de la aplicación de PMFA serán comparados con los conseguidos por estos algoritmos, utilizados como entrada del PMFA (ver figura 4.1). De este modo, se podrá contrastar la mejora conseguida al aplicar el PMFA propuesto respecto a estos algoritmos de referencia sobre los datos utilizados.

4.5.3. Corpus de referencia

Para evaluar el funcionamiento del algoritmo propuesto, es necesario disponer de uno o más corpus de voz con valores de F_0 o marcas de *pitch* fiables —concepto que en la literatura de PDA se suele denominar como *ground truth* (Yi y Glass, 2002; de Cheveigné y Kawahara, 2002; Sun, 2002; Sha, Burgoyne y Saul, 2004; Li, Malkin y Bilmes, 2004; Liu et al., 2005). Afortunadamente, existen varios corpus de voz que pueden ser utilizados para validar el funcionamiento de los PDA (ver (de Cheveigné y Kawahara, 2001; de Cheveigné y Kawahara, 2002)). Sin embargo, se trata de corpus no demasiado extensos (entre 5 y 45 minutos), comparados con los que habitualmente se utilizan en síntesis basada en corpus (ver sección 2.1.2). Normalmente, estos corpus de referencia son multilocutor, algunos contienen más de un idioma, y suelen estar pronunciados con estilo *neutro*. De entre ellos, se ha escogido el corpus denominado *Keele database* de la Universidad de Keele del Reino Unido (Plante, Meyer y Ainsworth, 1995), ya que es uno de los más usados en la literatura (de Cheveigné y Kawahara, 2002; Kasi y Zahorian, 2002; Dikshit, Zahorian y Nagulapati, S., 2005; Sun, 2002; Achan et al., 2005; Sha, Burgoyne y Saul, 2004; Sha y Saul, 2005; Bánhalmi et al., 2005) y acompaña la señal de voz con los valores de F_0 obtenidos por trama —evitando tener que estimar la periodicidad de las tramas a partir de la señal EGG, con el trabajo de validación que esto comporta (de Cheveigné y Kawahara, 2002; Dikshit, Zahorian y Nagulapati, S., 2005).

En cambio, no existe la misma variedad de posibilidades para validar los PMAs. En los trabajos recopilados que tratan sobre PMA, los autores suelen desarrollarse su propio corpus, validando sobre él los resultados de su algoritmo: p.ej. 1 minuto en (Harbeck et al.,

1995), 2.1 minutos en (Chen y Kao, 2001), o más de 500 segundos (unos 8.5 minutos) en (Lin y Jang, 2004). Otros trabajos, validan el funcionamiento del PMA de forma indirecta, sobre una aplicación específica (Laprie y Colotte, 1998; Veldhuis, 2000), o simplemente indican que el algoritmo funciona *bien* basándose en las pruebas empíricas realizadas (Goncharoff y Gries, 1998; Colotte y Laprie, 2002). El problema fundamental de estos corpus es que se trata de colecciones de voz muy reducidas, comparado con los utilizados en los sistemas de CTH basada en selección de unidades (CTH-SU)[22] (ver sección 2.1.2). Así pues, para poder obtener resultados significativos en el contexto de la síntesis basada en corpus, se utilizará un corpus de voz (de algo más de 2.5h de duración), que será útil para validar el funcionamiento de PMFA en el contexto para el que se ha desarrollado: el etiquetado robusto de corpus de voz (con distintos estilos de locución) para CTH-SU.

***Keele database*:** Es un corpus formado por 10 frases fonéticamente balanceadas en inglés de unos 35 segundos cada una, correspondientes a "*The North Wind Story*". Estas frases están pronunciadas con estilo de locución neutro por 10 locutores distintos[23], 5 hombres y 5 mujeres. Se trata de una señal con calidad de estudio muestreada con una $f_s = 20$KHz y resolución de 16 bits/muestra. Cada fichero de voz está acompañado por los valores de F_0 a nivel de trama (tramas obtenidas mediante ventanas de $25.6ms$ cada $10ms$) calculados a partir de la autocorrelación aplicada sobre la señal EGG. Estos valores de F_0 pueden ser utilizados como referencia para validar el funcionamiento de cualquier método de PDA, ya que han sido revisadas manualmente (de Cheveigné y Kawahara, 2002; Kasi y Zahorian, 2002; Sha y Saul, 2005; Achan et al., 2005; Dikshit, Zahorian y Nagulapati, S., 2005). De todos modos, los autores indican dentro del fichero *keele_pitch_database.htm* que acompaña al corpus que existen ciertos valores de F_0 críticos[24]:

- "*Where we know that there is voiced speech, but the lx trace has been corrupted the data is set to -1 (this happens sometimes because the measurements are based on two electrodes on the skin, which can loose contact as the speakers move around)*".

- "*For segments where there is a lx trace, but no obvious speech signal, we left the values but set them to negative values. So a -200 entry means that a 100Hz F_0 was measured but that this is not reflected in the speech data.*"

  donde "*lx trace*" corresponde a la señal EGG.

[22] Recientemente se ha presentado un primer estudio en (Vincent, Rosec y Chonavel, 2006) que analiza el funcionamiento del algoritmo de estimación del instante de cierre glotal que se propone utilizando un corpus (de estilo de locución neutro) utilizado para CTH-SU: CMU ARCTIC (ver sección 3.2.2), pero sin validar la fiabilidad de la señal de EGG utilizada como referencia.

[23] Aunque en el corpus original se disponía de 15 locutores (Plante, Meyer y Ainsworth, 1995), 5 de ellos niños, la versión actualmente disponible en la dirección ftp://ftp.cs.keele.ac.uk/pub/*pitch*/Speech sólo contiene los 10 locutores adultos.

[24] Estos puntos demuestran que la señal EGG no siempre es perfecta para estimar la periodicidad de la señal, como ya se ha comentado con anterioridad.

Corpus Publicitario: Como se ha descrito en el apartado 3.4 del presente trabajo de investigación, en el marco del proyecto MCYT PROFIT FIT-150500-2002-410 se grabó un corpus de voz en español de 2.5h (C_{Pub} en la tabla 3.8) en colaboración con el Departamento de Comunicación Audiovisual y Publicidad de la Universidad Autónoma de Barcelona (CAP-UAB). El corpus, grabado por una locutora profesional, está formado por 2590 frases extraídas de una base de datos publicitaria, que fueron grabadas con tres estilos de locución distintos: alegre, neutro y sensual. Se escogió el estereotipo para cada dominio según lo indicado por los expertos del CAP-UAB. Como resultado se dispone de una señal con calidad de estudio muestreada a $f_s = 16$KHz y 16 bits/muestra con tres estilos prosódicos y calidades vocales distintas. Este corpus permitirá evaluar el comportamiento de la propuesta sobre señales de mayor variabilidad prosódica que la que es habitual en los corpus neutros (típicamente utilizados en las evaluaciones de los PDA y PMA). De este modo, se podrá estudiar también el funcionamiento tanto de PMFA como de los algoritmos de referencia sobre señales de voz con características de F_0 muy distintas (en el marco del desarrollo de sistemas de conversión de texto en habla multidominio descrito en el capítulo 3).

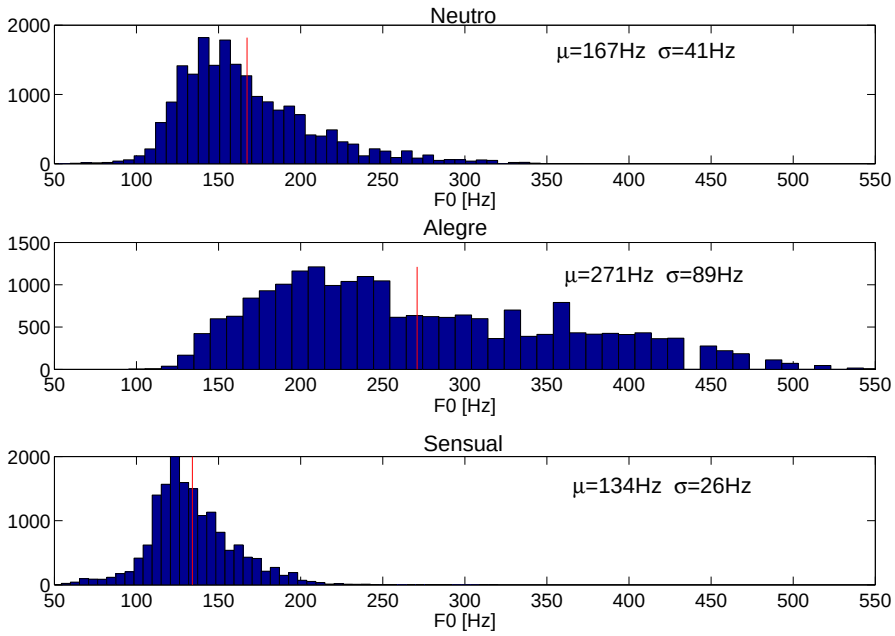


Figura 4.10: Distribución de los valores de F_0 para el corpus publicitario para cada uno de los estilos de locución que contiene: neutro, alegre y sensual.

En la figura 4.10 se presenta la distribución de valores de F_0 para cada uno de los subcorpus que se acaban de describir. Como se puede observar, el subcorpus alegre es el que presenta la mayor F_0 media y la mayor desviación de F_0 de los tres ($\mu = 271$Hz, $\sigma = 89$Hz), es decir, es el corpus más agudo y con mayor fluctuación de tono. Sin embargo, el subcorpus

sensual presenta los valores mínimos (μ = 134Hz, σ = 26Hz), es decir, es el corpus más grave y con menor variación de F_0. Finalmente, el corpus neutro presenta unos valores intermedios, tanto en media como en desviación de F_0 (μ = 167Hz, σ = 41Hz). Así pues, tanto el corpus alegre como el sensual serán más complejos de marcar que el neutro, como se verá en los experimentos que se describen a continuación, debido a la elevada variabilidad de F_0 del primero, y a la gran cantidad de señales que son prácticamente *susurros*, en el segundo —con la dificultad que esto conlleva para cualquier PDA o PMA.

Obtención de los valores de referencia: Para poder utilizar los tres subcorpus (neutro, alegre y sensual) que se acaban de presentar como referencia en la evaluación del algoritmo propuesto, es necesario disponer de sus marcas de *pitch* validadas previamente. Estas marcas se han obtenido aplicando el siguiente procedimiento[25] a los tres subcorpus:

1. Se obtiene una primera versión de las marcas de *pitch* utilizando el marcador RAPT de (Talkin, 1995). Se ajustan los parámetros de *get_f0* del paquete ESPS para abarcar el rango de frecuencias definido (de 50 a 550Hz), analizando, en este caso, la señal de voz mediante un ventaneo de $20ms$ con paso de $5ms$.

2. A continuación, se aplica un postprocesamiento automático simple sobre las marcas iniciales con el objetivo de eliminar las marcas excesivamente cercanas (sobremarcado) y rellenar las zonas sin marcas (inframarcado), normalmente debidos a tramas de transición de sonoridad. Para ello, se controla que las marcas se encuentren dentro del rango de frecuencias considerado (50 a 550Hz). La separación entre marcas de relleno en las zonas sordas se obtiene a partir de la interpolación lineal del valor de la periodicidad de las zonas sonoras vecinas.

3. Se aplica el programa *CorpusTester* (ver sección C.1) sobre las marcas obtenidas automáticamente. El objetivo es validar las marcas e indicar las zonas que necesitan ser revisadas manualmente, evitando así inspeccionar todas las marcas del corpus (como se ha comentado la duración total del corpus supera las 2.5h de duración). El programa *CorpusTester* se ha configurado según las características de cada uno de los subcorpus. En primer lugar, para evitar la presencia de marcas espurias se restringe el margen de frecuencias permitido en cada uno de ellos ([60,400] Hz para el neutro, [90 550] Hz para el alegre y [50, 300] Hz para el sensual), después de una primera revisión visual del rango de valores aproximado de F_0 en cada subcorpus. En segundo lugar, se controla la continuidad de la secuencia de marcas de *pitch* para evitar la presencia de saltos bruscos de T_0 mediante un *factor de detección de marca ausente*. Este valor define la variación máxima permitida de T_0 entre marcas consecutivas (umbral). Este es un elemento clave para detectar las zonas que han estado sobremarcadas (separación entre marcas menor al umbral), inframarcadas (separación entre marcas superior al umbral) o bien zonas donde no se han colocado marcas. En este caso, la variación máxima permitida entre marcas consecutivas se fijó experimentalmente a 38 % para los

[25] De algún modo, este proceso fue el embrión del algoritmo dinámico restringido mediante S_{max} sobre el que se basa el PMFA propuesto.

corpus neutro y sensual, y a 35 % para el alegre (inicialmente también se ajustó a 38 %, pero resultó necesario refinar el umbral al comprobarse la persistencia de errores).

4. Finalmente, se revisan manualmente las zonas de los ficheros indicadas por el *CorpusTester* para evitar la presencia de errores; proceso equivalente al realizado en otros trabajos, p.ej. (Sakamoto y Saito, 2000; Chen y Kao, 2001; Lin y Jang, 2004). La corrección manual de las marcas se ha realizado utilizando la herramienta ITP (ver sección C.1). Una vez modificadas dichas marcas se vuelve a aplicar el proceso de validación (*CorpusTester*) hasta conseguir eliminar los errores de marcado—según los criterios establecidos. Esta fase ha resultado ser un proceso bastante costoso, ya que el número total de marcas verificadas supera las 900K (aunque no todas tuvieron que ser modificadas, claro está), con un tiempo aproximado de revisión de varias semanas (unos tres de meses en total) con tres personas dedicadas a esta tarea.

4.6. Experimentos y resultados

En esta sección se presentan el conjunto de experimentos desarrollados para evaluar el funcionamiento del método de generación robusta de marcas de *pitch* propuesto, tanto en términos de GER[26] como de GPMER (debido a que PMFA no incluye ningún proceso de estimación de la sonoridad, el VER o UER están incluidos, de algún modo, en los resultados). De este modo, se analizan, por un lado, los valores de F_0 estimados por trama, y por otro, la posición de las marcas de *pitch* a lo largo de la señal de voz. El análisis del PMFA como PDA (GER) también será utilizado para analizar los resultados obtenidos mediante la nueva medida de evaluación de PMAs introducida (GPMER).

Todas las pruebas se realizan ajustando el margen de valores de F_0 para todos los algoritmos y todos los corpus a [50, 550] Hz, siendo éste un rango de valores representativo entre las distintas propuestas que se han encontrado en la literatura y aplicable, de entrada, para el marcado del abanico de frecuencias fundamentales presentes en los corpus utilizados. En este trabajo, y como primer paso para estudiar el funcionamiento del PMFA desarrollado, se estudian dos tamaños de ventana distintos: $5ms$ y $10ms$, que permiten albergar desde 0 a 3 marcas para $5ms$ y de 0 a 6 marcas para $10ms$ por trama de análisis, dado el margen de frecuencias considerado[27]. Por otro lado, los PDA y PMA utilizados han ajustado sus ventanas de análisis según cada corpus: para el caso del corpus *Keele* se ha utilizado una ventana de análisis de $25.6ms$ con un paso de $10ms$, siguiendo las especificaciones del mismo (Plante, Meyer y Ainsworth, 1995), para disponer de una comparativa directa respecto a los valores de F_0 de referencia. Asimismo, se adapta la frecuencia de muestreo de 20KHz a 16KHz para disponer de un análisis de los resultados obtenidos por las distintas

[26] En este trabajo, el valor de la F_0 por trama de los PMA se calcula como el valor medio del inverso de la distancia entre marcas de *pitch* dentro de la trama analizada, ya que el dato de partida utilizado son las marcas de *pitch* finales $m^f(n)$.

[27] En un futuro se pretenden analizar otros tamaños de ventana e incorporar cierto solapamiento entre las ventanas de análisis del PMFA para estudiar qué impacto tiene una mayor redundancia de los datos a la hora de ajustar de forma robusta las marcas de *pitch* de la señal a etiquetar.

configuraciones de PMFA sobre este corpus comparable con los obtenidos en el corpus publicitario ($f_s = 16$KHz). Para el corpus publicitario se ha escogido una ventana de análisis de $20ms$ para albergar, como mínimo, un periodo de la señal más grave (T_{0max}) del rango de frecuencias considerado. Las comparativas de algoritmos que se presentan a continuación se han realizado únicamente sobre las tramas etiquetadas en el corpus *Keele* con valores de F_0 no nulos, mientras que para el corpus publicitario, se han utilizado las tramas que corresponden a los fonemas sonoros —obtenidas a partir de la trascripción fonética y de las marcas de segmentación (ambas revisadas previamente de forma manual). De este modo, se están considerando tanto tramas completamente sonoras como tramas de sonoridad dudosa (transiciones de fonemas, fonemas fricativos sonoros, etc.) para evaluar el funcionamiento de los algoritmos estudiados. Por lo tanto, en lugar de centrar el estudio en las tramas claramente sonoras como se ha comentado, la evaluación sólo elimina las tramas claramente sordas, manteniendo las tramas de sonoridad ambigua dentro del intervalo, ya que éstas son las más complicadas de etiquetar y suelen provocar los mayores problemas durante la síntesis del habla *pitch*-síncrona (p.ej. utilizando TD-PSOLA (Moulines y Charpentier, 1990)).

Por otro lado, como se ha comentado en la descripción teórica del algoritmo, el PMFA sólo necesita del ajuste de los valores del parámetro de restricción de pendiente máxima, una vez definido el margen de frecuencias fundamentales contemplado y el criterio local de posicionamiento de marcas utilizado. Este parámetro define la variación máxima permitida trama a trama para la búsqueda dinámica de la secuencia óptima de valores de $\hat{T}_0(j)$, en la primera fase (S_{max}), y la posición de las marcas de *pitch*, en la segunda fase (S'_{max}). En (Goncharoff y Gries, 1998), se indica que el valor de S'_{max} de la segunda fase del algoritmo de programación dinámica restringido debe ser mayor que el de la primera, por lo que la variación entre ajustes de la posición estimada inicialmente para la marca dentro de cada trama para tramas consecutivas puede ser mayor para poder ajustar la posición final de las marcas según el criterio de ubicación local de las marcas escogido. Como se ha comentado anteriormente, en este trabajo se han seguido estos mismos criterios, ya que PMFA tiene el mismo objetivo que el algoritmo descrito en (Goncharoff y Gries, 1998), aunque trabaje con otro tipo de dato de entrada (las marcas de *pitch* o valores de F_0 estimados por otro PMA o PDA). Por ejemplo, en (Goncharoff y Gries, 1998) se indica una configuración de pendiente máxima $S_{max} = 3$ y $S'_{max} = 4$ (o $s34$ en las tablas), utilizando una ventana de $40ms$ y una $fs = 8$KHz, mientras en (Hosom, 2005) se selecciona $S_{max} = 2$, con un ritmo de ventaneo de $1ms$ para su algoritmo de programación dinámica (Viterbi, 1967). Del mismo modo, y siguiendo un criterio idéntico, se ha optado por una configuración sXY con $Y > X$ en los experimentos, donde sXY corresponde al valor de pendiente máxima para la primera fase (X= S_{max}) y segunda fase (Y= S'_{max}) de la aplicación del algoritmo de programación dinámica restringido en las tablas de resultados.

Las pruebas que se han llevado a cabo son:

1. Estudio del rendimiento de PMFA respecto a RAPT, YIN y SHRp mediante un barrido de los valores de S_{max} sobre el corpus de voz publicitario, utilizando las medidas GER y GPMER.

2. Validación de los resultados del PMFA en términos de GER, respecto a RAPT, YIN y SHRp sobre el corpus *Keele*.

En todos los experimentos se pretende comparar el funcionamiento del algoritmo de referencia (p.ej. RAPT) respecto al resultado obtenido de aplicar la propuesta de PMFA trabajando como postprocesamiento de éste mismo (p.ej. RAPT+PMFAsXY), como se muestra en las tablas de resultados presentadas a continuación.

Tabla 4.2: GER % sobre el corpus publicitario. En cursiva, los valores peores que los de referencia, y en negrita el mejor resultado en cada barrido.

MÉTODO	*Alegre*		*Neutro*		*Sensual*	
Ventana	$5ms$	$10ms$	$5ms$	$10ms$	$5ms$	$10ms$
RAPT	10.88		10.91		31.07	
RAPT + PMFAs13	*16.20*	*28.63*	*11.01*	*22.33*	24.05	*33.27*
RAPT + PMFAs24	8.88	*15.37*	7.28	10.51	21.06	23.48
RAPT + PMFAs34	7.61	10.42	6.61	7.98	20.64	21.65
RAPT + PMFAs26	9.17	*15.65*	7.32	10.47	20.88	23.23
RAPT + PMFAs37	**7.54**	10.83	**6.08**	8.01	19.93	21.25
RAPT + PMFAs48	7.88	9.17	6.59	7.43	20.21	20.79
RAPT + PMFAs68	7.76	**8.22**	6.21	6.76	**19.90**	**20.12**
RAPT + PMFAs79	7.87	8.24	6.19	**6.74**	20.02	20.20
RAPT + PMFAs912	8.59	8.89	6.38	6.84	20.46	20.35
YIN	17.44		22.35		36.86	
YIN + PMFAs13	16.82	*28.59*	12.10	*24.37*	23.18	32.79
YIN + PMFAs24	9.56	15.82	7.73	11.42	20.41	22.58
YIN + PMFAs34	**8.43**	11.16	7.14	8.39	20.26	21.03
YIN + PMFAs26	9.82	16.02	7.83	11.48	20.37	22.34
YIN + PMFAs37	8.61	11.50	**6.98**	8.60	20.10	20.70
YIN + PMFAs48	8.68	10.06	7.06	7.70	**20.09**	20.23
YIN + PMFAs68	8.75	**8.99**	7.14	**7.15**	20.16	**20.03**
YIN + PMFAs79	8.87	9.06	7.13	7.19	20.35	**20.03**
YIN + PMFAs912	9.60	9.73	7.44	7.55	20.76	20.54
SHRp	22.45		25.16		38.85	
SHRp + PMFAs13	18.28	*31.87*	12.94	24.98	25.11	35.24
SHRp + PMFAs24	9.49	16.87	8.64	12.02	**23.09**	24.47
SHRp + PMFAs34	**8.28**	11.30	**8.02**	9.09	23.11	22.97
SHRp + PMFAs26	9.73	17.11	8.63	12.05	23.16	24.47
SHRp + PMFAs37	8.87	11.71	8.15	9.17	23.82	23.11
SHRp + PMFAs48	8.67	9.62	**8.02**	8.20	23.30	22.93
SHRp + PMFAs68	8.85	**8.58**	8.16	**7.83**	23.69	**22.77**
SHRp + PMFAs79	9.21	8.73	8.32	7.86	24.04	22.89
SHRp + PMFAs912	10.37	9.73	8.91	8.35	24.91	23.76

Ajuste de PMFA sobre el corpus publicitario

En el primer experimento se analiza el funcionamiento de PMFA sobre el corpus publicitario con respecto a *(i)* distintas configuraciones del algoritmo (S_{max} y tamaño de ventana, ambos parámetros dependientes de la frecuencia de muestreo f_s utilizada), y *(ii)* los distintos estilos de locución presentes en el corpus (además del estilo neutro típicamente estudiado). Para ello se realiza un barrido de valores de S_{max} y S'_{max} para las dos configuraciones de análisis de las marcas de entrada consideradas, según la longitud de la ventana (L/f_s) de la fase de filtrado de errores: $5ms$ y $10ms$. Por el momento no se ha considerado solapar la información de las ventanas ($R = L$) —ver apartado 4.4.2.

La tabla 4.2 presenta los resultados obtenidos por los algoritmos estudiados sobre el corpus publicitario en términos de GER. Un análisis global de los resultados permite comprobar que PMFA mejora *claramente* los resultados obtenidos por los algoritmos de referencia, independientemente del PDA o PMA utilizado, el ventaneo o la configuración S_{max} usados —excluyendo valores de S_{max} como $s13$ o $s26$— y el estilo de locución estudiado. Por lo tanto, estos resultados confirman, por un lado, la viabilidad de la propuesta y por otro, su *fácil* configurabilidad. Es decir, excepto para algunos casos concretos, para cualquier configuración utilizada se consigue mejorar los resultados obtenidos por cualquiera de los métodos de referencia, con una mejora relevante en la gran mayoría de los casos.

No obstante, un análisis más detallado de los resultados permite observar como la configuración $5ms$ consigue, en general, unos mejores resultados que la el análisis con ventanas de $10ms$. Por lo tanto, un análisis más detallado de los datos permite mejorar el rendimiento de PMFA en términos de GER (así como se verá también para GPMER), aunque esto comporte un aumento del coste computacional del algoritmo —cuestión que en el contexto de la construcción de corpus para sistemas de conversión de texto en habla no es excesivamente crítica, ya que se trata de un proceso *off-line*. En cuanto a las configuraciones de S_{max}, la mayoría de valores de sXY presentan buenos resultados, mientras que las configuraciones $s13$ y $s26$ presentan, en general, los peores resultados (llegando a empeorar los valores de GER obtenidos con los métodos de referencia). Concretamente, $s13$ resulta ser una configuración demasiado restrictiva para poder seguir las variaciones de T_0 trama a trama (se dejan de observar valores locales de T_0 importantes, como sucedía en el ejemplo de la figura 4.3 que trabajaba con $S_{max} = 1$ respecto a utilizar $S_{max} = 2$ en la figura 4.4), mientras que $s26$ es una configuración descompensada (mal balanceada, con un ratio de 3 entre las dos pasadas —el mayor de todas las configuraciones). El resto de configuraciones presentan valores bastante cercanos, siendo las configuraciones $s34$ y $s37$ las mejores para ventanas de $5ms$, mientras $s68$ es el mejor para ventana de $10ms$ a lo largo de las pruebas, debido a que se dobla el paso entre ventanas.

Por otro lado, si se analiza el comportamiento de PMFA sobre los distintos estilos de locución considerados, se puede comprobar como las mayores tasas de error se obtienen para el subcorpus sensual, fundamentalmente por la dificultad de marcado que este estilo implica (presencia de suspiros, voz temblorosa,...), seguido por el subcorpus alegre, como resultado de su elevada F_0 media y desviación (mayores fluctuaciones de tono, períodos de señal muy próximos,...), y finalmente, se encuentra el subcorpus neutro, con las menores

Tabla 4.3: Mejoras relativas de GER (%) (mínima, media y máxima) conseguidas al aplicar PMFA sobre el corpus publicitario para el barrido de S_{max} estudiado, excluyendo la configuración $s13$, calculadas según la ecuación (4.15).

Mejora (%)	*Alegre*			*Neutro*			*Sensual*		
Ventana $5ms$	Min	Media	Max	Min	Media	Max	Min	Media	Max
RAPT+PMFA	15.72	24.97	30.71	32.88	39.66	44.28	32.79	34.36	35.93
YIN+PMFA	43.71	48.17	51.69	64.98	**67.31**	68.75	43.68	**48.89**	45.49
SHRp+PMFA	53.80	**59.09**	63.11	64.60	66.79	68.14	35.88	39.15	40.51
Ventana $10ms$	Min	Media	Max	Min	Media	Max	Min	Media	Max
RAPT+PMFA	0.48	14.55*	24.40	3.62	25.81	38.22	24.41	31.16	35.22
YIN+PMFA	8.16	33.82	48.44	48.64	61.14	68.01	38.73	**43.20**	45.66
SHRp+PMFA	23.80	**47.87**	61.79	52.12	**62.96**	68.87	37.01	39.71	41.38

*Sin considerar ni $s24$ ni $s26$, ya que empeoran los resultados de referencia.

tasas GER (es decir, es el más sencillo de etiquetar por sus características prosódicas). De todos modos, cabe destacar que las mejoras relativas conseguidas por PMFA sobre el subcorpus sensual (calculadas según la ecuación 4.15) son importantes y se asemejan a las obtenidas sobre el corpus alegre, demostrando así la versatilidad y robustez de la propuesta —aunque las máximas reducciones de GER se consiguen sobre el corpus neutro (ver tabla 4.3, de donde se excluyen las configuraciones que provocan $\Delta\text{GER}(\,\%) < 0$ (ecuación 4.15) para evitar sesgos de los resultados (son configuraciones extremas, que no presentan buen comportamiento, como se discute a lo largo de estos experimentos).

$$\Delta\text{GER}(\,\%) = \frac{\text{GER}_{ALG+PMFA} - \text{GER}_{ALG}}{\text{GER}_{ALG}} \cdot 100 \tag{4.15}$$

donde $\text{GER}_{ALG+PMFA}$ representa el GER conseguido por PMFA junto a uno de los algoritmos (ALG) de referencia, mientras GER_{ALG} denota el GER del algoritmo de referencia. Por lo que $\Delta\text{GER}(\,\%)$ representa la mejora relativa conseguida al aplicar PMFA sobre el algoritmo de referencia.

Finalmente, de la tabla 4.2 se puede observar que los mejores resultados (en términos absolutos) se obtienen mediante la combinación de RAPT+PMFA, con una mejora del orden de un 1 % en los subcorpus alegre y neutro respecto a los mejores resultados obtenidos por YIN+PMFA y SHRp+PMFA (p.ej. para $5ms$, PMFA con RAPT llega a un GER de 7.54 %, mientras que con YIN es de 8.43 % y con SHRp de 8.28 %, para distintas combinaciones de sXY), mientras que para el subcorpus sensual presenta unos resultados similares a YIN+PMFA y algo mejores que SHRp+PMFA (valores alrededor de GER = 20 % *vs.* valores alrededor de GER = 23 %). Sin embargo, PMFA no sólo mejora los resultados de RAPT, sino que también mejora considerablemente los resultados obtenidos por YIN y SHRp. Por ejemplo, las mayores reducciones relativas de GER se consiguen cuando PMFA utiliza SHRp o YIN (ver tabla 4.3).

Tabla 4.4: GPMER % sobre el corpus publicitario. En cursiva, los valores peores que los de referencia, y en negrita el mejor resultado en cada barrido.

MÉTODO	*Alegre*		*Neutro*		*Sensual*	
Ventana	5ms	10ms	5ms	10ms	5ms	10ms
RAPT	10.37		7.86		29.26	
RAPT + PMFAs13	*15.47*	*34.00*	6.86	*20.67*	13.76	24.58
RAPT + PMFAs24	6.04	*15.70*	2.86	7.08	10.93	14.28
RAPT + PMFAs26	6.49	*16.27*	3.04	7.46	11.22	14.71
RAPT + PMFAs34	**4.88**	8.64	2.30	3.72	10.37	12.45
RAPT + PMFAs37	5.39	9.55	**2.19**	4.06	**9.66**	12.99
RAPT + PMFAs48	5.89	7.56	2.43	3.24	10.28	12.55
RAPT + PMFAs68	5.88	**6.59**	2.28	**2.62**	9.81	**12.09**
RAPT + PMFAs79	6.54	7.09	2.32	2.64	9.83	12.16
RAPT + PMFAs912	8.95	9.41	2.61	2.95	10.18	12.57
YIN	8.06		5.16		22.06	
YIN + PMFAs13	*16.73*	*34.88*	*8.09*	*22.38*	13.97	*24.00*
YIN + PMFAs24	7.28	*17.07*	3.44	*8.03*	11.70	13.61
YIN + PMFAs26	7.91	*17.62*	3.68	*8.49*	12.06	13.97
YIN + PMFAs34	**6.21**	*10.04*	**2.87**	4.41	**11.36**	**12.05**
YIN + PMFAs37	7.61	*11.07*	3.07	4.83	12.33	12.50
YIN + PMFAs48	7.80	*9.25*	3.14	3.75	12.28	12.13
YIN + PMFAs68	8.04	***8.42***	3.20	**3.14**	12.49	12.14
YIN + PMFAs79	*8.71*	*8.95*	3.30	3.23	12.72	12.33
YIN + PMFAs912	*10.85*	*11.21*	3.75	3.80	13.44	13.09
SHRp	12.25		8.86		25.55	
SHRp + PMFAs13	*16.72*	*36.28*	8.28	*22.92*	15.97	*27.34*
SHRp + PMFAs24	5.96	*16.62*	4.00	8.17	**14.79**	15.48
SHRp + PMFAs26	6.57	*17.17*	4.27	8.68	15.20	16.02
SHRp + PMFAs34	**4.69**	8.83	**3.46**	4.63	14.98	15.20
SHRp + PMFAs37	6.68	9.86	4.04	5.12	16.88	**14.92**
SHRp + PMFAs48	6.48	7.58	3.93	4.13	16.10	15.07
SHRp + PMFAs68	6.91	**6.43**	4.10	**3.71**	16.74	15.14
SHRp + PMFAs79	7.91	7.13	4.33	3.80	17.35	15.44
SHRp + PMFAs912	10.76	9.74	5.09	4.46	18.47	16.78

Análisis de PMFA mediante GPMER sobre el corpus publicitario

A continuación se presentan los resultados obtenidos mediante la medida de evaluación propuesta: GPMER (ver ecuación (4.14)). Como se ha comentado, esta medida permite comparar el funcionamiento de PMAs que utilizan criterios locales distintos para posicionar las marcas dentro de los periodos de la señal de voz (en este caso, RAPT y PMFA) sin

tener que alinear las marcas previamente. La tabla 4.4 muestra los resultados obtenidos en términos de GPMER para el mismo estudio presentado en la tabla 4.2 utilizando GER. Como se puede comprobar del análisis de los resultados, GPMER muestra un comportamiento muy similar al obtenido mediante GER, es decir, *(i)* PMFA mejora considerablemente los resultados obtenidos con los métodos de referencia a lo largo de la tabla, *(ii)* el análisis con ventanas de $5ms$ presenta mejores resultados ($s34$ and $s37$, como los mejores pares de S_{max}) que con la ventana de $10ms$ ($s68$ como el mejor par de S_{max}, aunque en el subcorpus sensual los resultados son muy parejos con $s34$), y *(iii)* los valores extremos de sXY presentan los peores resultados, empeorando los de referencia también en $s912$ (poca restricción), además de $s13$ y $s26$ también presentes en GER.

No obstante, en la comparación también se puede observar que los valores absolutos de GPMER (tabla 4.4) son menores que los de GER (tabla 4.2). El número de datos considerados en el cálculo de GPMER es mqyor que en GER, ya que se comparan los resultados a nivel de marcas de *pitch* ($K^r - 1$ valores de perdiodicidad realtiva p_r, siendo K^r el número de marcas de *pitch* de referencia), mientras que en GER se comparan los resultados de F_0 a nivel de trama (T valores comparados, siendo $T \ll K^r$, generalmente). Por este motivo, GPMER se puede calificar como un GER *fino*, ya que las comparaciones de periodicidad se realizan a nivel local en lugar de hacerlo a nivel de trama. De este modo, GPMER *(i)* minimiza la probabilidad de dejar de computar errores que quedan enmascarados, por ejemplo, cuando en una trama la primera parte está sobremarcada mientras la segunda está inframarcada, dando como resultado un valor medio (periodicidad de la trama) cercano al de referencia e inferior a la desviación considerada; y *(ii)* imputa el error de marcado cometido en su justa medida —por ejemplo, si en una zona de la señal (que correspondería a una trama del GER) sólo hay una marca mal posicionada, GPMER sólo computará un error localizado, si la periodicidad local debida a esa marca supera el umbral de error, a diferencia de GER que computará *toda* la trama como errónea (y en consecuencia todas las marcas que contiene habrán sido etiquetadas como erróneas). Finalmente, el hecho que los valores de GPMER sean inferiores a los de GER indica que el caso *ii)* contribuye más que el *i)* —dado que *i)* hace aumentar el valor de GPMER, mientras que *ii)* lo hace disminuir.

Validación del funcionamiento de PMFA sobre el corpus *Keele*

El segundo experimento consiste en validar los resultados obtenidos sobre el corpus publicitario mediante el análisis de las configuraciones óptimas de PMFA aplicadas a los mismos algoritmos de referencia, pero en este caso validando su funcionamiento sobre el corpus *Keele*. Para ello, los diez ficheros de voz (5 locutores masculinos y 5 locutoras femeninas) se analizan con una ventana de $25.6ms$ y con un paso de $10ms$, para disponer de un ritmo de estimación de F_0 equivalente al que acompaña el corpus de referencia[28]. Asimis-

[28]No obstante, después de analizar los resultados, se optó por aumentar el tamaño de la ventana de análisis de los algoritmos de referencia (RAPT, YIN y SHRp) sobre el locutor M1 hasta los $40ms$, debido a que su F_0 media ronda los 50Hz, por lo que los algoritmos de PMA y PDA no disponían de suficiente información para estimar la periodicidad de las tramas. A pesar de ello, los resultados conseguidos por PMFA sobre este fichero de voz son los peores, debido a que se mantuvo el análisis con ventana de $5ms$ en PMFA, para ser

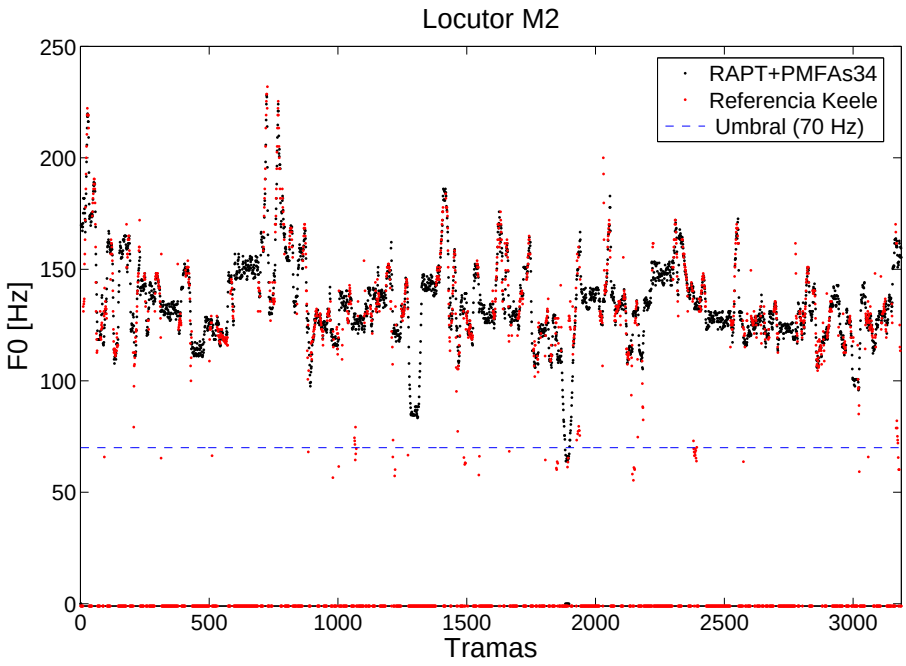


Figura 4.11: Distribución de los valores de F_0 por trama para el locutor masculino M2. La línea discontinua muestra el umbral de 70Hz considerado.

mo, los datos de periodicidad indicados en el corpus —en muestras— se han adaptado a $f_s = 16$KHz, una vez remuestreada la señal de voz de 20KHz a 16KHz para poder comparar de forma consistente los resultados obtenidos por PMFA en el corpus publicitario.

Además, después de revisar los valores de F_0 para los diez ficheros de voz, resulta evidente la existencia de valores erróneos (espurios) dentro del grupo de valores teóricamente revisados. Algunos de ellos son claramente inferiores o superiores a los valores de F_0 de las tramas vecinas, otros aparecen en medio de grupos de tramas sordas, etc. (ver los ejemplos de las figuras 4.11 y 4.12). Si se utilizaran estos valores para la comparación, implicaría cometer un error injusto en la validación del algoritmo analizado. Es por eso que resulta necesario eliminar estos valores espurios de la señal de referencia de F_0. En este caso, se ha seguido lo indicado en (Kasi y Zahorian, 2002), donde se eliminaron, después de una inspección visual de las curvas de *pitch*, los valores inferiores a 70Hz (ver figura 4.11), para las voces masculinas y los inferiores a 110Hz (en este trabajo se ha escogido 120Hz por parecer un valor más adecuado), para las femeninas (ver figura 4.12). En el trabajo de Kasi y Zahorian (2002) se asigna un valor de $F_0 = 0$ para todas estas tramas, es decir, no se consideran en la comparativa. Típicamente, la precisión de los PDAs sólo es evaluada en zonas etiquetadas como *claramente* sonoras, evitando tramas ambiguas (Bagshaw, 1994;

consistentes en las comparaciones, cuestión que provoca que, aproximadamente, sólo una de cada cuatro tramas de análisis contengan una estimación de la periodicidad debido al rango de periodicidades del locutor M1.

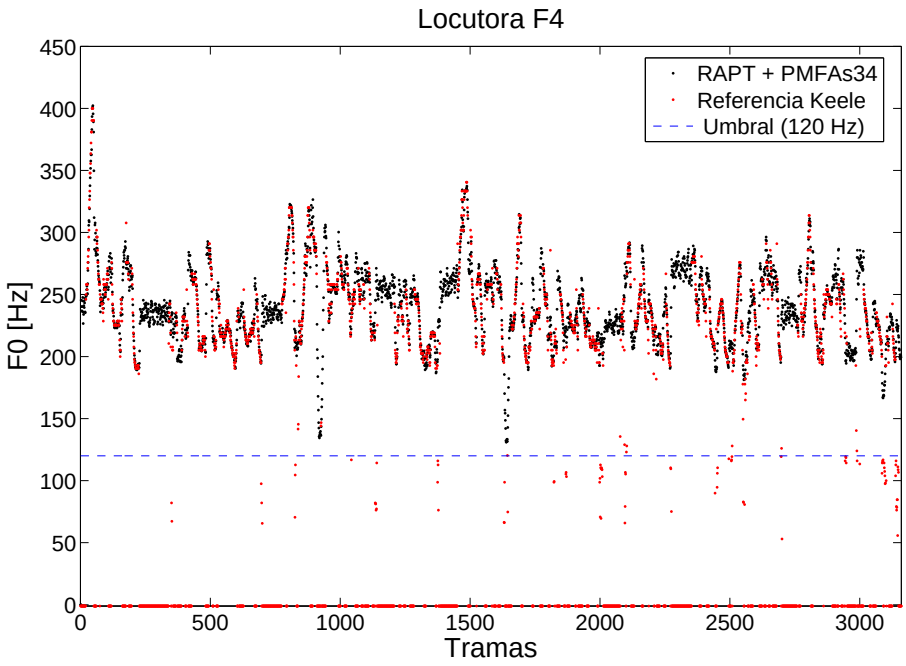


Figura 4.12: Distribución de los valores de F_0 por trama para la locutora femenina F4. La línea discontinua muestra el umbral de 120Hz considerado.

de Cheveigné y Kawahara, 2001; de Cheveigné y Kawahara, 2002; Sha, Burgoyne y Saul, 2004), que suelen ser computadas mediante la medida que evalúa la capacidad de acierto en la decisión de la sonoridad —VER o UER— (Sun, 2002). Sin embargo, en este trabajo estas tramas se tomarán en consideración. Concretamente, se pueden producir dos situaciones distintas en lo que se refiere al valor de F_0 entregado por el PDA o PMA evaluados en la trama donde la F_0 de referencia supera el umbral: *(i)* si éste se encuentra también fuera del umbral, se contabilizará como error (también es un valor espurio), en cambio, *(ii)* si éste se encuentra dentro del margen de valores válidos, esta trama no se contabilizará como errónea (es un valor teóricamente válido). De este modo, se pretende penalizar los valores que se encuentren fuera del margen de valores de F_0 considerados como válidos, sin sesgar los resultados hacia los valores que se encuentren dentro del umbral, ya que para esa trama no se dispone de un valor de F_0 fiable.

En las tablas 4.5 y 4.6 se presentan los resultados de GER obtenidos sobre el corpus *Keele* por la configuración PMFAs34 con ventaneo de $5ms$ sobre las marcas de pich obtenidas a partir de los algoritmos de referencia. Se utiliza el mismo rango de F_0 considerado en el primer estudio, es decir [50,550] Hz. En estas tablas se ha escogido la configuración $s34$ como representativa del barrido realizado, teniendo en cuenta los resultados del experimento sobre el corpus publicitario (en el Apéndice B.2 se presentan los resultados para todo el barrido de configuraciones estudiado de PMFA para el corpus *Keele*). Los resultados obtenidos demuestran, de nuevo, el buen funcionamiento de PMFA como sistema de

Tabla 4.5: GER % para los locutores masculinos (M1 a M5) del corpus *Keele* con PMFAs34 y ventana de $5ms$.

Método	*M1*	*M2*	*M3*	*M4*	*M5*	*Media*
RAPT	22.93	17.42	4.72	14.29	8.33	13.54
RAPT + PMFAs34	12.28	5.15	**0.89**	**2.47**	**3.10**	**4.78**
YIN	12.02	17.47	1.85	7.62	6.89	9.17
YIN + PMFAs34	**11.72**	**4.48**	**0.89**	2.60	6.19	5.18
SHRp	29.30	21.29	16.91	24.97	25.37	23.57
SHRp + PMFAs34	13.94	7.65	2.33	7.17	12.67	8.75

Tabla 4.6: GER % para las locutoras femeninas (F1 a F5) del corpus *Keele* con PMFAs34 y ventana de $5ms$.

Método	*F1*	*F2*	*F3*	*F4*	*F5*	*Media*
RAPT	6.62	4.29	5.44	7.68	2.01	5.21
RAPT + PMFAs34	**0.61**	**0.43**	**0.20**	**0.93**	0.44	**0.52**
YIN	3.72	1.07	1.88	4.21	0.38	2.25
YIN + PMFAs34	1.69	0.54	0.47	1.40	**0.27**	0.87
SHRp	10.85	6.53	10.56	20.71	8.15	11.36
SHRp + PMFAs34	0.88	0.86	0.95	4.38	1.14	1.64

Tabla 4.7: Mejoras relativas de GER (%) (mínima, media y máxima) conseguidas al aplicar PMFA sobre el corpus Keele para PMFAs34 y ventana de $5ms$, para las locutoras femeninas (F1 a F5) y los locutores masculinos (M1 a M5), calculadas según la ecuación (4.15).

Mejora (%)	*F1 a F5*			*M1 a M5*			*Global*
Método	Min	Media	Max	Min	Media	Max	Media
RAPT+PMFAs34	78.32	**89.97**	96.26	46.44	64.71	82.73	**71.71**
YIN+PMFAs34	28.57	61.21	74.75	2.45	43.54	74.33	47.03
SHRp+PMFAs34	78.87	85.54	91.86	52.43	**62.86**	86.21	70.24

filtrado y corrección de marcas, consiguiendo unas reducciones relativas promedio del 79 % y el 57 % de GER para las locutoras femeninas y los masculinos, respectivamente (ver tabla 4.7). Otra vez, la mejor combinación de algoritmos vuelve a ser RAPT+PMFA (con una mejora media de GER del 90 % y del 65 % de los resultados para las locutoras femeninas y los masculinos, respectivamente). No obstante, PMFA+YIN y PMFA+SHRp también presentan buenos resultados en las tablas 4.5 y 4.6, consiguiendo PMFA+YIN algunos de los mejores resultados absolutos de GER, mientras PMFA+SHRp presenta algunas de las mayores reducciones relativas de GER (ver tabla 4.7).

Por lo tanto, este experimento permite contrastar los resultados obtenidos sobre el corpus publicitario, ya que se obtiene un comportamiento similar al estudio anterior con mejoras

importantes de los resultados obtenidos para los distintos locutores del corpus de referencia (ver también los resultados para otras configuraciones de sXY en el anexo B.2).

4.7. Discusión

En este capítulo del trabajo de investigación se ha presentado una propuesta de algoritmo genérico para el filtrado y posterior posicionamiento de las marcas de *pitch* obtenidas a partir de un PDA o un PMA de entrada cualquiera. Según los resultados obtenidos, el algoritmo propuesto basado en la programación dinámica restringida, denominado PMFA (*Pitch Marks Filtering Algorithm*), ha conseguido los mejores resultados en términos de GER y GPMER, de lo que se deduce que consigue una secuencia de marcas *pitch* más robusta y suave a partir de la secuencia de marcas de entrada subministrada. Los experimentos han mostrado el buen funcionamiento de la propuesta respecto a distintos algoritmos de detección y marcado de *pitch* de referencia sobre los corpus de voz y estilos de locución estudiados. A continuación, para cerrar el presente apartado, se presentan algunas reflexiones sobre la propuesta realizada, los datos analizados y los resultados obtenidos.

Medidas de evaluación

En este capítulo se ha introducido una nueva medida que permite evaluar el funcionamiento de cualquier PMA independientemente del criterio local de ubicación de marcas utilizado. Para ello, se compara la secuencia de marcas estimadas respecto la secuencia de marcas de referencia según sus periodicidades locales (diferencia entre la posición de marcas consecutivas), determinando la correspondencia de marcas estimadas y marcas de referencia como se ha descrito en la expresión (4.13). Estos resultados pueden ser analizados mediante un tercer parámetro: el porcentaje de omisiones e inserciones que se obtiene con los algoritmos de referencia y con el posterior filtrado mediante PMFA. Esta información puede ser útil como referencia de comparación de los resultados obtenidos utilizando GPMER y GER. En el presente trabajo, debido a que las marcas comparadas utilizan criterios locales distintos, no se puede calcular la tasa de inserciones+omisiones de forma estricta (debido a las problemáticas de alineamiento que se han comentado en la sección 4.5.1). No obstante, el número de inserciones y omisiones se puede estimar como subproducto del algoritmo que implementa el cálculo de GPMER según lo descrito en las ecuaciones (4.12) y (4.13), evitando tener que alinear previamente las marcas. Concretamente, el porcentaje de inserciones se puede obtener a partir del número de periodicidades de referencia a las que les ha sido asignada más de una periodicidad evaluada ($I(\,\%)$ en la ecuación (4.16)), mientras que el porcentaje de omisiones se pueden detectar contando el número de periodicidades de referencia a las que no les ha sido asignada ninguna periodicidad evaluada ($O(\,\%)$ la ver ecuación (4.17))[29].

[29]La tasa de inserciones+omisiones se calcula para las zonas de interés, es decir, no se tienen en cuenta las zonas que son sordas, según los datos del corpus de referencia o las marcas de segmentación del corpus desarrollado.

Tabla 4.8: Porcentaje de omisiones + inserciones sobre el corpus publicitario. En cursiva, los valores peores que los de referencia, y en negrita el mejor resultado en cada barrido.

MÉTODO	*Alegre*		*Neutro*		*Sensual*	
Ventana	$5ms$	$10ms$	$5ms$	$10ms$	$5ms$	$10ms$
RAPT	6.64		4.99		22.53	
RAPT + PMFAs13	*7.41*	*15.96*	3.44	9.18	7.53	12.76
RAPT + PMFAs24	3.28	*7.29*	1.49	3.27	5.89	7.52
RAPT + PMFAs26	3.51	*7.44*	1.48	3.21	5.89	7.40
RAPT + PMFAs34	**2.78**	4.34	1.18	1.88	5.54	6.62
RAPT + PMFAs37	3.14	4.64	**1.01**	1.85	**5.04**	6.53
RAPT + PMFAs48	3.27	3.79	1.15	1.48	5.30	6.30
RAPT + PMFAs68	3.27	**3.47**	1.08	**1.22**	5.15	6.09
RAPT + PMFAs79	3.47	3.60	1.07	**1.22**	5.15	**6.03**
RAPT + PMFAs912	4.14	4.30	1.18	1.30	5.28	6.14
YIN	5.53		4.85		17.42	
YIN + PMFAs13	*7.92*	*16.12*	4.05	*10.12*	7.41	12.06
YIN + PMFAs24	3.92	*7.88*	1.79	3.76	6.09	7.08
YIN + PMFAs26	4.13	*7.99*	1.82	3.70	6.60	7.03
YIN + PMFAs34	**3.52**	4.99	1.47	2.20	**5.89**	6.28
YIN + PMFAs37	4.06	5.27	**1.43**	2.19	5.96	6.21
YIN + PMFAs48	4.03	4.51	1.52	1.74	5.99	5.97
YIN + PMFAs68	4.16	**4.23**	1.49	**1.52**	6.01	**5.86**
YIN + PMFAs79	4.35	4.36	1.50	1.55	6.06	5.89
YIN + PMFAs912	4.98	5.07	1.63	1.65	6.25	6.07
SHRp	9.87		7.14		19.10	
SHRp + PMFAs13	8.26	*18.03*	4.42	*10.58*	9.33	14.80
SHRp + PMFAs24	3.59	7.98	2.36	4.05	**8.56**	8.73
SHRp + PMFAs26	3.79	8.12	2.33	3.99	9.24	8.66
SHRp + PMFAs34	**3.11**	4.68	**2.12**	2.58	8.90	8.22
SHRp + PMFAs37	3.86	4.93	2.20	2.55	9.74	**8.17**
SHRp + PMFAs48	3.64	3.98	2.14	2.11	9.13	**8.17**
SHRp + PMFAs68	3.84	**3.50**	2.24	**1.99**	9.54	8.25
SHRp + PMFAs79	4.19	3.74	2.27	2.03	9.80	8.47
SHRp + PMFAs912	5.10	4.51	2.60	2.20	10.11	8.85

Las ecuaciones que aproximan el cálculo de las tasas de inserción y omisión son:

$$I(\,\%) = \frac{\#\Big((\#k' \in \mathcal{K}'_k) - 1 \,\Big|k' > 1\Big)}{K^r} \cdot 100 \tag{4.16}$$

$$O(\,\%) = \frac{\#k \,|\mathcal{K}'_k = \emptyset}{K^r} \cdot 100 \tag{4.17}$$

donde # indica *número de* o cardinalidad y K^r representa el número de marcas de referencia.

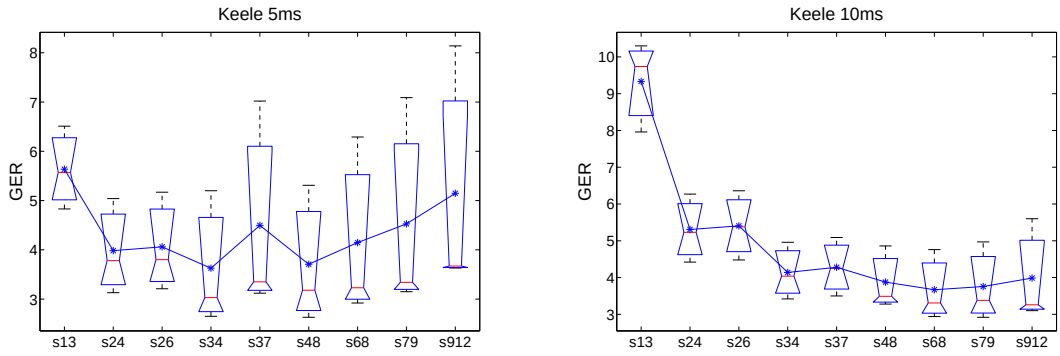
Si se compara la ordenación de valores obtenidos de GPMER y GER en relación con los estilos de locución estudiados, GPMER presenta unos valores que definen un patrón (Neutro < Alegre < Sensual) mejor correlado con la dificultad inherente del etiquetado de cada estilo, y más acorde con la distribución global de los errores (ver figura 4.17) y lo observado mediante la revisión experta de los resultados. Concretamente, si se toma en consideración la distribución de valores de la tasa de inserciones+omisiones como información de referencia (ver tabla 4.8), GPMER presenta una correlación de $\rho = 0.989$ respecto a la distribución de valores de inserciones+omisiones una vez aplicado PMFA (incluyendo las dos configuraciones de análisis para todas las parejas de S_{max} estudiadas). Sin embargo, esta correlación se reduce hasta $\rho = 0.909$ cuando se compara GER con la tasa de inserciones+omisiones. Caso a parte, es la relación de estas tres tablas en lo que se refiere a los resultados obtenidos por los algoritmos de referencia (RAPT, YIN y SHRp), donde, como se ha comentado, existe una correlación menos clara entre GER y GPMER ($\rho = 0.781$), con una correlación de GER y GPMER con el porcentaje de las inserciones+omisiones de $\rho = 0.827$ y $\rho = 0.994$, respectivamente. Por lo tanto, se puede concluir que GPMER es una medida que refleja mejor el funcionamiento del PMA analizado, en lugar de analizarlo mediante GER, que evalúa su comportamiento a nivel de trama, con la pérdida de detalle que esto puede conllevar. Asimismo, GPMER da información local de la fiabilidad del PMA sin necesitar de la previa alineación de las marcas como pasa en el cálculo estricto de la tasa de inserciones y omisiones.

Configurabilidad del PMFA

Una vez presentado el algoritmo de PMFA, queda claro que una de las características principales de la aproximación es su simplicidad: *(i)* no se utiliza ninguna función de coste compleja —sólo es necesario ajustar los valores de pendiente máxima (además del tradicional margen de frecuencias fundamentales ya comentado)—, *(ii)* se filtran los errores del PDA y el PMA mediante un esquema de elección de los valores candidatos de periodicidad por trama mediante una métrica entera (acumulación de presencia de periodicidades)[30], *(iii)* no realiza un seguimiento estricto de la periodicidad de las tramas de entrada, en la línea de (Veldhuis, 2000), que da menor importancia al valor de F_0 del PDA al calcular la autocorrelación de los periodos —pero a cambio de aumentar el coste computacional del proceso (Colotte y Laprie, 2002)—, y *(iv)* no incorpora ningún proceso de análisis de la sonoridad, evitando arrastrar los errores de estimación de la sonoridad de las tramas al proceso de marcado —en este caso, la periodicidad de las zonas sordas de la señal, habitualmente con ausencia de marcas, se convierte en una transición entre las zonas sonoras vecinas (ver ejemplo de la figura 4.6). En cuanto al valor de la variación máxima intertrama permitida (S_{max} y S'_{max}) para el algoritmo de programación dinámica implementado, al tratarse de un valor en muestras, éste dependerá de la frecuencia de muestreo utilizada (Goncharoff y Gries, 1998; Alías y Iriondo, 2001a; Hosom, 2005) y de la variabilidad prosódica (estilo de locución) del corpus analizado, como se ha observado en los experimentos desarrollados.

[30] En un futuro, se pretende estudiar métricas más sofisticadas con el objetivo de puede mejorar el funcionamiento del algoritmo desarrollado.

Los resultados obtenidos por PMFA mediante las distintas configuraciones de la restricción de pendiente máxima S_{max} permitida en el algoritmo de programación dinámica aplicado, permiten concluir que, prácticamente, para cualquier configuración de S_{max} (excluyendo los valores extremos y la configuración de ventana de $10ms$ para PMFA+YIN sobre el subcorpus alegre analizada según GPMER), PMFA consigue mejorar de forma significativa los resultados obtenidos por los algoritmos de entrada utilizados como referencia. En las figuras 4.13, 4.14, 4.15 y 4.16 se presenta el comportamiento global de las distintas configuraciones PMFAsXY (donde X indica el valor de S_{max} para la primera fase e Y el valor de S'_{max} para la segunda fase del algoritmo de programación dinámica restringido) agrupando todos los datos disponibles de medidas de evaluación (GER, GPMER y tasa de omisiones más inserciones), para los análisis con ventana de $5ms$ y $10ms$ (donde las configuraciones óptimas de sXY son distintas). Aunque se trata de representaciones de distribuciones de datos heterogéneos, todas ellas representan tasas de error y permiten tener una visión global del comportamiento de las distintas configuraciones de PMFA para las distintas pruebas (corpus y estilos de locución) realizadas.

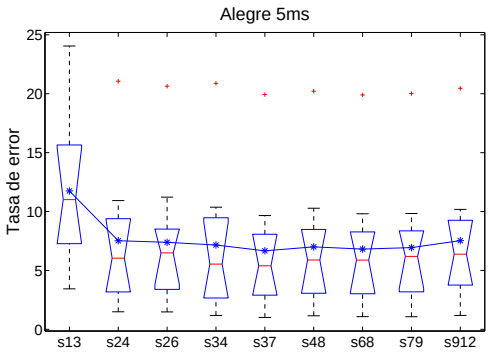


(a) Análisis con ventana de $5ms$ para todos los locutores.

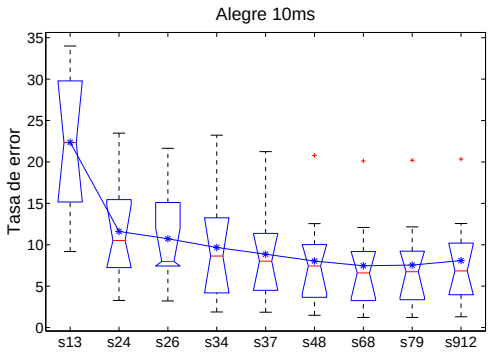
(b) Análisis con ventana de $10ms$ para todos los locutores.

Figura 4.13: Distribución de los valores de GER obtenidos a lo largo de las pruebas (RAPT+PMFA, YIN+PMFA y SHRp+PMFA) para las distintas configuraciones sXY de PMFA sobre el corpus *Keele*. La línea -∗- indica la media de las distribuciones.

De forma general, se puede observar como la configuración $s34$ y la configuración $s68$ son las que presentan unas menores tasas de error para los análisis con ventana de $5ms$ y $10ms$, respectivamente a lo largo de las figuras (con alguna excepción, por ejemplo, parece que la configuración $s37$ es algo mejor que $s34$ para el corpus alegre con análisis cada $5ms$). Sin embargo, encontrar la configuración *óptima* para un determinado locutor o un estilo de locución determinados puede depender de varios factores, entre ellos: el rango de frecuencias de esa voz, la variabilidad del tono de la locución, la configuración de análisis utilizada, la bondad del algoritmo de entrada, etc. Por ejemplo, aunque globalmente parece que la configuración PMFAs34 es la óptima cuando se utiliza una ventana de $5ms$, la tabla

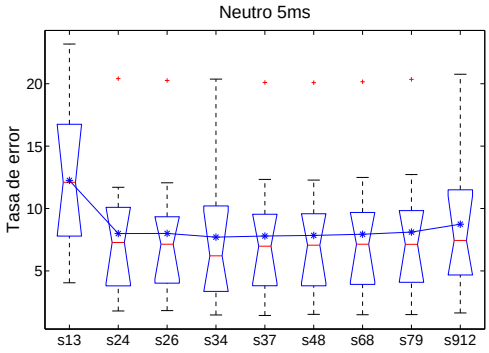


(a) Análisis con ventana de $5ms$ para el estilo alegre.

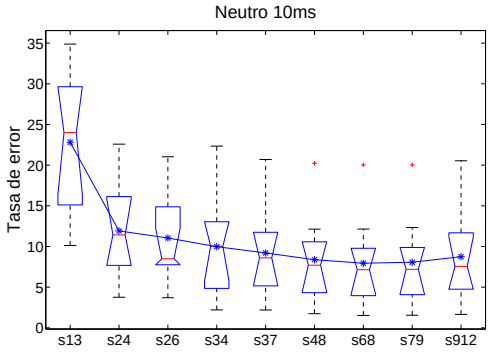


(b) Análisis con ventana de $10ms$ para el estilo alegre.

Figura 4.14: Distribución de las tasas de error (incorpora GER, GPMER y Omisiones + Inserciones) obtenidas a lo largo de las pruebas (RAPT+PMFA, YIN+PMFA y SHRp+PMFA) para las distintas configuraciones sXY de PMFA sobre el estilo alegre del corpus publicitario. La línea -$*$- indica la media de las distribuciones.



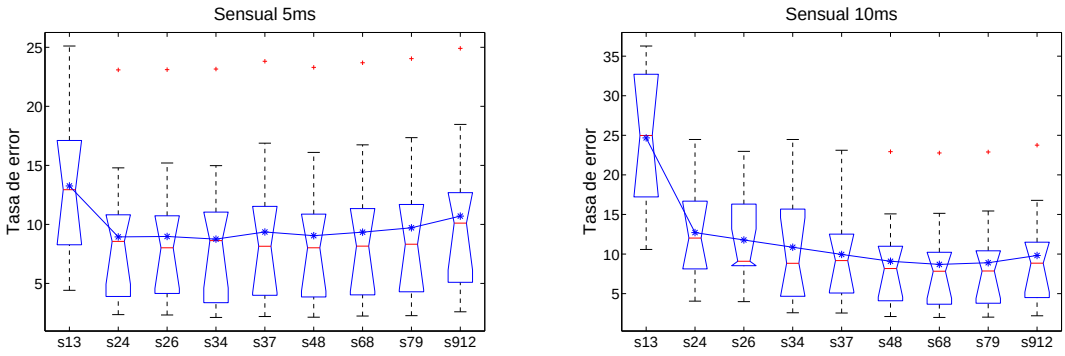
(a) Análisis con ventana de $5ms$ para el estilo neutro.



(b) Análisis con ventana de $10ms$ para el estilo neutro.

Figura 4.15: Distribución de las tasas de error (incorpora GER, GPMER y Omisiones + Inserciones) obtenidas a lo largo de las pruebas (RAPT+PMFA, YIN+PMFA y SHRp+PMFA) para las distintas configuraciones sXY de PMFA sobre el estilo neutro del corpus publicitario. La línea -$*$- indica la media de las distribuciones.

4.4 (GPMER) muestra como la configuración $s37$ presenta resultados algo mejores para los estilos neutro y sensual cuando RAPT es el algoritmo de referencia utilizado (aunque los resultados conseguidos con $s34$ son bastante parecidos). Por otro lado, para un paso de



(a) Análisis con ventana de $5ms$ para el estilo sensual.

(b) Análisis con ventana de $10ms$ para el estilo sensual.

Figura 4.16: Distribución de las tasas de error (incorpora GER, GPMER y Omisiones + Inserciones) obtenidas a lo largo de las pruebas (RAPT+PMFA, YIN+PMFA y SHRp+PMFA) para las distintas configuraciones sXY de PMFA sobre el estilo sensual del corpus publicitario. La línea -∗- indica la media de las distribuciones.

ventana de $10ms$, aunque se llegue a unos resultados algo peores, éstos son más estables, siendo $s68$ la mejor configuración, según las tablas de resultados y las figuras que se acaban de presentar. Finalmente, recordar que el incremento del valor de S_{max} provoca un aumento del coste computacional del PMFA, por lo que si se obtienen resultados parecidos con configuraciones menores de S_{max} y S'_{max} se acelerará el proceso de reetiquetado del corpus. De todos modos, como se ha comentado, este no es un factor crítico del algoritmo, ya que el etiquetado del corpus para conversión de texto en habla es un proceso *off-line*, por lo que el coste computacional no ha sido evaluado exhaustivamente en este capítulo.

En cuanto al criterio local, en este trabajo se ha escogido colocar las marcas de *pitch* siguiendo el máximo de amplitud de la señal en valor absoluto dentro de cada periodo, como en otros trabajos, p.ej. (Goncharoff y Gries, 1998; Veldhuis, 2000). No obstante, como se ha comentado a lo largo de este capítulo, el algoritmo desarrollado permite escoger otros criterios locales e incluso incorporar procesos que permitan decidir qué criterio local utilizar, como en los trabajos de (Lin y Jang, 2004) y (Chen y Kao, 2001), donde primero se escoge si colocar las marcas de *pitch* sobre los máximos o los mínimos de la señal y, a continuación, se ajustan las marcas al criterio local considerado.

Análisis del rendimiento del PMFA

Una de las particularidades del análisis de los experimentos, evaluados mediante GER y GPMER, es que se comparan todas las tramas de la señal de voz que pertenecen a fonemas no sordos (según las marcas de segmentación), incluidas las que son ambiguas en términos de sonoridad, a diferencia de otros trabajos que dejan estas tramas fuera de la comparativa

o bien las imputan en el cálculo de VER o UER. El motivo es doble, por un lado, se pretende obtener un resultado global del algoritmo analizado y, por otro, se busca analizar el comportamiento de los algoritmos sobre las tramas más complicadas de marcar. De este modo, se puede comprobar la capacidad de filtrado de errores de PMFA para este tipo de tramas, evitando sesgar los resultados al comparar sólo tramas claramente sonoras (Sun, 2002). No obstante, este análisis hace que los resultados no sean comparables exactamente (tasas de error mayores) con los presentados en la literatura para el corpus de referencia utilizado (de Cheveigné y Kawahara, 2002; Kasi y Zahorian, 2002; Dikshit, Zahorian y Nagulapati, S., 2005; Sun, 2002; Achan et al., 2005; Sha, Burgoyne y Saul, 2004; Sha y Saul, 2005; Bánhalmi et al., 2005).

Por otro lado, de los resultados obtenidos, se puede comprobar como el método PMFA consigue homogeneizar el rendimiento de los algoritmos utilizados como entrada —RAPT, YIN y SHRp—, tanto en términos de GER como de GPMER, para los dos corpus estudiados. Por ejemplo, para el estilo neutro del corpus publicitario, RAPT presenta un GER de 10.91 %, que mediante RAPT+PMFAs34 se convierte en 6.61 %, YIN parte de 22.35 % que con YIN+PMFAs34 pasa a ser de 7.14 %, y finalmente, SHRp pasa de 25.16 % a 8.02 %, para la misma configuración de PMFA. Asimismo, en términos de GPMER, estos valores son: para RAPT, 7.86 %, mientras RAPT+PMFAs34 consigue 2.30 %, para YIN, 5.16 %, mientras YIN+PMFAs34 llega a 2.87 % y, finalmente, PMFA parte de 8.86 % para SHRp y consigue un GPMER de 3.46 % mediante PMFAs34. Ejemplos parecidos pueden encontrarse a lo largo de las tablas presentadas en este capítulo así como en el apéndice B.2. Por lo tanto, se puede concluir que PMFA es capaz de conseguir tasas de error similares (en orden de magnitud), partiendo de tasas de error distintas, al conseguir mejoras relativas (calculadas según la ecuación (4.15)) mayores en los casos de mayor error, como se ha mostrado también a lo largo de los experimentos.

Fiabilidad estadística de los resultados: Como se ha comentado al describir los corpus utilizados para evaluar PMAs, en este trabajo se ha optado por trabajar con un corpus publicitario de dimensión considerable (más de 2.5h), dividido en tres subcorpus de unos 45 min para cada uno de los estilos analizados, a diferencia de los habitualmente utilizados para evaluar PMAs, los cuales no suelen pasar de los 10 minutos de duración. Aunque el trabajo previo de revisión de las marcas ha sido muy costoso, esto ha permitido obtener unos resultados estadísticamente mucho más fiables que los obtenidos por trabajos similares, o mediante *Keele*, por ejemplo. De este modo, se pueden validar los resultados obtenidos en términos del análisis de la varianza (en inglés, *Analysis of Variance* o ANOVA) de los mismos. Los resultados de PMFA obtenidos mediante el análisis con ventaneo de $5ms$ son significativamente mejores que los obtenidos con ventana de $10ms$ para el corpus publicitario, tanto para GER ($F(1,160) = 5.2, p < 0.0239$), GPMER ($F(1,160) = 12.43, p < 0.0006$), como para el porcentaje de omisiones e inserciones ($F(1,160) = 7.97, p < 0.0054$). Sin embargo, esto no es así para el corpus *Keele*, donde la mejora general conseguida por la configuración de $5ms$ respecto a $10ms$ (ver tabla B.1) no es estadísticamente fiable (por ejemplo, los mejores resultados de GER para PMFA+SHRp se consiguen con la configuración de $10ms$). Esto se debe, fundamentalmente al menor número de datos disponibles (duración muy reducida) y

al mayor número de locutores contemplados en ese corpus. No obstante, no hay que olvidar que PMFA consigue mejorar, en la gran mayoría de los casos, los resultados obtenidos por los PDA y PMA de referencia utilizados.

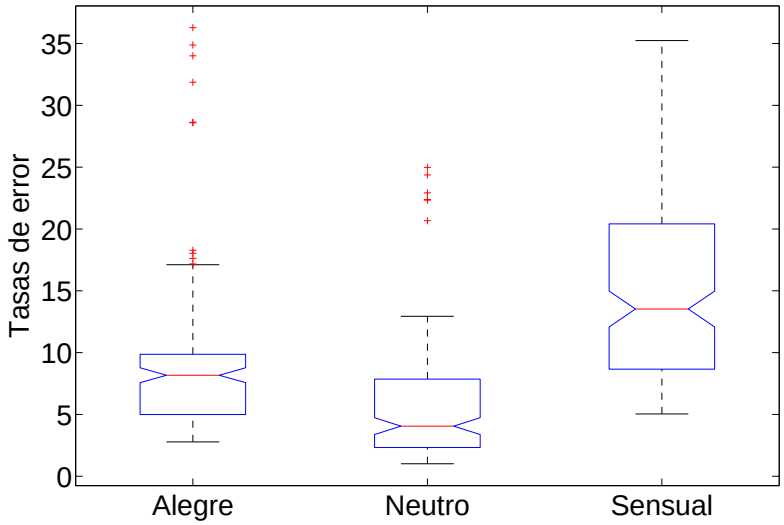


Figura 4.17: Distribución de las tasas de error (GER, GPMER y Omisiones+Inserciones) obtenidas a lo largo de las pruebas para los tres estilos de locución del corpus publicitario.

Estilos de locución

En general, los trabajos que han estudiado del funcionamiento de los algoritmos de detección de la frecuencia fundamental suelen trabajar con señales de voz de estilo neutro. No obstante, en estos trabajos no siempre se utilizan señales de calidad de estudio profesional, encontrándose en la literatura trabajos que analizan señales de calidad menor, por ejemplo, calidad telefónica (p.ej. (Wang y Seneff, 2000; Kasi y Zahorian, 2002), o señales degradadas (Kounoudes, Naylor y Brookes, 2002; Prasanna y Yegnanarayana, 2004), etc. En el camino iniciado en este trabajo de investigación hacia la síntesis multidominio (ver capítulo 3), ha sido necesario disponer de un sistema de etiquetado robusto de la periodicidad para corpus de voz de gran tamaño que presenten distintas características de la señal grabada (diversidad de calidades vocales). Es por ello que en este capítulo se ha estudiado el impacto en el funcionamiento de los PDA y los PMA de referencia (y el PMFA desarrollado) respecto a las distintas características de las señales de voz asociadas a distintos estilos de locución —en este caso, alegre, neutro y sensual—, en la línea de otros trabajos enfocados a evaluar procesos de etiquetado automático sobre corpus de estilos de locución distintos (Saito y Sakamoto, 2005). La figura 4.17 presenta las tasas de error obtenidas para cada uno de los estilos considerados, quedando demostrado la mayor dificultad de etiquetado del estilo sensual y el alegre respecto al neutro. Concretamente, los resultados demuestran

Tabla 4.9: Tasas de error sobre el corpus alegre para cuatro ventanas de análisis distintas. En cursiva, los valores peores que los de referencia, y en negrita el mejor resultado en cada barrido.

MÉTODO	GER (%)				GPMER (%)			
Ventana	$2.5ms$	$5ms$	$10ms$	$15ms$	$2.5ms$	$5ms$	$10ms$	$15ms$
RAPT	10.88				10.37			
RAPT+PMFAs13	9.10	*16.20*	*28.63*	*38.27*	5.73	*15.47*	*34.00*	*44.87*
RAPT+PMFAs24	7.38	8.88	*15.37*	*21.37*	**4.75**	6.04	*15.70*	*25.80*
RAPT+PMFAs26	7.73	9.17	*15.65*	*21.49*	5.06	6.49	*16.27*	*26.15*
RAPT+PMFAs34	**7.29**	7.61	10.42	*14.82*	4.78	**4.88**	8.64	*15.68*
RAPT+PMFAs37	7.71	**7.54**	10.83	*15.04*	5.31	5.39	9.55	*16.50*
RAPT+PMFAs48	7.81	7.88	9.17	*11.85*	5.85	5.89	7.56	*11.91*
RAPT+PMFAs68	7.76	7.76	**8.22**	9.22	5.90	5.88	**6.59**	**8.10**
RAPT+PMFAs79	8.04	7.87	8.24	**8.96**	6.62	6.54	7.09	8.20
RAPT+PMFAs912	8.66	8.59	8.89	9.47	9.13	8.95	9.41	10.12

que el estilo neutro es el más sencillo de etiquetar, seguido del alegre y, a mayor distancia, se encuentra el estilo sensual. Estos resultados son estadísticamente fiables en términos de ANOVA: Neutro $<$ Alegre con $F(1,322) = 36.45, p < 4.3210e^{-9}$, y Alegre $<$ Sensual con $F(1,322) = 57.59, p < 3.5027e^{-13}$.

Ventanas de análisis

Uno de los elementos clave de la propuesta es el tamaño de ventana de análisis de la secuencia de marcas de *pitch* iniciales para la estimación robusta de la periodicidad de la señal de voz trama a trama de la primera fase del PMFA. En los experimentos desarrollados, se han utilizado dos tamaños de ventana distintos: $5ms$ y $10ms$. Considerando $F_0 \in [50, 550]$ ($T_0^{50\text{Hz}} = 20ms$ y $T_0^{550\text{Hz}} = 1.8ms$), cada trama de análisis contiene entre 0 ($5ms/20ms = 0.25$) y 3 marcas ($5ms/1.8ms = 2.75$) o entre 0 ($10ms/20ms = 0.5$) y 6 marcas ($10ms/1.8ms = 5.55$), respectivamente. De los experimentos, se ha podido constatar que utilizando la ventana de $5ms$ se obtienen, en general, los mejores resultados en términos absolutos de la tasa de error (tanto en GER, GPMER como en la tasa de inserciones+omisiones) utilizando $s34$ como configuraciones de S_{max} y S'_{max} para las dos fases del algoritmo de programación dinámica (sXY).

A continuación, se presenta un nuevo experimento con el objetivo de obtener una visión más general de la relación del tamaño de la ventana con la configuración de los valores de pendiente máxima sXY de las dos fases en las que se aplica el algoritmo de programación dinámica restringido. Para ello, se estudia el rendimiento de PMFAsXY para dos nuevos tamaños de ventana de análisis: $2.5ms$ y $15ms$. El aumento del tamaño de la ventana (ventana de $15ms$) proporcionará mayor redundancia en el análisis de la periodicidad de cada trama, sobretodo a bajas frecuencias, mientras que la reducción del tamaño de la

Tabla 4.10: Tasas de error sobre el corpus neutro para cuatro ventanas de análisis distintas. En cursiva, los valores peores que los de referencia, y en negrita el mejor resultado en cada barrido.

MÉTODO	GER (%)				GPMER (%)			
Ventana	2.5ms	5ms	10ms	15ms	2.5ms	5ms	10ms	15ms
RAPT	10.91				7.86			
RAPT+PMFAs13	7.57	*11.01*	*22.33*	*32.41*	2.83	6.86	*20.67*	*32.77*
RAPT+PMFAs24	6.39	7.28	10.51	*15.54*	2.20	2.86	7.08	*13.59*
RAPT+PMFAs26	6.41	7.32	10.47	*15.41*	2.33	3.04	7.46	*14.28*
RAPT+PMFAs34	6.15	6.61	7.98	10.43	**2.09**	2.30	3.72	*6.97*
RAPT+PMFAs37	6.23	**6.08**	8.01	10.34	2.25	**2.19**	4.06	7.53
RAPT+PMFAs48	6.15	6.59	7.43	8.47	2.22	2.43	3.24	4.95
RAPT+PMFAs68	**6.01**	6.21	6.76	7.42	2.10	2.28	**2.62**	3.40
RAPT+PMFAs79	6.09	6.19	**6.74**	**7.19**	2.14	2.32	2.64	**3.15**
RAPT+PMFAs912	6.33	6.38	6.84	**7.19**	2.52	2.61	2.95	3.32

ventana (ventana de 2.5ms) permitirá un análisis más redundante de la periodicidad local, a cambio de intercalar un mayor número de tramas de análisis vacías (aumenta la probabilidad de no encontrar ninguna marca de *pitch*). Para este experimento, se ha escogido el algoritmo RAPT como PMA de entrada, ya que se trata del algoritmo de referencia a partir del que la aplicación de PMFA consigue obtener unos mejores resultados a lo largo de los experimentos.

En las tablas 4.9, 4.10 y 4.11, se presentan los resultados obtenidos mediante RAPT + PMFA utilizando las nuevas configuraciones (2.5ms y 15ms) respecto a los conseguidos con las ventanas de 5ms y 10ms, utilizando las medidas de evaluación GER y GPMER. Los resultados muestran una cierta tendencia a presentar las mínimas tasas de error sobre la diagonal de la matriz formada en filas por las distintas configuraciones sXY y en columnas por las cuatro ventanas de análisis. Es decir, a medida que aumenta el tamaño de la ventana de análisis, y consecuentemente, el paso entre ventanas, la configuración óptima de sXY también debe aumentar de valor. Por ejemplo, para el caso del corpus alegre, se puede observar claramente como, a medida que se aumenta el tamaño de la ventana, aparecen más valores por encima de la tasa de error del marcador de referencia (números en cursiva) (p.ej. para 10ms hasta $s26$ y para 15ms hasta $s68$). También se puede observar como GPMER presenta una tendencia algo más clara que GPER, ya que, como se ha discutido con anterioridad, GPMER es una medida algo más fina que GER. No obstante, es importante comentar que variaciones del orden de una décima en la tasa de error son poco significativas, para el volumen de datos con el que se trabaja. Asimismo, del análisis de las tasas de error obtenidas con cada ventana, se puede comprobar que, en general, a medida que aumenta el tamaño de la ventana, la tasa mínima conseguida (con una configuración de sXY mayor) también aumenta (con alguna excepción, p.ej. 2.5ms sobre el corpus sensual). Del mismo análisis se puede concluir que utilizando las ventanas de 5ms y 2.5ms se consiguen las mínimas tasas de error, sin que 2.5ms mejore de forma sustancial los resultados conseguidos con 5ms (incluso llega a empeorarlos sobre el corpus sensual según los resultados de la tabal

Tabla 4.11: Tasas de error sobre el corpus sensual para cuatro ventanas de análisis distintas. En cursiva, los valores peores que los de referencia, y en negrita el mejor resultado en cada barrido.

MÉTODO	GER (%)				GPMER (%)			
Ventana	2.5ms	5ms	10ms	15ms	2.5ms	5ms	10ms	15ms
RAPT	31.07				29.26			
RAPT+PMFAs13	21.73	24.05	*33.27*	*43.09*	12.60	13.76	24.58	*37.77*
RAPT+PMFAs24	20.72	21.06	23.48	27.99	12.16	10.93	14.28	19.11
RAPT+PMFAs26	20.37	20.88	23.23	27.73	12.33	11.22	14.71	19.78
RAPT+PMFAs34	20.36	20.64	21.65	23.48	**12.00**	10.37	12.45	14.31
RAPT+PMFAs37	20.09	19.93	21.25	23.05	12.31	**9.66**	12.99	14.91
RAPT+PMFAs48	**20.08**	20.21	20.79	21.68	12.42	10.28	12.55	13.55
RAPT+PMFAs68	20.21	**19.90**	**20.12**	20.87	12.54	9.81	**12.09**	12.59
RAPT+PMFAs79	20.19	20.02	20.20	20.80	12.83	9.83	12.16	**12.52**
RAPT+PMFAs912	20.65	20.46	20.35	**20.73**	13.36	10.18	12.57	13.12

4.11). De algún modo, se puede intuir que para señales de voz con valores de F_0 elevados y variabilidad elevada (como es el caso del corpus alegre), se conseguirán resultados óptimos con ventanas de análisis pequeñas y configuraciones sXY bajas (p.ej. $2.5m$ + $s24$ o $5ms$ + $s34$ o $s47$), mientras que para señales de voz con valores de F_0 bajos y poca variabilidad (como es el caso del corpus sensual) es necesario trabajar con tamaños mayores de ventana de análisis[31] y, consecuentemente, configuraciones sXY más altas. No obstante, resulta necesario realizar un estudio más exhaustivo para poder obtener unas conclusiones mejor sustentadas que las descritas aquí. El solapamiento entre ventanas es otro de los parámetros con los que se puede jugar en un futuro para añadir mayor redundancia al análisis de la periodicidad de la señal (repetición de marcas en tramas consecutivas). Si se disminuye el paso entre ventanas se aumentará el número de tramas de análisis, disponiendo de mayor redundancia en el proceso de estimación de la periodicidad de la señal de voz. Aunque no se han realizado pruebas en este sentido[32], se puede intuir que aumentar en exceso el solapado puede generar demasiados valores candidatos de periodicidad, teniendo en cuenta que se utiliza una métrica binaria para seleccionar el camino óptimo sobre la matriz de periodicidades por trama. No obstante, este estudio queda para futuras investigaciones.

Así pues, todavía queda camino por recorrer para intentar mejorar los resultados obtenidos y estudiar más exhaustivamente el funcionamiento de la presente propuesta sobre nuevos estilos de locución y otros marcadores o estimadores de la frecuencia fundamental.

[31] En esta misma línea se encuentran los resultados obtenidos sobre el corps Keele, que, aunque mediante PMFA mejoran los valores de referencia, estos dependen de la dinámica de la voz analizada. Por ejemplo M1, se ha ventaneado con $40ms$ en lugar de $25.6ms$ como se indica en la información que acompaña al corpus, mientras PMFA se ha mantenido a $20ms$, por los que los resultados son los peores de la tabla.

[32] En (Alías y Iriondo, 2001a) se realizó un primer estudio para distintas configuraciones de análisis (tamaño de ventana y paso) sobre un corpus de unos 5 min, evaluando los resultados a partir de la desviación de la F_0 media de los fonemas sonoros —equivalente a haber promediado el GER de cada fonema.

El objetivo final será disponer de un proceso de filtrado y remarcado de *pitch* fácilmente configurable y con un comportamiento lo más estable posible ante la variabilidad de los datos de entrada.

Posibles elementos de mejora

En los experimentos se ha optado por trabajar con un único rango de frecuencias para todos los corpus, locutores y algoritmos (PDA y PMA) utilizados —en este caso, [50,550] Hz. Como se ha podido comprobar a lo largo de las tablas presentadas, los resultados obtenidos por PMFA presentan unos resultados bastante buenos, mejorando en general de forma clara los obtenidos por los algoritmos de referencia. No obstante, parece lógico pensar que si el rango de frecuencias considerado se ajusta mejor a las características de la voz analizada, seguramente se podrán conseguir resultados aún mejores (p.ej. ver las tablas de resultados de *Keele*, con diez locutores distintos y resultados particulares por locutor). En este mismo contexto, se podría restringir algo más la búsqueda de posiciones candidatas para la ubicación temporal de las marcas de *pitch*. Siguiendo lo indicado en (Goncharoff y Gries, 1998), se construye una matriz **S** de señal formada por tramas *pitch*-síncronas de tamaño $2 \cdot T_{0max}$ alrededor de la estimación inicial de la posición de las marcas. A partir de esta misma información, se podría trabajar con un tamaño de ventana ajustado al valor medio estimado de la periodicidad[33], reduciendo el coste computacional del proceso y manteniendo la probabilidad de encontrar la posición más adecuada, según el criterio local utilizado, para ubicar la marca de pitch dentro del periodo.

Por otro lado, la simplicidad de ajuste de PMFA, que permite mejorar los resultados de partida sin tener que perder mucho tiempo en ajustar el algoritmo, tiene un coste. El PMFA propuesto es sensible a grandes errores de los algoritmos de entrada (valores erróneos presentes durante muchas tramas seguidas), que lo pueden arrastrar hacia valores de periodicidad erróneos (p.ej. alrededor de la trama 1900 en la figura 4.11 o alrededor de la trama 900 en la figura 4.12). No obstante, mientras estos valores espurios se encuentren en zonas sordas, el impacto será mínimo en la síntesis, como se puede observar de las buenas prestaciones conseguidas por PMFA. El problema aparece cuando estos errores arrastran a PMFA a valores erróneos en las zonas sonoras, sobretodo cuando el camino óptimo se aleja demasiado de la periodicidad media, por lo que necesita diversas tramas para volver a situarse en los valores correctos. Este problema, debido a que PMFA es un proceso ciego, ya que no dispone de información de la bondad de los valores de entrada —por ejemplo, en (de Cheveigné y Kawahara, 2001) se considera el nivel de confianza de la periodicidad—, se puede minimizar, o bien, ajustando el rango de frecuencias fundamentales considerado al rango de F_0 aproximado del habla estudiada, en lugar de trabajar siempre con un rango amplio como en los experimentos, o bien, incorporando en el proceso de filtrado información sobre el nivel de confianza de los valores candidatos de periodicidad (p.ej. utilizando la información que YIN proporciona (de Cheveigné y Kawahara, 2001)).

[33]Como resultado de la primera fase del método, se obtiene una primera estimación de periodicidad por trama $\hat{T}_0(j)$, por lo que se puede obtener el valor medio de la periodicidad de forma muy fácil. No obstante, sería prudente dejar un cierto margen de seguridad para absorber posibles errores de estimación.

Asimismo, se pretende estudiar otros criterios de posicionamiento local de las marcas de *pitch*, estudiando su impacto en la calidad de la síntesis mediante estudios y análisis subjetivos de la calidad. No obstante, la calidad subjetiva obtenida por el sistema ha sido evaluada indirectamente dentro de las pruebas realizadas en el capítulo 3, ya que el sistema de conversión de texto en habla basado en selección de unidades utiliza el corpus publicitario etiquetado mediante RAPT+PMFA. No obstante, quedan para un trabajo futuro estudios estrictamente centrados en la evaluación subjetiva de la calidad sintética obtenida cuando se apliquen modificaciones de la frecuencia fundamental de la señal de forma *pitch*-síncrona, como en (Sakamoto y Saito, 2000; Lin y Jang, 2004).

Todas estas reflexiones quedan abiertas para ser estudiadas en futuros trabajos de investigación, con el objetivo de mejorar aún más la robustez y la fiabilidad de la propuesta.

Capítulo 5

Conclusiones y trabajo futuro

El trabajo de investigación que se ha presentado en esta tesis se enmarca en el diseño de nuevas estrategias para la mejora de la calidad y la flexibilidad de los sistemas de conversión de texto en habla basados en corpus o selección de unidades (CTH-SU). Para ello se han abordado tres elementos fundamentales para este tipo de sistemas de síntesis: *(i)* el *módulo de selección de unidades*, a través de una nueva estrategia que permite ajustar subjetivamente los pesos de la función de coste a partir de la que se eligen las unidades del corpus, *(ii)* el *corpus de voz*, mediante el diseño de una nueva técnica para el etiquetado robusto de las marcas de *pitch*, y *(iii)* la *filosofía* de la conversión de texto en habla (CTH), mediante la definición e implementación de la que se ha denominado CTH multidominio, estrategia que abre la puerta al desarrollo de nuevos sistemas de CTH basado en selección de unidades (u otras estrategias de síntesis) más flexibles y adaptables a las necesidades planteadas por las aplicaciones en las que se enmarcan. Aunque las particularidades de cada uno de los tres problemas estudiados son distintas, todos ellos han sido abordados con el objetivo de avanzar en la consecución de sistemas de CTH de mayor naturalidad y flexibilidad, introduciendo un nuevo punto de vista para la resolución de los problemas planteados en el contexto de los CTH-SU. El ajuste subjetivo de pesos eficiente junto al mejor etiquetado del corpus permite mejorar la calidad de la conversión de texto en habla basada en selección de unidades, mientras que la conversión de texto en habla multidominio, fundamentalmente, permite mejorar la flexibilidad de los sistemas de CTH. Asimismo, la capacidad del método de ajuste robusto de marcas de *pitch* para trabajar sobre cualquier algoritmo de detección o marcado de *pitch* de entrada, dota de mayor flexibilidad al proceso de etiquetado del corpus de voz. Así pues, el conjunto de las aportaciones permite dar un paso más hacia la consecución del objetivo de cualquier CTH: la síntesis genérica perfecta (ver figura 3.1).

A lo largo de los distintos capítulos, y fundamentalmente en sus apartados de discusión, se han ido detallando algunos de los elementos claves de cada una de las propuestas realizadas sobre las tres líneas de investigación desarrolladas. No obstante, a continuación se presentan las conclusiones globales del trabajo realizado, junto a algunas de las diversas líneas de trabajo que quedan abiertas a partir de los resultados obtenidos hasta el momento.

5.1. Sobre el ajuste de pesos

En lo referente al problema del ajuste de pesos de la función de coste de selección, en los últimos años se han presentado varios métodos —además de la aproximación básica del ajuste manual de los pesos— que permiten entrenar de forma automática las ponderaciones que definen la importancia que tienen los subcostes de selección. Algunos de ellos trabajan con medidas objetivas en la comparación de las unidades utilizadas durante el entrenamiento, mientras otros tratan de optimizar subjetivamente —a partir del resultado de alguna prueba perceptual— el valor de los pesos de la función de coste. Como se refiere en la literatura, a lo largo del presente trabajo, se ha podido comprobar la complejidad del problema, sobretodo por la dificultad de relacionar satisfactoriamente la percepción humana con la selección de unidades.

A partir del análisis de las propuestas existentes en la literatura, se planteó la necesidad de disponer de un método que permitiera incorporar de forma explícita la percepción humana en el proceso de entrenamiento de los pesos, y que fuera capaz de superar las restricciones planteadas por las aproximaciones clásicas descritas en la literatura específica del tema, cambiando el enfoque habitual de las propuestas realizadas hasta el momento (no se pretende modelar la percepción humana, sino que ésta se incorpora directamente en el proceso). Para ello, se recorrió al campo de la Inteligencia Artificial para buscar una estrategia que puediera afrontar con garantías la resolución de un problema de optimización en un dominio multimodal complejo. Este tipo de problemas han sido abordados durante años con éxito por los *algoritmos genéticos* (AG), que pertenecen a este campo del conocimiento. Aunque existen diferentes ejemplos de cooperación entre estas técnicas, como se ha comentado a lo largo de la memoria, hasta el inicio de este trabajo de investigación, no se había planteado la aplicación de los algoritmos genéticos con el objetivo de ajustar eficientemente los pesos involucrados en el cálculo de la función de coste de selección. En esta línea, se han presentado dos nuevas aproximaciones para el entrenamiento de los pesos basadas en los AG, una de ellas, basada en una distancia objetiva —pretende codificar la similitud entre las unidades, generalmente, se utiliza la distancia Euclídea cepstral—, mientras que la otra está basada en incorporar el dominio perceptivo (subjetividad) de los usuarios al proceso (basada en los AG interactivos). Asimismo, para ambas aproximaciones, se ha definido un entorno de trabajo que permite entrenar conjuntamente los pesos de unidad y de concatenación (w_j^t y w_j^c), cuestión que permite dejar de lado el enfoque de entrenamiento independiente típicamente utilizado. En cuanto al método automático basado en AG, éste ha demostrado superar las restricciones de los métodos automáticos clásicos, por el hecho de representar una adaptación no lineal al problema, aprovechándose de su robustez ante entornos ruidosos y multimodales. En cuanto al método subjetivo, este ha evolucionado desde un sistema inicial de ajuste interactivo —utilizando, un algoritmo genético interactivo simple (AGI)— hasta la propuesta final de esta tesis, en la que se hace uso de un algoritmo genético interactivo activo (aAGI), adaptado a las características y las necesidades del problema. El método de ajuste subjetivo de pesos ha demostrado, en las pruebas desarrolladas, conseguir la mejor calidad sintética, junto a una menor fatiga del usuario y una mayor consistencia de sus evaluaciones de entre los métodos (objetivos y subjetivos) analizados.

Dado que el sistema de CTH-SU de nuestro grupo utiliza los difonemas y trifonemas como unidades básicas, las pruebas realizadas para el entrenamiento de los pesos han considerado estas unidades como dato de entrenamiento, considerando las parejas de difonemas (o trifonemas) para el entrenamiento conjunto de los pesos de unidad y concatenación. Los experimentos se han llevado a cabo a nivel de unidad (ajuste de pesos para cada una de las unidades básicas) y a nivel global (ajuste de pesos subjetivo a partir de un conjunto de frases de entrenamiento). En estas pruebas se ha podido observar como el método automático de ajuste de pesos basado en un AG diseñado supera las restricciones de los métodos clásicos, al permitir un ajuste sin restricciones lineales (a diferencia de MLR) y con un coste computacional asumible (a diferencia de WSS), aunque no consigue mejorar la calidad sintética obtenida. Por otro lado, las pruebas desarrolladas mediante la plataforma *web* diseñada incorporando el algoritmo genético interactivo (AGI) permiten observar que los patrones conseguidos mediante los métodos automáticos, incluso el basado en AG, distan del patrón perceptual definido por los usuarios. Asimismo, se constata la dificultad de poner la tarea de ajuste de los pesos en manos de los usuarios, cuestión que provoca la aparición de nuevos problemas que es necesario abordar —fundamentalmente, la fatiga y la consistencia del usuario. Para ello, una vez definidos los *algoritmos genéticos interactivo activos* (aAGIs) en (Llorà et al., 2005), se sustituye el AGI de la plataforma interactiva de ajuste de pesos por un aAGI, adaptado a las necesidades que plantea el problema de ajustar subjetivamente los pesos de la función de coste del módulo de selección de unidades del CTH. Por una parte, resulta necesario controlar la consistencia de los usuarios que realizan las evaluaciones, y por la otra, se adapta el aAGI definido por (Llorà et al., 2005), que trabaja con datos binarios, a trabajar con datos continuos (vectores de pesos). En este trabajo, también se propone un nuevo método para evaluar la *consistencia* de los usuarios durante el proceso de interacción basado en la detección automática de ciclos del grafo definido según sus preferencias ante las soluciones propuestas, junto a una nueva medida que permite cuantificar la consistencia de las evaluaciones. Como resultado del nuevo algoritmo, se observa que el usuario sólo necesita la mitad del tiempo de un AGI convencional para llegar a soluciones equivalentes, mejorando considerablemente la consistencia —y por lo tanto, la bondad— de las soluciones. Asimismo, se consigue que la calidad sintética obtenida a partir de los pesos entrenados mediante esta estrategia mejore significativamente respecto a los métodos objetivos y subjetivos de referencia. Por lo tanto, las pruebas validan satisfactoriamente la propuesta, abriendo un camino interesante para futuras investigaciones.

En el momento de escribir este documento, el estado actual del trabajo de investigación en el ámbito del ajuste de pesos para selección de unidades está encaminado hacia el estudio y mejora de las aproximaciones presentadas a partir de un nuevo conjunto de experimentos. El trabajo a desarrollar a partir de este momento debe permitir validar, y posiblemente generalizar, los resultados obtenidos en las pruebas desarrolladas en contextos más ambiciosos, tanto en lo que se refiere a la aproximación basada en un algoritmo genético (ajuste objetivo de los pesos) como en la basada en incorporar la percepción humana al proceso mediante un algoritmo genético interactivo activo (ajuste subjetivo eficiente de los pesos). En estas pruebas se pretende involucrar a un mayor número de evaluadores durante la fase de entrenamiento para poder determinar más claramente los patrones subjetivos distintos

que se han obtenido en las pruebas realizadas hasta el momento. Asimismo, es necesario reforzar la intuición inicial de que se trata de un problema de carácter multimodal, es decir, debido a la percepción humana, se pueden hallar diversos criterios de valoración de los resultados según los distintos perfiles de usuario que realizan las pruebas. Si esto se confirma, puede ser interesante estudiar las soluciones que se proponen desde el ámbito de los algoritmos genéticos a este tipo de problemas (como p.ej. la especiación) (ver sección 2.5), así como buscar estrategias para integrar los distintos perfiles en un perfil común, por ejemplo, ajustando de forma cooperativa las mismas pruebas entre distintos usuarios o bien, definiendo técnicas de integración de los resultados de los usuarios —p.ej. agrupando los grafos normalizados de comparaciones en un único grafo.

En un futuro se pretende hacer uso de otros corpus de voz con mayor cobertura y variabilidad de las unidades (difonemas y trifonemas, en este caso) para ser utilizados en el entrenamiento de los pesos para síntesis basada en selección de unidades. De este modo, se podrá estudiar los métodos propuestos sobre un mayor número de unidades, ya que se dispondrá de mayor robustez estadística para el entrenamiento de sus pesos. Por ejemplo, se estudia repetir las pruebas sobre un corpus de voz con distintas calidades vocales, como el corpus publicitario utilizado en la implementación de la propuesta de sistema de CTH multidominio (ver tabla 3.8). El objetivo es analizar el impacto de las características acústicas de cada subcorpus en el entrenamiento de los pesos. Parece lógico, de entrada, que, por ejemplo, la importancia de la frecuencia fundamental en la selección de unidades para el subcorpus de estilo alegre, debe ser mayor que en el corpus neutro, dada la distribución estadística de este parámetro en los corpus (ver figura 4.10). Por otro lado, también se pretende realizar las pruebas sobre el nivel intermedio de agrupación de los datos (ni a nivel de unidad ni a nivel global, sino agrupándolos según ciertos criterios: fonéticos, acústicos, etc.), para disponer así de un contexto que permita abordar el ajuste subjetivo eficiente de los pesos mediante los aAGIs, y disponer de patrones distintos para unidades de características distintas —la agrupación de datos también permite abordar el problema de la falta de redundancia estadística en el contexto de los corpus de tamaño reducido.

Como se ha descrito en el capítulo 2, las pruebas realizadas se han desarrollado utilizando una función de coste que contempla fundamentalmente información prosódica: subcostes de frecuencia fundamental, energía, duración y subcostes espectrales (mediante los MFCC). Será necesario incluir en un futuro factores fonéticos y lingüísticos del texto (acento, posición en frase, etc.) dentro de la función de coste, al estilo de otros trabajos (Blouin et al., 2002; Peng, Zhao y Chu, 2002; Campillo, Alba y Rodríguez Banga, 2005), estudiando así el impacto de estos parámetros en la función de coste. Asimismo, se pretenden estudiar otras funciones de coste distintas a la combinación lineal de los subcostes típicamente utilizada, en la línea de lo descrito en (Toda, 2003; Toda, Kawai y Tsuzaki, 2004), con el objetivo de disponer tanto de un proceso de ajuste de los pesos como de una función de coste sin restricciones lineales. En esta misma línea de trabajo, también se pretende estudiar detalladamente el impacto de la modificación prosódica de las unidades candidatas antes de la comparación con la unidad objetivo, por ejemplo, en la línea de lo descrito en (Chen, Chen y Kao, 2006), para modelar el impacto de esta modificación sin tener que incorporarla durante el proceso de selección. De este modo, se puede realizar el entrenamiento de los

pesos de selección independientemente del método de modificación prosódica utilizado, lo que permite reusar el bloque de selección de unidades diseñado para otras técnicas de procesamiento digital de la señal —es decir, se desacoplan los procesos.

Otro de los elementos clave del esquema de entrenamiento subjetivo definido es el tiempo necesario para realizar las pruebas. Aunque, gracias a la incorporación del aAGI en el proceso, la duración global de las pruebas se ha reducido significativamente, todavía existe un cierto lapso temporal entre iteración e iteración del proceso evolutivo para disponer de las nuevas frases sintéticas de la nueva generación. Para ello se estudiará cómo aplicar las distintas estrategias de agrupación o poda de los datos (descritas en la sección 2.1.2) para aumentar la eficiencia del proceso de selección de unidades sin disminuir la capacidad de cobertura en la elección de unidades del mismo. En esta misma línea, se pretende mejorar la plataforma *web* desarrollada incorporando nuevas funcionalidades de control de la consistencia del evaluador en tiempo de ejecución, en la línea de lo descrito en (Llorà et al., 2006), ya que por el momento, esta información se obtiene una vez finalizado el proceso de entrenamiento de los pesos para cada usuario.

Dado que el objetivo de este trabajo era el de plantear una nueva estrategia eficiente para el ajuste subjetivo de los pesos de la función de coste de selección, se han tomado valores típicos en la literatura para sus parámetros, que han sido ajustados mediante simples pruebas preliminares de viabilidad. Por lo tanto, resulta necesario estudiar de forma más exhaustiva las configuraciones de los AG utilizados, en términos de probabilidad de cruce, mutación y tamaño de la población. Asimismo, resulta necesario realizar un estudio exhaustivo del impacto de la resolución de los individuos de la población —los pesos de la función de coste— en el proceso de selección de unidades, ya que este elemento no ha sido analizado detalladamente a lo largo del presente trabajo de investigación —simplemente, se ha constado que el *ratio* entre el número de secuencias sintéticas generadas a partir del número de configuraciones de pesos considerado es razonable.

Finalmente, se pretende estudiar el método desarrollado en el contexto del entrenamiento de pesos para otras tareas involucradas en el ámbito del ajuste de los sistemas de conversión de texto en habla, p.ej. como en (Oversdotter y Llorà, 2006), con el objetivo de definir un proceso automático de ajuste subjetivo eficiente de los parámetros del CTH, controlando la consistencia de los usuarios (expertos en tecnologías del habla, generalmente) y reduciendo la fatiga que este tipo de procesos provoca (los ajustes manuales suelen ser bastante tediosos).

5.2. Sobre la CTH-MD

Uno de los elementos críticos de los sistemas de CTH-SU es la gran dependencia que estos sistemas tienen respecto al corpus que utilizan. Como se ha discutido a lo largo de esta tesis, los CTH-SU son capaces de generar voz sintética de alta calidad con las mismas características de la señal almacenada en el corpus, reduciéndose considerablemente la naturalidad de la señal generada cuando sus especificaciones de síntesis (prosodia, calidad vocal, estilo de locución, etc.) difieren de las que dispone el sistema de CTH-SU. En este trabajo se ha presentado una nueva propuesta que permite aumentar la flexibilidad de este

tipo de sistemas, manteniendo la calidad que consiguen en el contexto de su funcionamiento óptimo. Esta estrategia o filosofía de CTH se ha denominado CTH multidominio (CTH-MD), ya que permite sintetizar textos de distintos dominios con el estilo más apropiado de forma automática. Esta estrategia se basa en la filosofía de otros sistemas orales multimodales (sistemas de diálogo, reconocimiento automático, etc.) que permiten la interacción con el usuario sobre distintos dominios automáticamente, determinando el dominio de comunicación de forma explícita (indicado por el usuario) o de forma implícita (extraído del mensaje). Así pues, adaptar el sistema de interacción al dominio de comunicación permite mejorar las prestaciones del mismo, como se ha indicado en la sección 3.1.1.

Esta filosofía parte de la hipótesis de la relación existente entre el contenido y estructura del texto y el estilo de locución más apropiado para sintetizarlo. Por el momento, la CTH-MD se ha implementado siguiendo la filosofía de CTH-SU bajo la restricción de la correspondencia entre el dominio y el estilo de locución, siguiendo una estructura de corpus formada por varios subcorpus independientes (estrategia *tiering*) —debido a las particularidades vocales de los estilos de voz considerados. En este contexto, la elección de dominio implica elegir el modelo prosódico correspondiente para determinar la prosodia del texto de entrada, así como indicar al módulo de selección de unidades el subcorpus donde realizar la búsqueda de la secuencia óptima de unidades. No obstante, la filosofía CTH-MD permite ser implementada mediante otras estrategias, tanto de síntesis (p.ej. basada en HMM), como mediante otras tipologías de corpus multidominio (p.ej. un único corpus formado por un núcleo principal de voz con estilo neutro más pequeños subcorpus de dominio —estrategia *blending*), como se ha comentado en la sección 3.5. Asimismo, la filosofía CTH-MD permite reutilizar distintos corpus orientados a dominio disponibles, por ejemplo, para un mismo locutor, evitando así, tener que trabajar con múltiples CTH en paralelo.

Para llevar a la práctica esta metodología, resulta imprescindible incorporar un módulo que sea capaz de escoger el dominio o los dominios más adecuados para llevar a cabo la síntesis. En este contexto, se han estudiado distintas herramientas (representación, parametrización y modelado de los textos) procedentes del mundo de la Clasificación automática de Textos (CT) para implementar el módulo de clasificación automática de dominio del CTH-SU. Para ello se ha desarrollado un método de clasificación de textos adaptado a las necesidades que plantea la CTH-MD, fundamentalmente: trabajar con textos cortos y clasificar con un bajo coste computacional. A diferencia del enfoque de clasificación temática, resulta necesario que el sistema de CT para CTH-MD trabaje con todas las palabras del texto (no elimina palabras ni filtra su flexión), al estilo de las aplicaciones de CT estilísticas (no temáticas). En las pruebas preliminares se pudo observar la dificultad que esta tarea implica para métodos habitualmente aplicados a la clasificación temática de documentos, fundamentalmente por no reducir el espacio de datos de clasificación y por trabajar con textos extremadamente cortos (p.ej. una frase *vs.* artículos periodísticos). A raíz de estos resultados, se decidió incorporar nuevos parámetros para considerar la estructura y secuencialidad del texto (las coocurrencias entre palabras y las secuencias de palabras coincidentes entre el texto de entrada y el dominio a clasificar, denominada como longitud del patrón) en el proceso de clasificación, así como la importancia de las palabras relativas al texto a clasificar (denominado como *inverse word frequency*), cuestiones muy importantes dada la

inherente naturaleza secuencial del habla. Por este motivo, y con el objetivo de considerar toda la información de los textos, se ha propuesto un método de CT basado en la representación de los textos mediante un grafo de nodos ponderados denominado Red Relacional Asociativa (RRA), utilizando un modelo de espacio vectorial (MEV) para la clasificación de los textos. En este trabajo, se han presentado dos sistemas de CT utilizando el modelo RRA + MEV : el basado en la modelado de los datos mediante un espacio de representación global o RRA F, y el basado en el modelado de los textos mediante un espacio de representación reducido o RRA R —justificado algebraicamente como una aproximación de RRA F, según el criterio de los mínimos cuadrados. Ambos métodos de CT han demostrado su buen funcionamiento a lo largo de las pruebas objetivas desarrolladas, destacando el método de CT basado en RRA R por su buena eficiencia de clasificación, bajo coste computacional —minimizando la sobrecarga que la CT añade al proceso de CTH-MD— y robustez frente a la reducción de la longitud del texto de entrada y el ruido de los datos de entrenamiento. Estos resultados se constatan mediante el análisis del binomio *coste de computación - eficiencia de clasificación*, donde RRA R presenta también el mejor compromiso de los métodos estudiados. Es por ello que la estrategia RRA R ha sido la escogida para la implementación final del sistema de CTH-MD basado en selección de unidades, evaluado mediante el conjunto de pruebas subjetivas descritas. De todos modos, los moderados resultados conseguidos para pocas frases por documento (tanto para el corpus completo como el reducido) dejan abierto el camino para posibles mejoras del sistema de clasificación de dominios en futuras investigaciones.

En cuanto a las pruebas subjetivas realizadas, éstas se han centrado en validar el impacto del funcionamiento del CT, así como de la filosofía de la CTH-MD en general, sobre la calidad sintética obtenida. De las pruebas se puede concluir que gracias a la eficiencia del CT desarrollado, el sistema de CTH-MD implementado (basado en CTH-SU y corpus *tiering*) funcionará, en general, con la misma calidad que la obtenida por distintos CTH-DR trabajando en paralelo. Asimismo, se ha podido constatar en las pruebas la complejidad de la relación entre el mensaje y el mejor estilo de locución, sobretodo para textos tan cortos como los utilizados (en este caso, eslóganes publicitarios). En futuros trabajos de investigación se pretende estudiar con más profundidad la relación entre el texto de entrada y la mejor manera de pronunciarlo, tal y como se ha comentado en la introducción del capítulo 3, aunque parece esta línea de investigación pertenece más al ámbito semántico o cognitivo de la comunicación entre las personas.

A partir del trabajo desarrollado en esta tesis en el contexto de la definición y diseño de la estrategia de CTH-MD, se abren distintas de líneas de investigación. En cuanto a la filosofía de CTH-MD, una de las cuestiones que quedan abiertas pasa por determinar el contenido del corpus de voz. Como se ha comentado en esta tesis, la CTH-MD define una arquitectura de corpus multidominio totalmente flexible, que permite ser adaptado a distintas estrategias de CTH, desde la de propósito general (un único corpus genérico), pasando por la de dominio restringido (un único corpus de dominio), hasta la multidominio (implementada según estrategia *tiering* o *blending*), con el objetivo de disponer de sistemas de síntesis de alta calidad que incorporen diferentes emociones (alegre, triste, enfado, etc.), estilos de locución (sensual, publicitario, etc.), calidades vocales (susurro, carraspeo, etc.)

o dominios temáticos (periodístico, literario, etc.) en un mismo corpus. Por un lado, parece obvio que no tiene sentido intentar incorporar explícitamente todos estos dominios, pero por otro, también parece lógico que no se podrá obtener síntesis de alta calidad y flexibilidad a partir de un corpus únicamente de propósito general (neutro). En futuros trabajos de investigación, será necesario estudiar *qué* es necesario grabar y *qué* se puede modelar satisfactoriamente (mediante modificación prosódica, como por ejemplo la emoción triste desde el estilo neutro (Iriondo et al., 2004) o mediante transformación de la señal de voz (Yamagishi et al., 2005)). Esta es una línea muy interesante a estudiar en el ámbito de los sistemas de CTH, con el objetivo de disponer de los estilos o emociones básicas acompañados por los algoritmos de interpolación o mezcla capaces de llegar al estilo deseado con una buena calidad sintética. No obstante, todavía queda mucho camino por recorrer en esta línea de investigación.

Por otro lado, como se ha podido observar en este trabajo, el análisis del texto de entrada, sobretodo en el contexto de la CTH-MD, es muy complejo, por lo que todavía queda camino por recorrer para mejorar el funcionamiento del sistema de CT. Hasta el momento, se ha entrenado el CT a partir de los textos correspondientes al corpus de voz. Aunque los resultados obtenidos por el momento son satisfactorios, se plantea estudiar el impacto que puede tener en la tarea, por un lado, entrenar el sistema con muchos más textos de los que se dispone en el corpus y, por otro, incorporar corpus de información lingüística tipo WordNet al proceso (Ovesdotter, Roth y Sproat, 2005). Asimismo, se pretenden estudiar otros parámetros que permitan modelar mejor la información contenida en el texto, como por ejemplo, información morfosintáctica obtenida del módulo de procesamiento del lenguaje natural , como en (Ovesdotter, Roth y Sproat, 2005) o información contextual a partir del histórico de frases sintetizadas —no obstante, este enfoque, a menudo sugerido por expertos en el procesamiento del lenguaje natural, no siempre puede dar buenos frutos en el contexto de la CTH, donde, en principio, cada mensaje puede ser independiente del anterior.

Otra de las líneas de trabajo que se pretende abordar pasa por estudiar otras estrategias de clasificación distintas a las basadas en el modelo de espacio vectorial, como por ejemplo, el modelado probabilístico de los textos. Como se ha comentado en la sección 3.5, se deduce que el modelo RRA puede ser modelado, con cierta facilidad, como sistema de clasificación probabilístico, dada la información que contiene la red (información sobre la frecuencia de aparición y de coocurrencia de las palabras). Asimismo, se pretende estudiar si la combinación de los distintos clasificadores (clasificación por concurso o *boosting*, en inglés) estudiados puede ser útil para conseguir el objetivo de mejorar la eficiencia de clasificación cuando se trabaja con muy pocas frases por documento, en la línea de lo estudiado para agrupación no supervisada de documentos en (Sevillano et al., 2006a; Sevillano et al., 2006b).

Por otro lado, aunque la filosofía de CTH-MD define un corpus estructurado jerárquicamente, por el momento, las pruebas de clasificación se han desarrollado al nivel inferior de la estructura jerárquica. Se pretende explotar la estructura jerárquica obtenida de la aplicación del análisis en componentes independientes (ICA) al contenido del corpus, así como la clasificación del texto de entrada en distintos dominios (*soft classification* (Sebastiani, 2002). Ambas estrategias permitirán explotar completamente la organización de los datos

del corpus definida. Asimismo, a partir de las conclusiones obtenidas de la aplicación de ICA al problema, se pretende continuar trabajando en esta línea para complementar las características fundamentalmente estilísticas del método de CT basado en RRA desarrollado, con las meramente temáticas del método de CT basado en ICA. A partir de la organización de la colección de documentos en los tópicos temáticos independientes que lo han generado, se pretende aplicar la RRA para modelar cada uno de los dominios definidos por el algoritmo de CT basado en ICA, de entrada, en cada uno de los niveles de la jerarquía. De este modo, se puede trabajar con corpus no etiquetados previamente, ya que el algoritmo ICA es capaz de organizar los textos de forma semisupervisada, como se ha descrito en este trabajo.

Del mismo modo, se pretende continuar explotando la estrategia de CT basada en RRA en otras aplicaciones, desde otras de análisis de textos (p.ej. asignación de la autoría de un texto, determinación del género, etc.) a coninuar mejorando el funcionamiento del CTH, como por ejemplo, ayudando a desambiguar el mensaje en la fase de normalización del texto —según las características y particularidades del CTH sobre el que se incorpore el módulo de clasificación de dominio, éste tendrá un impacto u otro en el funcionamiento del sistema de síntesis. La información que la RRA contiene puede ser utilizada para distintas tareas de clasificación, desde las puramente temáticas (no se utilizan las relaciones de los textos y se eliminan las palabras temáticamente ambiguas, como p.ej., artículos o conjunciones) a otras tareas estilísticas, como por ejemplo, determinar el género del texto —incluso se pretende estudiar si puede ser útil para detectar el estado emocional del autor a partir del texto, como en (Tao y Tan, 2004; Ovesdotter, Roth y Sproat, 2005), cuestión muy interesante en el contexto de la CTH expresiva.

Finalmente, se quiere incorporar el módulo de CT en el flujo del sistema de CTH-SU implementado, ya que por el momento se ha evaluado de forma aislada (sistema independiente). Para ello se pretende incorporar la información de la clasificación del texto a sintetizar mediante el etiquetado correspondiente. Se estudiará la aplicación de etiquetas —p.ej. tipo SSML— a alto nivel, como en (Alías et al., 2005) o a bajo nivel como en (Hamza et al., 2004), entre otros.

5.3. Sobre el PMFA

El etiquetado y la revisión fiable de los corpus de voz es una de las tareas fundamentales en el contexto de la conversión de texto en habla basada en corpus. En este contexto, la extracción de las marcas de *pitch* es una de las tareas más complejas y costosas, sobretodo para los corpus de tamaño considerable típicamente utilizados en CTH-SU. El objetivo de este trabajo ha sido el de definir un método que permita obtener un marcado fiable para facilitar el etiquetado robusto de corpus de voz para CTH-SU. Una de las conclusiones fundamentales del trabajo es que resulta prácticamente imposible diseñar métodos de detección y marcado de *pitch* universales, es decir, que funcionen óptimamente para cualquier voz, estilo de locución, calidad de la señal, etc. sin tener que ser adaptados a las características de la señal de voz con las que se trabaja. Aunque el uso de la señal electroglotal (EGG) ha simplificado, en parte, el análisis y el marcado del *pitch* de las señales de voz, todavía

existen ciertas circunstancias en las que, o bien, la calidad de la señal EGG no es del todo satisfactoria (Ferencz et al., 2004), o bien, no se dispone de esta señal acompañando a la voz.

El trabajo de investigación desarrollado se ha centrado en el diseño de un método de ajuste robusto de las marcas de *pitch* a partir de la información obtenida de un algoritmo de detección o de marcado de *pitch* de entrada cualquiera (independientemente de la señal que se haya utilizado para extraer la información relacionada con la frecuencia fundamental de la señal). Una de las características principales de la propuesta, denominada *Pitch Marks Filtering Algorithm* (PMFA), es su fiabilidad y flexibilidad, ya que ha demostrado su buen funcionamiento (mejora significativa de las tasas de error) sobre los tres métodos de referencia utilizados (RAPT (Talkin, 1995), YIN (de Cheveigné y Kawahara, 2002) y SHRp (Sun, 2000; Sun, 2002)) y los dos corpus de voz analizados (Keele database (Plante, Meyer y Ainsworth, 1995) y el corpus publicitario del CAP-UAB, ver tabla 3.8). No obstante, como se ha comentado, encontrar la configuración óptima para un determinado locutor o corpus de voz necesita de un cierto ajuste de los parámetros considerados por PMFA, por ejemplo, del rango de valores de frecuencia fundamental considerados —información que se puede extraer de un pequeño análisis previo de la señal de voz del corpus. Otro de los elementos clave de PMFA es que consigue obtener una secuencia de marcas robusta (localmente, bien posicionada dentro de los periodos de voz y globalmente, con una evolución suave a lo largo de la señal de voz) a partir del ajuste de muy pocos parámetros, a diferencia de otras propuestas que necesitan de funciones de coste más complejas (ver sección 4.3). Además de definir el criterio local utilizado para el posicionamiento temporal de las marcas (por el momento el máximo de amplitud de la señal en valor absoluto), la propuesta de PMFA necesita sólo concretar las restricciones de variación máxima permitidas entre estados consecutivos sobre los que se aplica el algoritmo de programación dinámica restringido que incorpora, elemento fundamental de la propuesta. De este modo se evitan las transiciones bruscas entre periodicidades (de las tramas de análisis) vecinas, permitiendo una evolución suave de la frecuencia fundamental a lo largo de la señal de voz.

Las pruebas realizadas han permitido estudiar el funcionamiento del PMFA más allá de los análisis típicos en el contexto de la investigación en el marcado automático de *pitch*: pequeños corpus de voz neutros, y evaluación centrada en zonas claramente sonoras. En este trabajo, se ha evaluado el funcionamiento de la propuesta sobre un corpus de grandes dimensiones (comparado con los típicamente utilizados en otros trabajos) y con distintos estilos de locución (características acústicas de la señal y rangos de frecuencia distintos), considerando todas las zonas claramente no sordas (a diferencia de otros trabajos centrados sólo en las zonas claramente sonoras), ya que las transiciones sonora-sorda y sorda-sonora son zonas críticas de marcado y afectan decisivamente a la calidad de la señal sintética generada por el conversor de texto en habla basado en técnicas de síntesis *pitch*-síncronas, p.ej. PSOLA. De este modo, se puede verificar con más garantías el comportamiento de los algoritmos estudiados y las mejoras conseguidas por la propuesta, sobretodo teniendo en cuenta el contexto en el que se enmarca: la CTH basada en selección de unidades.

Otro de los elementos clave del marcado de *pitch* es su evaluación. Si las marcas comparadas se ubican siguiendo el mismo criterio local (p.ej. máximo de energía), resulta re-

lativamente sencillo determinar la fiabilidad el algoritmo evaluado. No obstante, no todos los métodos de marcado de *pitch* siguen el mismo criterio local, por lo que se dificulta su evaluación. En este trabajo, se ha definido una nueva medida de evaluación denominada *Gross Pitch Marks Error Rate* (GPMER) capaz de comparar el funcionamiento de algoritmos de marcado de *pitch* independientemente del criterio local utilizado para ubicar las marcas de *pitch*. De este modo, se evita tener que alinear previamente las marcas, con los errores que este proceso puede comportar. Los estudios realizados demuestran que es una medida capaz de dar información detallada del comportamiento de cada algoritmo sobre los distintos corpus utilizados para la evaluación.

De todos modos, aunque este método ya ha sido aplicado satisfactoriamente para el etiquetado de nuevos corpus (p.ej. el corpus de la aplicación meteorológica descrita en el anexo D.1), todavía existen diversas líneas de trabajo abiertas con el objetivo de mejorar y generalizar el funcionamiento del PMFA propuesto. Entre otras, resulta necesario realizar un barrido más exhaustivo de los parámetros de análisis utilizados, tanto, los que controlan la pendiente máxima permitida, como, el tamaño y solapado de la ventana de análisis de las marcas de entrada. El objetivo final es encontrar una relación (más allá de las definidas en este trabajo) capaz de determinar la configuración más adecuada del PMFA para el tipo de señal de voz analizado. Asimismo, se pretenden incorporar restricciones de segundo orden (no sólo la restricción de pendiente sino la restricción de aceleración) para controlar mejor algunas oscilaciones que provocan valores de frecuencia fundamental espurios sobre PMFA. En esta misma línea, parece interesante incorporar información sobre la fiabilidad de los valores de entrada al método propuesto, o bien, mediante información extraída del propio método de entrada de análisis de la periodicidad (p.ej. YIN (de Cheveigné y Kawahara, 2002)), o a partir de un esquema de votación en el que se utilizan distintos métodos de detección o marcado de *pitch*, para poder desambiguar los valores binarios de la matriz de análisis de periodicidad utilizada.

Por otro lado, el análisis de la propuesta para distintos estilos de locución continúa siendo uno de las líneas de trabajo más interesantes del método, bajo el enfoque del desarrollo de nuevos sistemas de CTH-MD o CTH expresiva.

Apéndice A

Algoritmo de agrupación de unidades para ajuste de pesos

Dividir el espacio de unidades en grupos (*clusters*) ofrece un nivel de precisión intermedio entre el ajuste global (un solo vector de pesos para todo el corpus) y el ajuste local (una vector por unidad) (Hunt y Black, 1996; Meron y Hirose, 1999). Esta aproximación permite disponer de configuraciones de pesos ajustadas a las características de las unidades que forman cada grupo. La agrupación de unidades para el ajuste de los pesos de selección se suele realizar a partir de la información fonética de las unidades (Meron y Hirose, 1999; Campillo y Rodríguez Banga, 2003; Campillo, Alba y Rodríguez Banga, 2005). Bajo este mismo enfoque, se ha llevado a cabo la agrupación de los difonemas (y trifonemas), unidades mínimas de síntesis para el CTH basado en corpus desarrollado. En este caso, se toma en consideración toda la información fonética de las unidades. Para el catalán, las etiquetas fonéticas consideradas son:

- *Tipo de unidad*: vocal, consonante, semivocal, silencio.
- *Modo de articulación*:
  - Consonante: oclusiva, fricativa, lateral, vibrante, nasal.
  - Vocal: cerrada, abierta, semiabierta.
- *Punto de articulación*:
  - Consonante: bilabial, dental, velar, alveolar, palatal, labiodental, interdental, prepalatal
  - Vocal: frontal, central, posterior.
- *Sonoridad*: sorda, sonora.

A partir de la información a nivel de fonema, se procede a etiquetar al difonema simplemente mediante la combinación de los parámetros que cada fonema presenta por separado.

Para el caso de los trifonemas, el fonema central (/r/, /l/, ...), que suele presentar una importante coarticulación con la vocal que le acompaña, no es considerado para el etiquetado de la unidad. Como resultado, se obtiene, por ejemplo, el siguiente etiquetado:

```
/pa/=[CONSONANTE-VOCAL, OCLUSIVA-ABIERTA, BILABIAL-CENTRAL, SORDA-SONORA]
```

siguiendo el mismo orden indicado en la enumeración de parámetros fonéticos considerados.

El método de agrupación de unidades propuesto toma en consideración tanto las categorías fonéticas como el número de realizaciones de las unidades. De este modo, se pretende no sólo agrupar las unidades fonéticamente parecidas, sino que se busca que éstas queden *bien* distribuidas dentro de los grupos (grupos equilibrados en número de realizaciones). Es decir, se pretende evitar la aparición tanto de *clusters* sobrepoblados como de *clusters* residuales (con unidades con insuficiente número de realizaciones para el ajuste de pesos). Por lo tanto, el algoritmo deberá dividir los datos del corpus en grupos estadísticamente representativos, para conseguir un entrenamiento fiable de los pesos. De este modo, unidades con muchas realizaciones (p.ej. /@l/, /l@/, /@n/, etc.), con un abanico de realizaciones muy amplio, prácticamente constituirán un único grupo. En cambio unidades con muy pocas realizaciones (1 o 2, p.ej. /tt/, /pt/, etc.) necesitan compartir grupo para disponer de suficiente variabilidad estadística. Además, muchas de estas unidades quedarán distribuidas a lo largo de los grupos, reduciendo los inconvenientes que provoca el entrenamiento de unidades con un número de realizaciones insuficiente.

Según estas premisas, se deberá escoger adecuadamente tanto el algoritmo de *clustering* como el número óptimo de *clusters* para distribuir las unidades en grupos uniformes (tan equidistribuidos como el corpus permita).

Método propuesto

Siguiendo una estrategia similar a la indicada por Meron y Hirose (1999) (i.e. un árbol de decisión), se ha decidido trabajar con adaptación del algoritmo *Classification and Regression Tree* (CART) (Breiman et al., 1984) para la agrupación de difonemas y trifonemas en el marco del ajuste de pesos. Se trata de un método de aprendizaje automático inductivo que organiza los datos mediante árboles de decisión binarios. Primero, a partir del conjunto de parámetros considerados, se construye el árbol completo (*growing*) mediante particiones binarias que minimizan la impureza de los nodos del árbol. A continuación, se realiza la poda (*pruning*) del árbol hasta llegar al número de nodos deseado (N), evitando la aparición de grupos residuales (*isolated clusters*). Destacar que uno de los puntos claves de este algoritmo es su capacidad de tratar implícitamente la dispersión de los datos (en este caso, la distribución de las unidades del corpus), ya que sólo dividirá un nodo en dos si éste dispone de suficientes datos y la partición permite reducir la impureza del nodo de partida (Black y Taylor, 1997a). Comentar que el algoritmo CART ha sido utilizado con anterioridad dentro del ámbito de la CTH basada en corpus. Entre otras: la agrupación acústica de las unidades del corpus como alternativa al cálculo de la función de coste objetivo (Black y

Taylor, 1997a), o el modelado de duraciones de las unidades (Black y Taylor, 1997b; Syrdal et al., 1998).

En el presente trabajo, se ha modificado el algoritmo CART para agrupar los difonemas y trifonemas en grupos lo más *equidistribuidos* posible a partir de información fonética categórica (son valores discretos, por lo que la fase de regresión no es aplicable). Por lo tanto, se adapta un método de aprendizaje supervisado (clasificador) a la agrupación automática de datos según criterios estadísticos (aprendizaje no supervisado). En este caso, se trata de un CART categórico, que sólo permite preguntas sobre valores discretos independientes X, es decir, $(X \in A)$? y $(X \notin A)$?. El objetivo es determinar, para cada posible partición (nodo), la pregunta o grupo de preguntas A_i $(1 \leq i \leq N)$, según el caso, que minimiza la entropía (ϕ), calculada sobre la agrupación de unidades que consigue (ver ecuaciones (A.1) y (A.1)). Esta es una medida de impureza típica para este tipo de métodos de clasificación ((Duda, Hart y Stork, 2001), cap. 8.4.3).

$$\phi(A) = min_{[1..N]}(\phi(A_i)) \tag{A.1}$$

$$\phi(A_i) = -\frac{\sum Unidades_{A_i}}{\sum Unidades} * log\left(\frac{\sum Unidades_{A_i}}{\sum Unidades}\right) \tag{A.2}$$

A continuación, se describe un ejemplo práctico del funcionamiento del algoritmo basado en CART implementado a partir de un nodo que contiene 696 unidades.

1. Para a cada atributo se calcula la entropía que se obtiene de su partición, siguiendo la ecuación A.2:

   - ***Tipo de unidad:*** $\phi(A_1) = 0{,}231$
     `CONSONANTE-VOCAL`: 539 unidades
     `VOCAL-CONSONANTE`: 157 unidades

     $$\phi(A_1) = -\left(\frac{539}{696} \cdot log\left(\frac{539}{696}\right) + \frac{157}{696} \cdot log\left(\frac{157}{696}\right)\right) = 0{,}231$$

   - ***Punto de articulación:*** $\phi(A_2) = 0{,}5484$
     `ALVEOLAR-CENTRAL`: 292 unidades
     `CENTRAL-ALVEOLAR`: 157 unidades
     `INTERDENTAL-CENTRAL`: 124 unidades
     `VELAR-CENTRAL`: 123 unidades

     $$\phi(A_2) = -\left(\frac{292}{696} \cdot log\left(\frac{292}{696}\right) + \frac{157}{696} \cdot log\left(\frac{157}{696}\right)\right) -$$
     $$-\left(\frac{124}{696} \cdot log\left(\frac{124}{696}\right) + \frac{123}{696} \cdot log\left(\frac{123}{696}\right)\right) = 0,5484$$

- ***Modo de articulación:*** $\phi(A_3) = 0{,}6279$
  `LATERAL-SEMIABIERTA`: 169 unidades
  `SEMIABIERTA-NASAL`: 157 unidades
  `FRICATIVE-SEMIABIERTA`: 124 unidades
  `NASAL-SEMIABIERTA`: 123 unidades
  `OCLUSIVA-SEMIABIERTA`: 123 unidades

$$\phi(A_3) = -\left(\frac{169}{696}\cdot log\left(\frac{162}{696}\right) + \frac{157}{696}\cdot log\left(\frac{157}{696}\right)\right) - \\ -\left(\frac{124}{696}\cdot log\left(\frac{124}{696}\right) + 2\cdot\frac{123}{696}\cdot log\left(\frac{123}{696}\right)\right) = 0,6279$$

- ***Sonoridad:*** $\phi(A_4) = 0{,}2170$
  SONORA-SONORA: 573 unidades
  SORDA-SONORA: 123 unidades

$$\phi(A_4) = -\left(\frac{573}{696}\cdot log\left(\frac{573}{696}\right) + \frac{573}{696}\cdot log\left(\frac{123}{696}\right)\right) = 0{,}2170$$

2. Se escoge el atributo que consigue una partición de los datos con menor entropía, en este caso A_4 (Sonoridad), y se procede a dividir los datos por su valor dominante (SONORA-SONORA). Por lo tanto, se obtiene la partición ($X \in$ [SONORA-SONORA]) y ($X \notin$ [SONORA-SONORA]).

Este proceso, se repite hasta conseguir la división máxima de los datos, es decir, que todos los nodos tengan entropía mínima. Una vez finalizado el proceso de crecimiento del árbol, se procede a su poda. A cada nivel de la estructura arbórea se agrupa el nodo con menor número de unidades (`MIN`) con su nodo complementario, subiendo un nivel en la jerarquía. Este proceso se repite hasta que se llega al número de nodos (N) deseado.

Número de *clusters*

Finalmente, resulta necesario definir el número de grupos necesario para repartir suficientemente las unidades del corpus de voz. Este valor, denominado en este trabajo como número *óptimo* de *clusters* (N^*), se define como el número de grupos que permite una división suficiente de los datos, consiguiendo, a la vez, una partición del espacio lo más uniforme posible. Por lo tanto, para determinar N^* es necesario evaluar el grado de uniformidad de la distribución de las unidades. Para ello, se escogió, inicialmente, calcular la kurtosis de la distribución obtenida, ya que este parámetro mide lo apuntada o aplanada que es una distribución respecto a la normal. Desafortunadamente, el número de muestras que se toma de la distribución ($N \leq 100$) es insuficiente para que el resultado de este momento de orden cuarto sea fiable. Por lo tanto, se decide trabajar con momentos de orden inferior, combinándolos entre sí mediante un multicriterio estadístico. Se mide:

1. El número de unidades del grupo menos poblado (`MIN`),
2. El número de unidades del grupo más poblado (`MAX`),
3. La desviación estándar del número de unidades por grupo (`STD`),
4. La diferencia `MAX-MIN`,
5. La pendiente de la distribución ordenada de unidades por grupo (`SLOPE`).

Así pues, según este multicriterio, el número óptimo de grupos para el ajuste de pesos (N^*) se puede definir como el valor de N que consigue maximizar la uniformidad de los datos (`STD`, `SLOPE` y `MAX-MIN` mínimos), evitando la aparición de grupos residuales (`MIN` máximo) o predominantes (`MAX` mínimo).

Ajuste y evaluación de la propuesta

Les pruebas se han llevado a cabo sobre un corpus de voz en catalán formado por 1520 frases, formadas por 9863 realizaciones de las 1207 unidades (difonemas y trifonemas). Aunque no es un corpus diseñado explícitamente para CTH basada en selección de unidades (20 minutos de duración), es un buen banco de pruebas para aplicar el método de agrupamiento de unidades, ya que muchas de ellas están pobremente representadas en el mismo.

Los experimentos siguientes evalúan tres aspectos del proceso de agrupación de los difonemas del corpus. Primero, se selecciona el mejor conjunto de preguntas fonéticas para el CART, evaluando la capacidad de cada configuración para distribuir uniformemente las unidades. Seguidamente, se compara el algoritmo basado en CART diseñado con dos métodos clásicos de *clustering* para evaluar la viabilidad del método propuesto. Finalmente, se determina el número óptimo de grupos siguiendo el multicriterio estadístico que se acaba de describir. Las pruebas se realizan para $N \in [3, 100]$.

Elección del conjunto de preguntas fonéticas más adecuado

Como ya se ha comentado, la división de los datos basada en CART se realiza a partir de un conjunto de preguntas fonéticas: el tipo de unidad, la sonoridad, el modo y el punto de articulación de los difonemas. A continuación se presenta el estudio que analiza qué características fonéticas debe tener en cuenta el algoritmo CART para funcionar de forma eficiente. Inicialmente, parece lógico dejar al algoritmo que escoja qué preguntas fonéticas debe considerar, por lo que se consideran los 4 grupos de preguntas al proceso (CART-4q). Sin embargo, se ha querido contrastar esta hipótesis mediante un pequeño experimento. En él se comparan el funcionamiento del CART-4q respecto a las cuatro combinaciones posibles del CART de 3 grupos de preguntas (CART-3q):

- CART-3q0: tipo de unidad, punto de articulación, modo de articulación

- CART-3q1: tipo de unidad, punto de articulación, sonoridad
- CART-3q2: tipo de unidad, modo de articulación, sonoridad
- CART-3q3: punto de articulación, modo de articulación, sonoridad

Grupos menores de preguntas presentan resultados muy deficientes para la tarea propuesta, por lo que no se estudia en detalle su funcionamiento en este experimento.

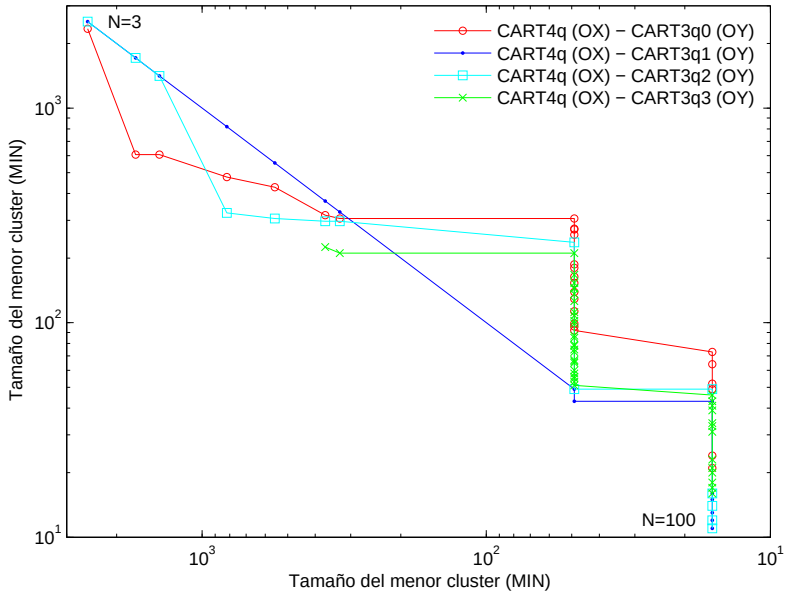


Figura A.1: Comparativa de la estadística `MIN` en escala logarítmica entre CART-4q (4 preguntas) *vs.* CART-3q0 (sin sonoridad), CART-3q1 (sin modo de articulación), CART-3q2 (sin punto de articulación) y CART-3q3 (sin tipo de unidad). $N = 3$ corresponde al punto superior izquierdo de la gráfica y $N = 100$ al punto inferior derecho de cada pareja de valores ($N \in [3, 100]$).

La figura A.1 presenta una comparativa entre CART-4q junto a dos de las cuatro combinaciones posibles del CART-3q, para la estadística que evalúa el número de unidades en el grupo menos poblado. En este caso, cuanto mayor sea el valor `MIN`, mejor será el agrupamiento, para cada valor de N.

Después de analizar los resultados obtenidos a partir de los indicadores estadísticos considerados, se puede concluir que CART-4q presenta un comportamiento más estable a lo largo de los N *clusters* —presenta un grupo de unidades mínimo de mayor tamaño que las variantes que contemplan sólo 3 preguntas. Por lo tanto, la configuración 4q es la escogida para realizar las siguientes pruebas. Sin embargo, cabe destacar que CART-3q también presenta un buen funcionamiento para valores pequeños de N, cuando no se incluye

el modo de articulación, y una ligera mejora de los resultados para valores mayores de N cuando se excluye la sonoridad. Además, se puede concluir de este experimento que el tipo de unidad es fundamental para obtener una buena partición del espacio de datos. CART-3q es incapaz de agrupar los datos cuando $N < 8$ si se excluye este dato del conjunto de preguntas a considerar por el algoritmo (ver figura A.1).

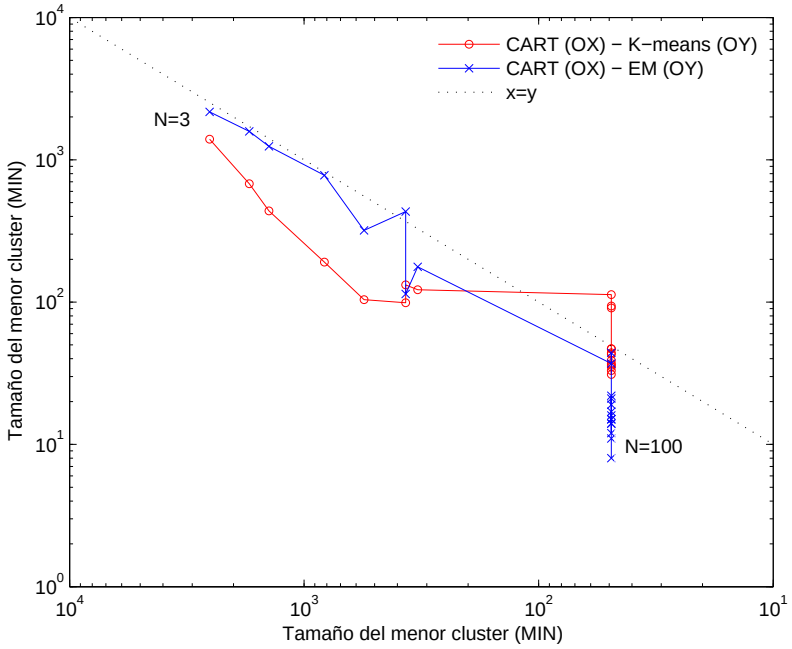


Figura A.2: Comparativa de la estadística `MIN` en escala logarítmica entre CART vs. *K-means* y CART vs. EM, donde $N = 3$ corresponde al punto superior izquierdo de la gráfica y $N = 100$ al punto inferior derecho de cada pareja de valores ($N \in [3, 100]$).

Comparación de la propuesta respecto otros métodos de agrupamiento

En el siguiente experimento se evalúa el funcionamiento del algoritmo de agrupamiento presentado basado en CART respecto a dos algoritmos típicos de *clustering*: *K-means* y *Expectation-Maximization* (EM). Ambos métodos serán aplicados al problema de la agrupación uniforme de difonemas a partir de información categórica, utilizando el *software* `WEKA` (Witten y Frank, 2000). En ambos métodos, la distribución final de los datos en grupos depende de la inicialización al problema. Por ello, se analiza su comportamiento medio para 10 inicializaciones (*semillas*) distintas. De los experimentos desarrollados se concluye que el algoritmo CART presenta un mejor comportamiento a lo largo del experimento, según

el multicriterio estadístico contemplado, ya que maximiza la uniformidad de la distribución de los grupos respecto a los otros dos métodos (ver (Formiga, 2005) para más detalles). En la figura A.2 se presenta la comparativa según el estadístico `MIN`.

Número óptimo de grupos

Aunque el multicriterio estadístico se diseñó para definir claramente el número óptimo de grupos, desafortunadamente, los indicadores estadísticos no coinciden en un único valor para determinar N^* inequívocamente. Por lo tanto, el valor de N^* ha sido seleccionado mediante un criterio heurístico. Para el corpus en cuestión, el número óptimo de grupos se ha fijado en 10, ya que presenta el mejor comportamiento a nivel del multicriterio estadístico. Concretamente, en $N = 11$ aparece un cluster residual (con unidades que disponen de 1 o 2 realizaciones, con un total de 49 realizaciones en el mismo) (ver figuras A.1 y A.2). Como se ha comentado en este apartado, es imprescindible evitar la aparición de grupos residuales de unidades. Por lo tanto, finalmente, se escoge $N^* = 10$, ya que se consigue una división suficiente de los datos evitando la aparición de uno de los factores críticos: unidades sin realizaciones suficientes en un mismo grupo.

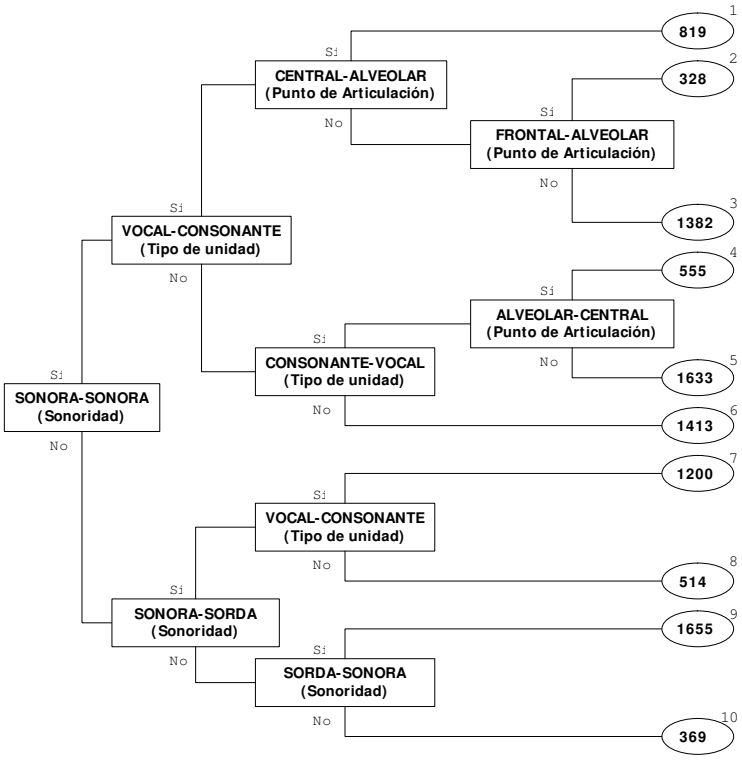


Figura A.3: CART obtenido para el número óptimo de grupos $N^* = 10$. Se indica el número de realizaciones de difonemas por grupo.

La figura A.3 presenta el árbol resultante para el número de grupos óptimo escogido. Comentar que, aunque el árbol ha sido construido a partir de la configuración CART-4q, finalmente sólo tres tipos de preguntas fonéticas son necesarias para N^*: el tipo de unidad, la sonoridad y el punto de articulación. Por lo tanto, así como ya se ha comentado en el apartado de elección del conjunto de preguntas, en este caso, la configuración CART-3q1 es suficiente para particionar el espacio de difonemas en 10 grupos para este corpus de voz.

Apéndice B

Sobre el ajuste robusto de marcas de *pitch*

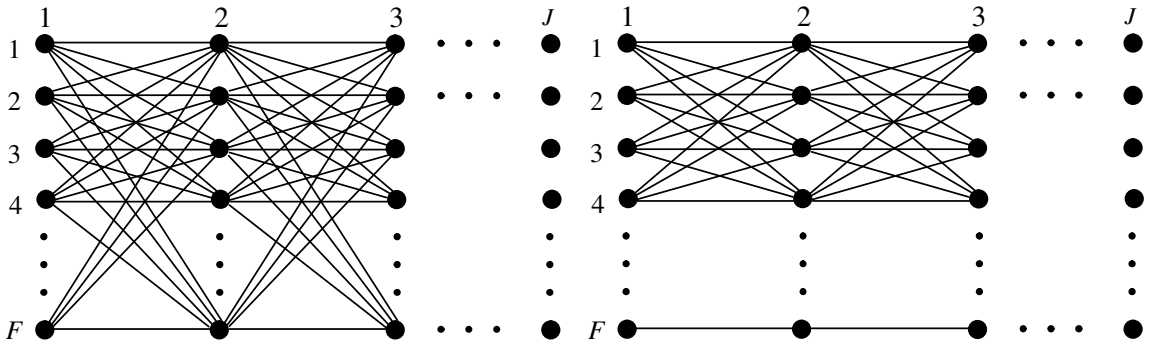
B.1. Número de caminos de la estructura *trellis* restringida

En el contexto del algoritmo de ajuste robusto de marcas de *pitch* desarrollado, se incluye un algoritmo de programación dinámica restringido. Esta restricción se centra en el número de estados conectados entre pasos consecutivos de la búsqueda. El parámetro que controla este número de transiciones es S_{max} o pendiente máxima. El valor de S_{max} influye directamente sobre el coste computacional del PMFA, ya que el tamaño de la estructura *trellis* generada durante la primera fase del algoritmo de programación dinámica depende de este valor. Por ejemplo, en la figura 4.3 se presenta un fragmento de los caminos generados con $S_{max} = 1$ partiendo de la casilla p_{51}, por lo que cada paso dentro de la estructura trellis consigue explorar 3 filas de la siguiente trama, es decir, permite $c(j) - c(j-1) = \{-1, 0, 1\}$. En la figura 4.4 se presenta, para el mismo ejemplo de la figura 4.3, el resultado que se obtendría considerando $S_{max} = 2$, dado que cada paso del barrido dentro de la estructura *trellis* consigue, en este caso, explorar 5 filas de la siguiente trama, es decir, permitiría $c(j) - c(j-1) = \{-2, -1, 0, 1, 2\}$; y así sucesivamente, siguiendo la relación $2 \cdot S_{max} + 1$, en lo que se refiere a la diferencia de periodicidades permitida entre tramas contiguas. No obstante, debido a las condiciones de contorno, el número de transiciones contempladas por el algoritmo de programación dinámico restringido entre dos tramas consecutivas será:

$$(T_{0\,max} - T_{0\,min} + 1) \cdot (2 \cdot S_{max} + 1) - S_{max}(S_{max} + 1), \tag{B.1}$$

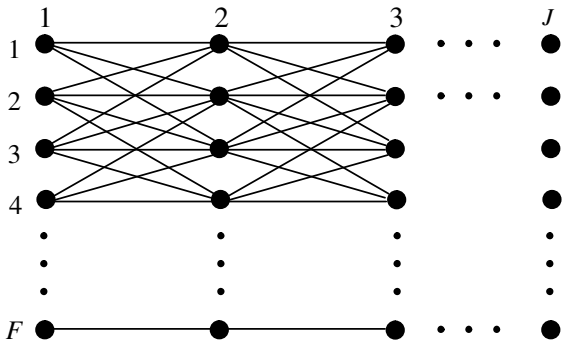
a diferencia de un algoritmo de programación dinámica de conectividad total entre estados que permite $(T_{0\,max} - T_{0\,min} + 1)^2$ conexiones (ver figura B.1(a)). Por ejemplo, para $S_{max} = 3$ se pierden 12 conexiones por condiciones de contorno (ver figura B.1(b), mientras que en $S_{max} = 2$, se pierden 6 (ver figura B.1(c)) y para $S_{max} = 1$ se pierden sólo 2 conexiones (ver figura B.1(d)), respecto a las $(T_{0\,max} - T_{0\,min} + 1) \cdot (2 \cdot S_{max} + 1)$ conexiones que teóricamente habría si se permitieran $2 \cdot S_{max} + 1$ posibles conexiones entre estados consecutivos

(variaciones de periodicidad entre intertrama). De este modo, el número total de caminos $c^s(j)$ posibles pasa de ser $(T_{0\,max} - T_{0\,min} + 1)^J$ para el caso no restringido (donde J indica el número de columnas -o tramas- de la estructura *trellis*), a ser:

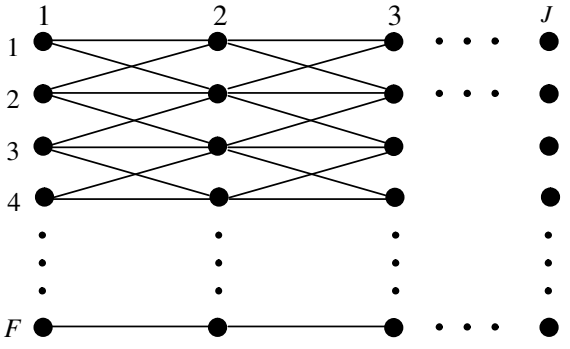


(a) Estructura *trellis* completa.

(b) Estructura *trellis* para $S_{max} = 3$.



(c) Estructura *trellis* para $S_{max} = 2$.



(d) Estructura *trellis* para $S_{max} = 1$.

Figura B.1: Representación de los cuatro caminos óptimos que se obtienen del ejemplo presentado en la figura on $S_{max} = 1$.

$$(F - 2 \cdot S_{max}) \cdot (2 \cdot S_{max} + 1)^{(J-1)} + 2 \sum_{n=1}^{S_{max}} (S_{max} + n)^{(J-1)}, \quad F \geq 2 \cdot S_{max} \tag{B.2}$$

para el caso de transiciones restringidas según S_{max}, donde F representa el número de filas de la estructura *trellis*, en este caso, $F = (T_{0\,max} - T_{0\,min} + 1)$. En la expresión (B.2) el primer término del sumatorio computa el número de estados capaces de generar en cada paso $2 \cdot S_{max} + 1$ conexiones (todo el abanico de posibilidades según la restricción de pendiente máxima S_{max}), mientras que el segundo término corresponde a los estados que pueden generar $(S_{max} + n) < (2 \cdot S_{max} + 1)$ transiciones entre columnas consecutivas. Aunque esta

Tabla B.1: GER (%) medio para todos los locutores (M1 a M5 y F1 a F5) del corpus *Keele*.

MÉTODO	RAPT		YIN		SHRp		*Media*	
Ventana	$5ms$	$10ms$	$5ms$	$10ms$	$5ms$	$10ms$	$5ms$	$10ms$
Configuración	9.37		5.71		17.46		10.85	
+ PMFAs13	4.83	7.96	5.57	*9.74*	6.51	10.30	5.63	9.33
+ PMFAs24	3.13	4.42	3.78	5.23	**5.04**	6.27	3.98	5.31
+ PMFAs26	3.21	4.48	3.80	5.38	5.17	6.36	4.06	5.41
+ PMFAs34	2.65	3.42	**3.03**	4.04	5.20	4.96	**3.62**	4.14
+ PMFAs37	3.12	3.50	3.35	4.25	7.02	5.09	4.50	4.28
+ PMFAs48	**2.63**	3.28	3.18	3.49	5.31	4.86	3.71	3.88
+ PMFAs68	2.92	2.94	3.23	3.31	6.29	**4.76**	4.14	**3.67**
+ PMFAs79	3.15	**2.92**	3.34	3.38	7.09	4.97	4.53	3.76
+ PMFAs912	3.63	3.10	3.67	**3.26**	8.14	5.60	5.15	3.99

expresión se ha deducido por inducción para $F \geq 2 \cdot S_{max}$, normalmente se trabaja con $F \gg 2 \cdot S_{max}$ para el algoritmo descrito en el presente capítulo —es decir, el margen de periodicidades es mucho mayor que el margen de variación trama a trama permitido. No obstante, la formulación sirve para cualquier combinación que cumpla la restricción. Por ejemplo, para $F = 6$ y $J = 3$, con $S_{max} = 3$ se obtienen 154 caminos posibles, de los cuales no hay ninguno que consiga cubrir $2 \cdot S_{max} + 1$, al ser $F < 2 \cdot S_{max} + 1$. No obstante, desde los estados intermedios se obtienen $2[(3+1)^2 + (3+2)^2 + (3+3)^2] = 154$ caminos. Para el caso de trabajar con $S_{max} = 2$, para $F = 6$ y $J = 3$, se obtienen 100 caminos, con $(6 - 2 \cdot 2)(2 \cdot 2 + 1) = 50$ caminos que permiten $(2 \cdot S_{max} + 1)^1 = 5$ conexiones entre tramas y $2[(2+1)^2 + (2+2)^2] = 50$ que permiten 3 o 4 conexiones. Finalmente, si se trabaja con $S_{max} = 1$ también para $F = 6$ y $J = 3$, se construyen 44 caminos, debidos a $(6 - 2 \cdot 1)(2 \cdot 1 + 1) = 12$ transiciones que dan lugar a $(2 \cdot 1 + 1) = 3$ transiciones de $J = 2$ a $J = 3$ (36 caminos en total), más 8 caminos ligados a los estados extremo (1 y 6), es decir $2[(1+1)^2] = 8$ caminos.

B.2. PMFA sobre *Keele database*

A continuación se presenta el barrido completo del estudio del funcionamiento de la propuesta de método de ajuste robusto de marcas de *pitch* para el corpus Keele, dados los algoritmos de referencia utilizados en el presente trabajo de investigación (ver capítulo 4).

Tabla B.2: GER (%) para los locutores masculinos (M1 a M5) del corpus *Keele* con una ventana de análisis de PMFA de $5ms$.

Método	*M1*	*M2*	*M3*	*M4*	*M5*	*Media*
RAPT	22.93	17.42	4.72	14.29	8.33	13.54
RAPT + PMFAs13	18.01	8.02	1.37	6.00	*10.99*	8.88
RAPT + PMFAs24	14.46	**4.65**	1.37	2.84	5.45	5.75
RAPT + PMFAs26	14.36	5.02	1.51	2.91	5.69	5.90
RAPT + PMFAs34	12.28	5.15	0.89	**2.47**	**3.10**	4.78
RAPT + PMFAs37	11.27	7.02	1.10	3.02	**3.10**	5.10
RAPT + PMFAs48	11.28	5.14	**0.82**	2.54	3.53	**4.66**
RAPT + PMFAs68	**10.53**	5.75	1.37	2.90	3.38	4.79
RAPT + PMFAs79	10.92	6.87	0.96	3.33	3.15	5.05
RAPT + PMFAs912	11.78	7.15	1.30	4.38	3.91	5.70
YIN	12.02	17.47	1.85	7.62	6.89	9.17
YIN + PMFAs13	*18.94*	8.10	1.30	*7.94*	*15.13*	10.28
YIN + PMFAs24	*15.20*	6.44	1.58	3.90	*7.32*	6.88
YIN + PMFAs26	*15.35*	6.37	1.51	3.96	*7.32*	6.90
YIN + PMFAs34	11.72	**4.48**	**0.89**	**2.60**	6.19	**5.18**
YIN + PMFAs37	**9.45**	7.85	1.37	4.05	4.54	5.45
YIN + PMFAs48	10.07	6.25	1.44	4.00	5.03	5.36
YIN + PMFAs68	10.03	7.36	1.44	3.76	4.20	5.36
YIN + PMFAs79	9.68	7.56	1.37	3.81	4.35	5.35
YIN + PMFAs912	10.41	9.16	1.65	4.18	**3.87**	5.86
SHRp	29.30	21.29	16.91	24.97	25.37	23.57
SHRp + PMFAs13	23.22	7.95	**1.37**	8.66	16.59	11.56
SHRp + PMFAs24	15.37	7.95	2.26	**5.26**	**12.53**	**8.68**
SHRp + PMFAs26	15.53	7.87	2.40	5.50	12.87	8.84
SHRp + PMFAs34	13.94	7.65	2.33	7.17	12.67	8.75
SHRp + PMFAs37	11.27	8.55	4.73	12.93	14.54	10.40
SHRp + PMFAs48	12.07	7.80	2.95	7.85	12.73	**8.68**
SHRp + PMFAs68	**10.76**	**7.12**	3.64	12.12	14.34	9.60
SHRp + PMFAs79	11.28	8.55	4.73	12.62	14.49	10.33
SHRp + PMFAs912	11.89	9.07	4.94	12.12	15.57	10.72

Tabla B.3: GER (%) para las locutoras femeninas (F1 a F5) del corpus *Keele* con una ventana de análisis de PMFA de $5ms$.

Método	*F1*	*F2*	*F3*	*F4*	*F5*	*Media*
RAPT	6.62	4.29	5.44	7.68	2.01	5.21
RAPT + PMFAs13	0.61	1.18	0.54	1.05	0.49	0.78
RAPT + PMFAs24	0.61	0.32	0.41	**0.88**	**0.27**	**0.50**
RAPT + PMFAs26	0.68	0.32	0.41	0.99	0.36	0.51
RAPT + PMFAs34	0.61	0.43	**0.20**	0.93	0.44	0.52
RAPT + PMFAs37	0.68	0.86	1.01	2.65	0.44	1.13
RAPT + PMFAs48	**0.54**	**0.32**	**0.20**	1.57	0.38	0.60
RAPT + PMFAs68	0.75	0.96	1.02	2.14	0.38	1.05
RAPT + PMFAs79	0.88	1.07	1.15	2.66	0.49	1.25
RAPT + PMFAs912	0.81	1.66	1.82	2.54	0.92	1.55
YIN	3.72	1.07	1.88	4.21	0.38	2.25
YIN + PMFAs13	**0.68**	*1.13*	0.61	1.58	0.27	0.85
YIN + PMFAs24	0.88	**0.43**	**0.47**	**1.28**	0.27	**0.67**
YIN + PMFAs26	0.88	0.48	**0.47**	**1.28**	0.33	0.69
YIN + PMFAs34	1.69	0.54	**0.47**	1.40	0.27	0.87
YIN + PMFAs37	2.35	0.54	0.74	2.21	*0.44*	1.26
YIN + PMFAs48	1.75	0.70	0.61	1.75	**0.22**	1.01
YIN + PMFAs68	1.95	0.86	**0.47**	1.87	0.33	1.10
YIN + PMFAs79	2.62	0.59	0.68	2.27	*0.44*	1.32
YIN + PMFAs912	2.75	0.70	1.08	2.32	*0.60*	1.49
SHRp	10.85	6.53	10.56	20.71	8.15	11.36
SHRp + PMFAs13	**0.82**	1.40	**0.75**	**3.33**	0.98	1.45
SHRp + PMFAs24	**0.82**	**0.64**	0.88	4.03	**0.71**	**1.41**
SHRp + PMFAs26	0.88	0.75	0.88	4.14	0.82	1.50
SHRp + PMFAs34	0.88	0.86	0.95	4.38	1.14	1.64
SHRp + PMFAs37	2.24	1.99	1.70	9.16	3.15	3.65
SHRp + PMFAs48	1.42	0.86	0.81	5.54	1.09	1.95
SHRp + PMFAs68	2.10	1.56	1.49	7.19	2.56	2.98
SHRp + PMFAs79	2.51	1.99	1.83	9.40	3.53	3.85
SHRp + PMFAs912	3.05	3.76	2.51	13.10	5.38	5.56

Tabla B.4: GER (%) para los locutores masculinos (M1 a M5) del corpus *Keele* con una ventana de análisis de PMFA de $10ms$.

Método	*M1*	*M2*	*M3*	*M4*	*M5*	*Media*
RAPT	22.93	17.42	4.72	14.29	8.33	13.54
RAPT + PMFAs13	*23.85*	14.24	2.47	9.71	18.31	11.18
RAPT + PMFAs24	16.70	6.67	**1.23**	4.64	10.11	5.66
RAPT + PMFAs26	16.54	6.90	**1.23**	4.45	10.40	5.75
RAPT + PMFAs34	14.03	5.17	1.58	4.70	5.95	4.35
RAPT + PMFAs37	14.29	4.87	1.58	4.82	6.34	4.40
RAPT + PMFAs48	14.04	**4.42**	1.65	3.65	6.07	3.95
RAPT + PMFAs68	12.55	5.15	1.30	2.78	3.73	3.24
RAPT + PMFAs79	11.72	4.92	1.23	**2.72**	3.67	**3.14**
RAPT + PMFAs912	**10.63**	5.22	1.30	2.96	**3.38**	3.22
YIN	12.02	17.47	1.85	7.62	6.89	9.17
YIN + PMFAs13	*25.12*	14.92	6.72	14.33	21.54	16.52
YIN + PMFAs24	*18.40*	7.20	2.47	6.34	12.74	9.43
YIN + PMFAs26	*18.63*	7.27	2.54	6.77	12.98	9.64
YIN + PMFAs34	16.87	5.55	1.58	4.19	8.24	7.28
YIN + PMFAs37	17.25	5.70	1.85	4.99	8.43	7.64
YIN + PMFAs48	12.92	**4.65**	1.85	4.06	7.07	6.11
YIN + PMFAs68	11.49	6.19	1.44	3.93	**3.33**	5.28
YIN + PMFAs79	11.10	6.02	**1.17**	4.12	4.54	5.39
YIN + PMFAs912	**11.00**	6.19	1.44	**3.93**	**3.33**	**5.18**
SHRp	29.30	21.29	16.91	24.97	25.37	23.57
SHRp + PMFAs13	23.38	*21.96*	2.33	13.67	23.64	17.00
SHRp + PMFAs24	17.78	9.67	**1.78**	9.77	16.30	11.06
SHRp + PMFAs26	18.11	9.75	**1.78**	9.59	16.50	11.14
SHRp + PMFAs34	15.52	8.17	1.92	**5.26**	12.43	8.66
SHRp + PMFAs37	16.02	8.17	2.06	5.50	12.53	8.86
SHRp + PMFAs48	15.91	7.65	2.88	5.44	10.52	8.48
SHRp + PMFAs68	13.01	**6.67**	2.47	7.42	**9.20**	**7.76**
SHRp + PMFAs79	**11.94**	7.05	2.61	7.79	10.23	7.92
SHRp + PMFAs912	12.11	6.97	2.95	9.09	**9.20**	8.06

Tabla B.5: GER (%) para las locutoras femeninas (F1 a F5) del corpus *Keele* con una ventana de análisis de PMFA de $10ms$.

Método	*F1*	*F2*	*F3*	*F4*	*F5*	*Media*
RAPT	6.62	4.29	5.44	7.68	2.01	5.21
RAPT + PMFAs13	**0.48**	1.83	1.29	2.51	4.95	2.21
RAPT + PMFAs24	0.61	1.45	0.54	1.46	0.76	0.96
RAPT + PMFAs26	0.54	1.56	0.95	1.40	0.82	1.05
RAPT + PMFAs34	0.61	0.70	**0.34**	**0.93**	0.22	**0.56**
RAPT + PMFAs37	0.61	0.86	**0.34**	1.11	0.22	0.63
RAPT + PMFAs48	0.81	**0.38**	0.47	1.11	**0.16**	0.59
RAPT + PMFAs68	0.88	0.59	0.41	1.74	0.22	0.77
RAPT + PMFAs79	1.42	0.70	0.54	2.09	0.22	0.99
RAPT + PMFAs912	1.75	1.34	0.54	3.11	0.76	1.50
YIN	3.72	1.07	1.88	4.21	0.38	2.25
YIN + PMFAs13	2.58	*2.47*	2.37	2.98	4.40	2.96
YIN + PMFAs24	**0.68**	1.18	0.68	2.33	**0.33**	1.04
YIN + PMFAs26	**0.68**	1.34	0.88	2.28	0.44	1.12
YIN + PMFAs34	0.88	0.70	**0.41**	**1.63**	**0.33**	**0.79**
YIN + PMFAs37	0.95	0.86	0.47	1.69	**0.33**	0.86
YIN + PMFAs48	1.08	**0.64**	**0.41**	1.81	0.38	0.87
YIN + PMFAs68	1.89	0.86	1.35	2.10	0.54	1.35
YIN + PMFAs79	1.69	1.29	1.28	2.04	0.60	1.38
YIN + PMFAs912	1.89	0.86	1.35	2.10	0.54	1.35
SHRp	10.85	6.53	10.56	20.71	8.15	11.36
SHRp + PMFAs13	3.60	1.61	1.22	6.01	5.55	3.60
SHRp + PMFAs24	**0.61**	1.45	0.88	3.50	0.92	1.47
SHRp + PMFAs26	0.75	1.56	1.15	3.33	1.14	1.59
SHRp + PMFAs34	0.82	0.91	**0.54**	3.38	**0.60**	1.25
SHRp + PMFAs37	0.88	0.97	0.61	3.38	0.82	1.33
SHRp + PMFAs48	1.09	**0.75**	0.68	**3.09**	0.60	**1.24**
SHRp + PMFAs68	0.95	0.91	0.75	4.55	1.69	1.77
SHRp + PMFAs79	1.22	1.13	0.88	5.19	1.63	2.01
SHRp + PMFAs912	1.96	1.83	1.22	8.32	2.34	3.13

Apéndice C

Herramientas e interfaces

En este apéndice se describen las distintas herramientas e interfaces diseñadas e implementadas a lo largo del presente trabajo de investigación. Asimismo, se presentan distintas aplicaciones en las que alguno de los sistemas desarrollados han tomado parte de forma más o menos directa. A continuación se describen, a grandes rasgos, sus características y elementos más particulares.

C.1. Interfaz de Tratamiento del Habla

Bajo el nombre de Interfaz de Tratamiento del Habla (en catalán *Interfície de Tractament de la Parla*, ITP) se ha desarrollado una aplicación que permite agrupar diferentes algoritmos y procesos relacionados con el etiquetado de un corpus de voz para síntesis concatenativa. Esta herramienta, programada inicialmente mediante Visual C++ 6.0©, partió de una interfaz generada originalmente en MATLAB ®, (Alías, 1999), que sólo permitía el etiquetado manual del corpus de voz.

ITP es una aplicación con el típico formato de ventanas (ver figura C.1), que incluye las siguientes funciones para el tratamiento de señales de voz y el diseño de corpus:

- Visualización y reproducción de la señal de voz (toda la señal, un tramo, etc.)
- Cálculo automático de las marcas de segmentación de la señal.
- Cálculo automático de las marcas de *pitch* de la señal.
- Visualización de la curva de *pitch*.
- Visualización del sonograma de la señal de voz (a partir de cálculo de la FFT de las tramas de la señal).
- Edición manual de las marcas de *pitch* y de segmentación, trabajando sobre la forma de onda de la señal.

- Consulta y modificación de la trascripción fonética de una frase.
- Cálculo de parámetros prosódicos de las unidades que forman la señal de voz: duración, energía, *pitch* medio y MFCC, a nivel de fonema, de unidad (difonema o trifonema) o alrededor del punto de concatenación.
- Algoritmo de selección las frases para el diseño de un corpus de voz, de un conjunto de pruebas, etc.

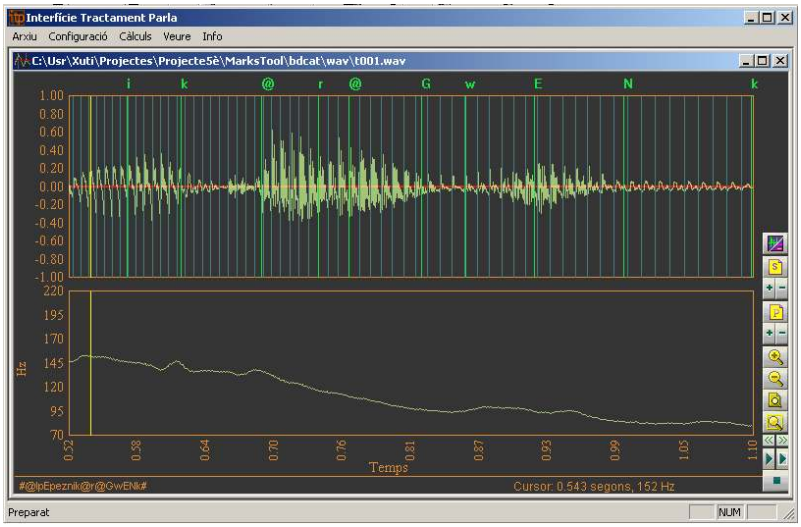


Figura C.1: Pantalla ejemplo de ITP. Se presentan distintas informaciones: les marcas de segmentación, las de *pitch*, la trascripción fonética y la curva de *pitch*.

Como se puede deducir de las principales funcionalidades que ofrece ITP, esta aplicación tiene como objetivo principal proporcionar al usuario (experto en tecnologías del habla) una herramienta completa, versátil y modular para el etiquetado completo de un corpus de voz destinado a la síntesis del habla. Además, ITP ha sido diseñada para que pueda ser ampliada en un futuro o utilizada tanto en su totalidad como sólo alguna de sus funcionalidades. Por esta razón, está formada por diversos módulos independientes, que pueden ser sustituidos por nuevos módulos.

Marcas de Segmentación

Esta librería permite etiquetar un fichero de voz con sus marcas de segmentación, de manera automática. Las marcas de segmentación son las encargadas de delimitar las unidades que componen la señal de voz, en este caso, a nivel de fonemas. De esta manera, se determina la posición temporal de cada unidad de la frase del corpus para su posterior recuperación durante el proceso de síntesis y para la extracción de sus características prosódicas, como por ejemplo, su duración.

Este módulo dispone de un algoritmo basado en los Modelos Ocultos de Markov (*Hidden Markov Modelos, HMM*). Inicialmente, se basaba en la aplicación de los modelos discretos (DHMM) (Alías y Iriondo, 2001b), pero gracias a la incorporación de la librería *Hidden Markov Modelos ToolKit* (HTK, 2001), actualmente el segmentador automático funciona mediante modelos continuos (CHMM), más precisos que los discretos. Sea cuál sea la opción escogida, el proceso de etiquetado se divide en dos fases: una primera, en la que se entrenan los modelos de cada una de las unidades del idioma (a partir de un grupo de frases etiquetadas y revisadas manualmente o de forma automática), y una segunda, donde se realiza la segmentación propiamente dicha de todo el corpus (fase de explotación del sistema).

En la actualidad, a pesar de ser un proceso automático que presenta un buen funcionamiento general, la revisión de las marcas de segmentación todavía resulta necesaria para evitar errores en el etiquetado que puedan afectar dramáticamente a la calidad de la señal sintética. A pesar de ello, se consigue acelerar notablemente el proceso de segmentación de los corpus de voz para selección de unidades, con respecto al proceso manual equivalente. ITP permite, pues, tanto el etiquetado automático como el proceso de revisión, generalmente realizado por un experto.

Marcas de *pitch*

En la aplicación ITP, la librería que calcula el *pitch* de la señal de voz parte de un algoritmo basado en dos conceptos fundamentales: la energía de la señal de la voz y el posicionamiento óptimo de las marcas mediante programación dinámica restringida (Alías y Iriondo, 2001a). En este caso, el sistema sólo precisa de las muestras de señal de voz para obtener las marcas de *pitch* de forma automática. No obstante, actualmente, se está trabajando con el algoritmo de ajuste robusto de marcas de *pitch* —*Pitch Marks Filtering Algorithm* (PMFA)— descrito en el capítulo 4, que en breve será incorporado a la interfaz. Por el momento, se ha desarrollado una interfaz *web* sobre la que aplicar el PMFA desarrollado.

CorpusTester

El etiquetado robusto del corpus de voz de los sistemas de conversión de texto en habla basado en selección de unidades es fundamental para evitar la presencia de problemas durante la fase de la síntesis. Debido al gran tamaño de este tipo de corpus, se ha desarrollado una herramienta que permite detectar las inconsistencias y errores de etiquetado del corpus para comprobar de forma automática la fiabilidad del etiquetado automático de un corpus de voz. Esta herramienta permite constatar la estructura del corpus, la trascripción fonética, las marcas de segmentación (delimitación de fonemas) y las marcas de *pitch*. Entre otras utilidades, da información de las marcas de segmentación excesivamente cercanas (se define un umbral, en este caso, de $20ms$), informa de las zonas donde las marcas de *pitch* presentan saltos bruscos (p.ej. se permiten transiciones del orden de $\pm 30\,\%$ entre la periodicidad local de marcas consecutivas) y controla que los fonemas no presenten valores de F_0 fuera de un

rango establecido (p.ej. [50, 550]Hz, como se explica en el capítulo 4) y contengan, como mínimo 2 marcas de segmentación (prerrequisito del módulo de síntesis basado en PSOLA utilizado (Iriondo et al., 2003)). Toda esta información se guarda en unos ficheros de salida que el revisor experto utiliza para catalogar los ficheros a revisar, evitando tener que revisar todo el corpus manualmente.

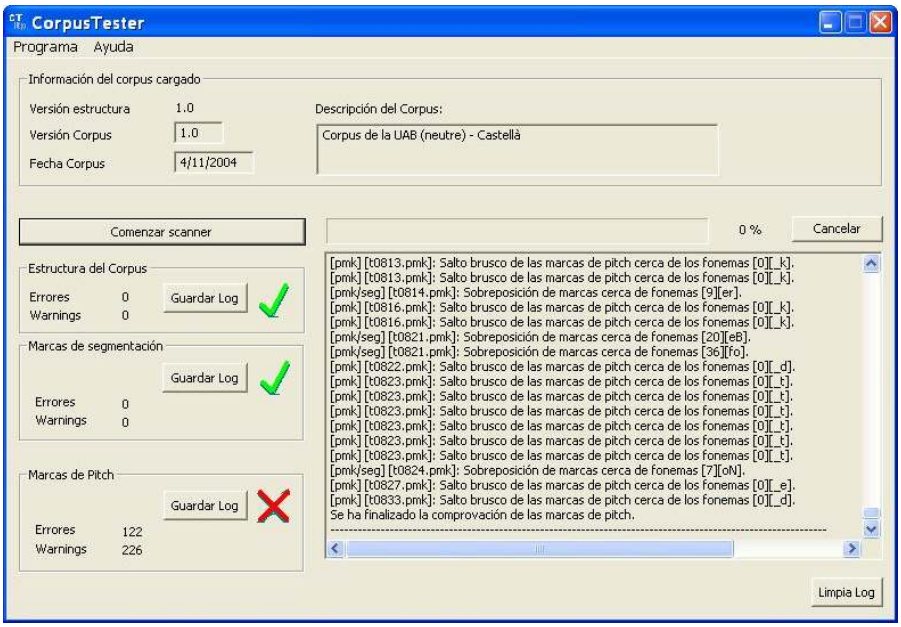


Figura C.2: Ejemplo del funcionamiento del *CorpusTester*. Se presentan distintas informaciones: análisis de la estructura del corpus, análisis de las marcas de segmentación y de las de *pitch*, presentando los errores en pantalla.

Extracción de parámetros prosódicos

El cálculo de los parámetros prosódicos se puede hacer tanto por fonemas como por difonemas o trifonemas. La segmentación en difonemas o trifonemas, actualmente, se realiza mediante un reparto simple de la duración del fonema en partes proporcionales. Para el caso de los difonemas (o trifonemas) el cálculo prosódico se puede realizar sobre toda la unidad, la mitad izquierda o la mitad derecha. De este modo se puede obtener la información prosódica que requiere el módulo de selección de unidades: los valores medios serán consultados para calcular los subcostes de *target* y los valores locales (izquierdos y derechos) serán necesarios para obtener los subcostes de concatenación (ver sección 2.1.2 para una explicación detallada de los conceptos de *subcoste* y de *selección de unidades*).

A continuación, se describe brevemente el proceso de cálculo de los parámetros prosódicos que contempla la ITP. Estos cálculos se pueden realizar tanto para un fichero de voz (por ejemplo, una frase), como para todos los ficheros que forman parte de la base de datos

analizada.

- La duración se obtiene, simplemente, como la resta entre la posición de las marcas de segmentación que delimitan la unidad a parametrizar. Se obtienen del fichero correspondiente a la frase analizada, escogiendo, en cada caso, las marcas adecuadas, según si se trabaja a nivel de fonema o a nivel de difonema o trifonema.

- Para obtener el valor del *pitch* medio, se parte de las marcas de la posición de las marcas de *pitch* (calculadas con el módulo automático) que corresponden a la unidad parametrizada (información extraída de las marcas de segmentación). La diferencia entre dos marcas de *pitch* consecutivas aporta el valor de *pitch* local a lo largo del fonema (o difonema y trifonema, según haga falta). A continuación, se promedian todos los valores para obtener el periodo medio (T_0). Si la información se quiere representar en el dominio frecuencial, habrá que tener en cuenta que $F_0 = 1/T_0$.

- El cálculo de la energía es muy sencillo, ya que sólo se utilizan las muestras de la señal de voz correspondientes a la unidad parametrizada. Se parte de las marcas de segmentación que limitan la unidad parametrizada y se calcula la raíz cuadrada del sumatorio del valor absoluto de las muestras de voz al cuadrado. A continuación, este valor se normaliza respecto de la duración del fonema (difonema o trifonema, según el caso).

Algoritmo de selección de frases

ITP también incorpora un sistema de selección de frases (a partir del texto) para el diseño del contenido de corpus (para síntesis basada en selección de unidades, para pruebas subjetivas, etc.). Este sistema está basado en un algoritmo *greedy* (van Santen y Buchsbaum, 1997; François y Boëffard, 2002) que escoge las frases a partir de unas determinadas especificaciones (número de fonemas, posición del fonema en la frase, acentuación, etc.). Este método permitirá escoger las frases a considerar en el corpus a fin de que éste presente una cobertura óptima para el que debe ser usado (p.ej. el dominio de funcionamiento del sintetizador – de propósito general o limitado).

En cuanto a los sistemas de síntesis de voz por concatenación basados en selección de unidades es esencial disponer de un corpus de voz con la máxima variabilidad prosódica, lingüística y fonética. La situación *ideal* sería poder disponer de un corpus que contuviera *todas* las realizaciones sonoras de un locutor para un idioma, pero esta posibilidad, evidentemente es inviable. Por esta razón, el corpus de voz se tiene que diseñar a partir de un conjunto finito de frases seleccionadas cuidadosamente (van Santen y Buchsbaum, 1997; Black y Lenzo, 2001b). El proceso de selección parte de una colección de frases (corpus de entrada) a la que se aplica un algoritmo encargado de elegir el subconjunto (corpus de salida) más adecuado. Es decir, un grupo de frases que consigan cubrir los requisitos exigidos según los criterios escogidos durante el diseño del corpus de voz (ver figura C.3). Este mismo proceso es aplicable para el diseño de corpus con un número de frases menor. Por ejemplo,

las pruebas subjetivas que se llevan a cabo en este trabajo de investigación hacen uso de este algoritmo para diseñar el grupo de frases que deben formar parte de las mismas.

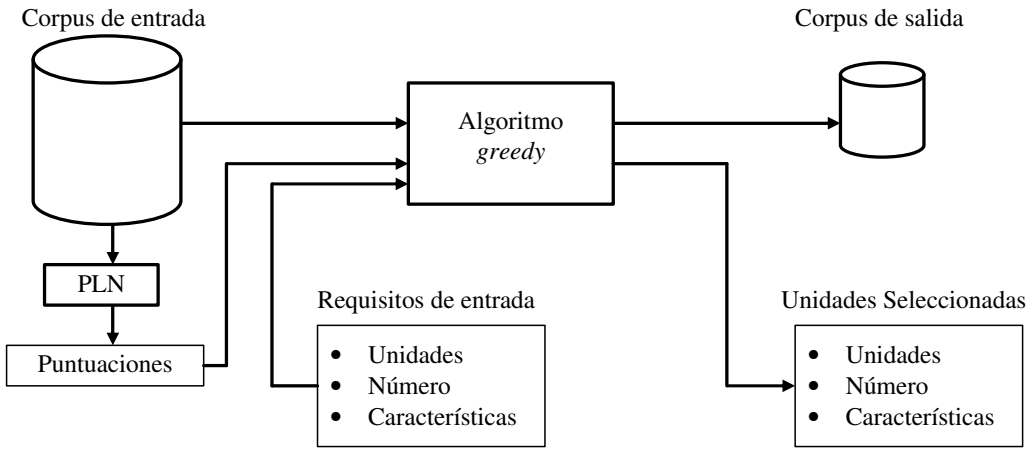


Figura C.3: Diagrama de bloques del proceso de selección de les frases del corpus de voz mediante un algoritmo *greedy*.

El algoritmo escoge, en cada paso del proceso iterativo, una frase del corpus de entrada para pasar a formar parte del corpus de salida. Esta frase es la que obtiene una mejor *puntuación* de entre las frases del corpus inicial. Una vez escogida, se contabilizan las unidades de la frase escogida (a nivel de fonemas o de difonemas y trifonemas) y se comprueba si ya se han alcanzado los requisitos de diseño del corpus. La puntuación de cada frase se obtiene a partir del módulo de procesamiento del lenguaje natural (PLN) de las frases del conjunto de entrada del algoritmo. El módulo de PLN se encarga de extraer las características de cada unidad de la frase que mejor permiten estimar los atributos considerados (prosódicos, lingüísticos y fonéticos), entre otros, la posición en la frase, la acentuación, la trascripción fonética, reglas para la estimación de la duración de la unidad, ... Estas características (Chu et al., 2001) serán las utilizadas para definir los requisitos y para comprobar, durante la ejecución del algoritmo, si estos requisitos se van cumpliendo o no. Actualmente, se están estudiando otros parámetros a considerar, en cuanto a número y tipo. La salida del algoritmo será un conjunto de frases, subconjunto de la base de datos inicial, que contiene el conjunto de unidades que mejor cumplen los requisitos (no siempre se consigue dar cumplimento a todas las especificaciones de partida debido al carácter finito del corpus), tanto en lo que se refiere al número como la cobertura de las características requeridas (Chu et al., 2001).

Así pues, el funcionamiento, a grandes rasgos, del algoritmo implementado hasta el momento (ver figura C.3) es el siguiente:

- Como parámetros de entrada recibe el corpus de frases original y los requisitos de

diseño, en cuanto al número de repeticiones por cada unidad y sus características: hasta el momento, sólo se considera la de posición de la unidad en la frase (al principio, en el medio o al final). Esta información será ampliada próximamente mediante los datos que se pueden obtener del módulo de PLN del CTH (ver sección 2.1.2) y la información aportada por lingüistas expertos.

- Iterativamente, el algoritmo escoge las "mejores" frases según el grado de satisfacción de los requisitos fijados para el diseño del corpus de voz.

- A medida que se añaden frases al subconjunto de salida, se van completando los requisitos de partida.

- En el momento que se cumplen todos los requisitos, el algoritmo se detiene, obteniendo el subconjunto de frases que tienen que constituir el corpus de voz. Sin embargo, el algoritmo presentará la lista de las unidades correspondientes a las frases escogidas junto con sus características.

C.2. Plataforma para el ajuste subjetivo de pesos

En este apartado se describe, a grandes rasgos, la plataforma desarrollada para el ajuste de pesos mediante evaluación subjetiva[1]. Se escoge trabajar con una plataforma *web*, como en (Jilka y Syrdal, 2002; Black y Tokuda, 2005), con el fin de ofrecer un acceso cómodo y amigable a los usuarios de la aplicación, junto a una gestión automática de los resultados. Las especificaciones a partir de las que se diseñó la plataforma fueron:

- **Autenticación:** el acceso a las pruebas es controlado. El administrador es el que decide qué usuarios tienen acceso al test, definiendo su perfil de usuario (descrito a continuación). De esta manera se dispone de un control estricto de los resultados obtenidos, ya que un acceso no restringido podría provocar resultados ruidosos.

- **Usuario:** se define el perfil de la persona que interactúa con la plataforma. En principio sólo hay dos tipos de usuarios: administrador (con acceso y control total) y evaluador (con acceso sólo a las pruebas).

- **Niveles:** las pruebas estarán divididas en diferentes grupos de usuarios según su experiencia en el ámbito de las tecnologías del habla. Es decir, la plataforma permite estudiar el ajuste de los pesos de selección según los conocimientos de los usuarios sobre las posibilidades de los CTH-SU actuales (en cuanto a calidad de voz). En principio se dispone de tres niveles: *inexperto*, *medio* y *experto*. La asignación de cada usuario a un nivel determinado será trabajo del administrador de la plataforma de test.

[1] http://www-sinev.salle.url.edu

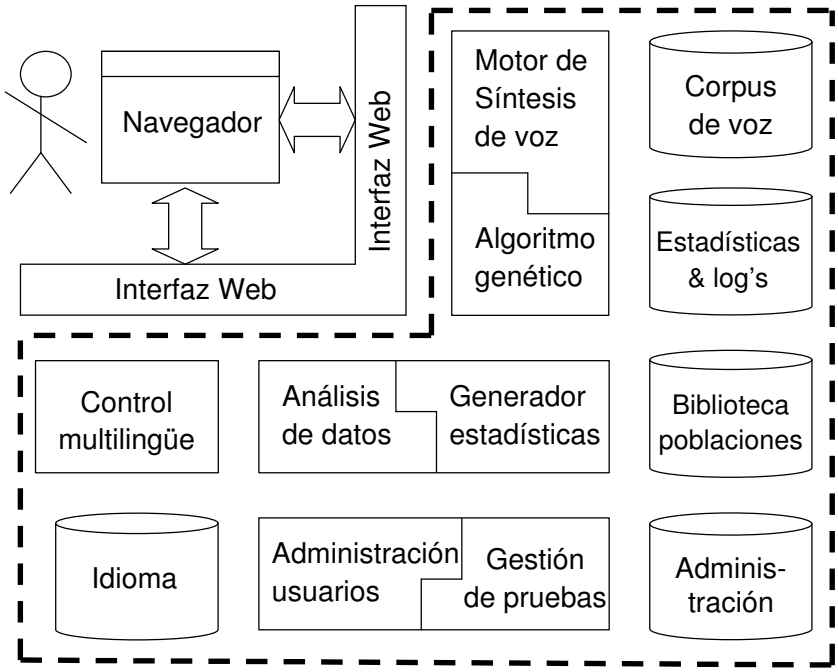


Figura C.4: Diagrama de los módulos y bases de datos que constituyen la plataforma interactiva de ajuste de pesos desarrollada.

- **Distribuida:** el diseño permite que la aplicación pueda funcionar sobre varios servidores, aumentando su capacidad de servir peticiones (síntesis solicitadas durante las pruebas).
- **Concurrente:** la plataforma debe permitir la coexistencia de distintas sesiones ejecutándose al mismo tiempo.

A continuación se detalla el diseño de la plataforma y todos los módulos que la componen, así como las diversas tecnologías que intervienen:

- **Diseño de la plataforma:** se ha desarrollado una plataforma que dispone de diferentes módulos de proceso, acción y gestión. Entre otros, la plataforma dispone de una interfaz de usuario que permite el acceso remoto a las pruebas (vía *Internet*), un sistema para la síntesis del habla o un algoritmo genético interactivo. En la figura C.4 se presentan todos estos módulos, junto con los diferentes corpus y bases de datos que los acompañan. Concretamente, la plataforma dispone de un corpus de voz que contiene las unidades (difonemas y trifonemas) para la selección de unidades y su posterior síntesis, y la base de datos que contiene información de las pruebas realizadas, así como de otras informaciones referentes a la gestión de la plataforma. A continuación, se describen todos estos módulos, indicando, sus funcionalidades básicas en la plataforma.

- *Interfaz web:* se ha desarrollado mediante un conjunto de *scripts* que cargan distintas plantillas *web* en función del perfil (privilegios) de cada usuario (administrador o evaluador), el idioma de trabajo (catalán, castellano o inglés) y la acción que se está realizando en cada momento (análisis, realización o administración de las pruebas).

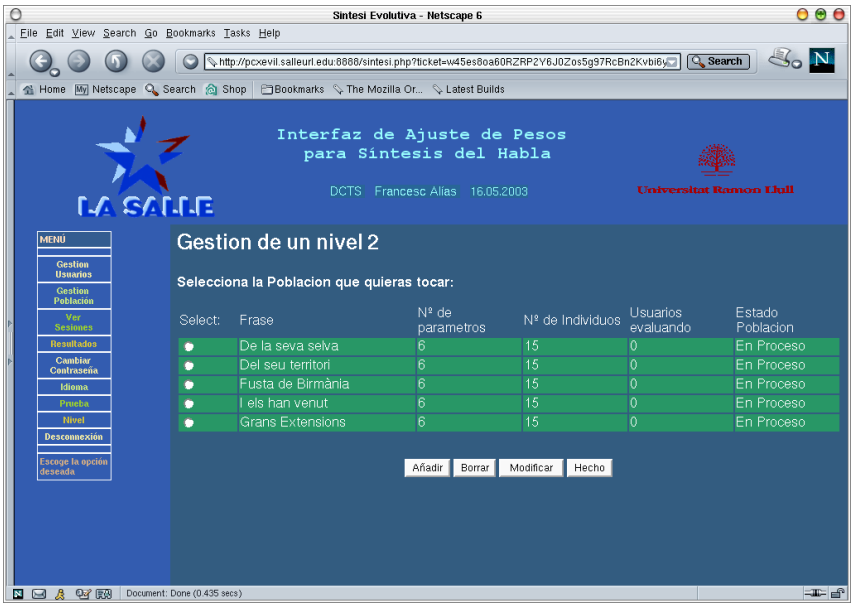


Figura C.5: Pantalla de *Gestión de pruebas* de la plataforma.

- *Motor de síntesis:* Está constituido por el bloque de procesamiento digital de la señal del conversor de texto en habla basado en selección de unidades desarrollado durante el presente trabajo de investigación (ver sección 2.1.2). Este módulo parte de la información prosódica del texto de entrada y, mediante la elección de las unidades óptimas en términos de la función de coste (guiada por el valor de los pesos de los subcostes), escoge la secuencia de difonemas y trifonemas del corpus (incorporado también en la plataforma) a sintetizar. A continuación, estas unidades son concatenadas para generar los ficheros de voz correspondientes a los pesos indicados por el algoritmo genético interactivo.
- *Algoritmo genético:* se trata del algoritmo genético interactivo descrito en el apartado 2.2.3, que almacena el estado del ciclo evolutivo en una base de datos relacional tipo MySQL[2].Esta información es accesible por el administrador, que puede disponer de la historia del proceso evolutivo que ha seguido la población según el criterio de cada usuario. Recientemente, se ha colaborado en el desarrollo de nuevas técnicas para poder controlar visualmente la consistencia de los usuarios a lo largo de la interacción (Llorà et al., 2006), con el objetivo de

[2]http://www.mysql.com

incorporarlas en la plataforma.

- *Análisis de datos:* la plataforma incorpora un bloque que permite analizar de forma gráfica los resultados obtenidos a lo largo de las diferentes pruebas realizadas mediante distintas estadísticas: histogramas, correlaciones, etc. Los resultados pueden evaluarse a nivel de prueba (para una frase o expresión), a nivel de usuario (para todas las frases de una prueba), para un grupo de usuarios (dentro de un tipo de prueba), e incluso, para todas las pruebas en global. De este modo, se pueden analizar las distintas tendencias dentro de cada uno de los grupos de datos que se acaba de describir. Este bloque de la plataforma todavía se encuentra en fase de desarrollo.
- *Generador de estadísticas:* es el encargado de obtener las estadísticas de uso de la plataforma. Por ejemplo, se pueden consultar: el número de accesos, las pruebas que se están activas, los archivos sintetizados, la memoria disponible de los servidores, etc. Esta información sólo podrá ser consultada por el administrador de la plataforma.

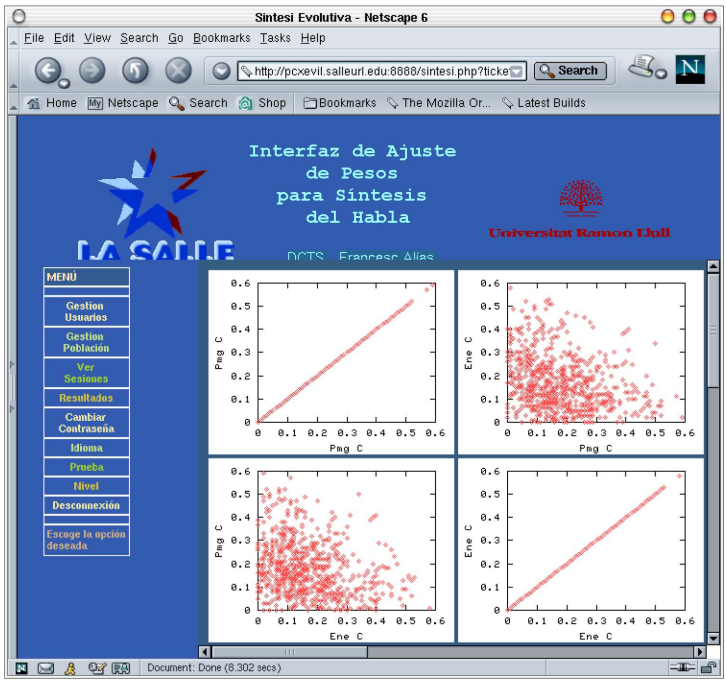


Figura C.6: Ejemplo de pantalla de análisis de resultados.

- *Gestión de pruebas:* mediante este módulo, el administrador puede gestionar todos los parámetros que intervienen en la configuración de las pruebas: el texto de la expresión sintetizada, la configuración del nivel de la prueba, los parámetros del algoritmo genético interactivo (p_c, p_m, etc.), ...
- *Administración de usuarios:* es el encargado de definir y gestionar a los usuarios

que tienen acceso a la aplicación. Además, permite asignar el perfil de usuario (idioma, nivel de prueba, etc.) entre otros parámetros configurables de la interfaz de usuario (p.e. la información presentada en la pantalla de las pruebas).

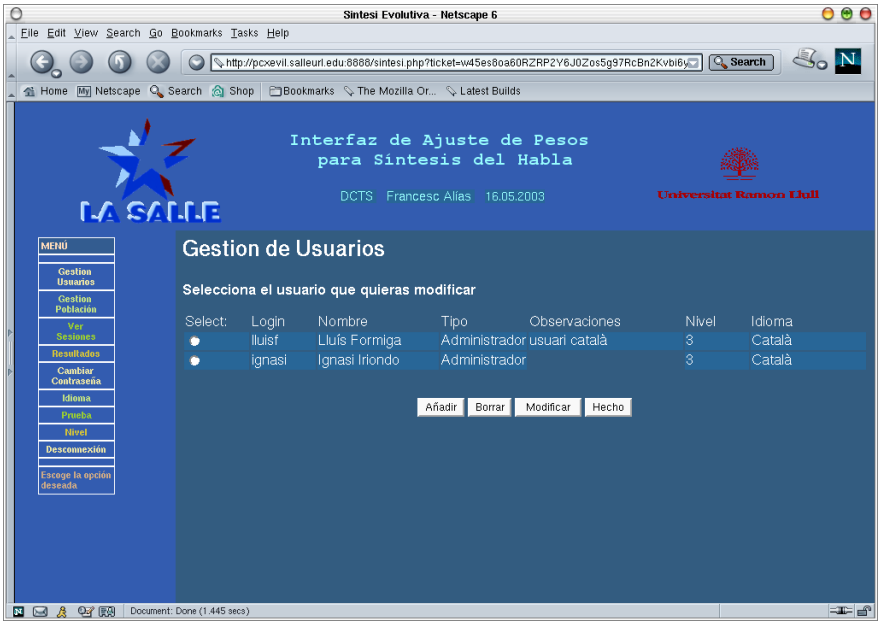


Figura C.7: Pantalla de *Gestión de usuarios* de la plataforma.

- **Tecnologías utilizadas:** a modo de breve reseña (ver (Formiga, 2003) para conocer más detalles), las tecnologías en las que se basa el funcionamiento de la plataforma son:

  - *CGI* (*Common Gateway Interface*) programadas mediante PHP para la interfaz *web*.
  - Base de datos relacional (MySQL, 2003) para almacenar toda la información que contiene la plataforma
  - Código en C ++ compilado con gcc (GNU C Compilier) sobre sistema operativo Linux, con el que se implementa el bloque de síntesis.

En las figuras C.5, C.6 y C.7 se presentan algunas de las pantallas de la plataforma implementada, a modo de ejemplo ilustrativo de la apariencia y el funcionamiento de la plataforma de test desarrollada. Como se puede observar, la interfaz se divide en dos *frames* independientes, uno que contiene la barra de menús (a la izquierda) y la cabecera (en la parte superior), y el otro, que varía su contenido según la acción activa en cada una de las figuras presentadas.

Apéndice D

Aplicaciones

D.1. Conversión de texto en habla meteorológica

Una de las aplicaciones más interesantes de los sistemas de CTH es su integración dentro de sistemas de comunicación persona-máquina, donde la CTH formará parte del canal de salida de la interfaz, encargándose de generar el mensaje oral que el usuario espera recibir del sistema. Esta sección del trabajo de investigación se presenta un sistema de CTH diseñado para una aplicación real, adaptando algunos de los resultados conseguidos de la investigación descrita en los apartados anteriores del presente trabajo. Concretamente, este CTH se ha desarrollado en el seno de un proyecto de investigación titulado "*Personajes Virtuales*" en el que participan la Corporación Catalana de Radio y Televisión (CCRTV), como líder del proyecto, el Grupo de Tecnologías Interactivas (GTI) de la Universidad Pompeu Fabra (UPF) y nuestro grupo de investigación. Este proyecto financiado en parte por la CCRTV y el CIDEM (Centre d'Innovació i Desenvolupament Empresarial)[1], tiene como principal objetivo la creación de un escenario que permita la generación de productos audiovisuales automáticos para diferentes entornos multimedia: televisión, Internet y dispositivos móviles. Estos productos están basados en personajes sintéticos animados dotados de habla sintética. Se pretende poner al alcance del diseñador multimedia las herramientas para crear personajes virtuales, dotarlos de movimiento, expresividad, incorporarlos en espacios virtuales y que, mediante un guión y unas descripciones, puedan hablar y moverse por un escenario virtual, etc.

Como primer paso de este ambicioso proyecto, se ha desarrollado una aplicación para dar un servicio de información meteorológica local, denominada *Meteorólogo Virtual* (Alías et al., 2005). En esta aplicación, un hombre del tiempo, ubicado dentro de un entorno virtual 3D, ofrece al usuario la previsión meteorológica para una ciudad concreta (seleccionada por el usuario) (ver figura D.1) —actualmente, da información de ciudades españolas y capitales europeas, sobre un total de 175 ciudades[2].

[1] Proyecto RDITSCON04-0005

[2] http://www.meteosam.com

Figura D.1: Aspecto visual de la aplicación del *Meteorólogo virtual*.

En este contexto, ha resultado necesario adaptar las investigaciones desarrolladas en el ámbito de la CTH basada en selección de unidades a las restricciones y especificaciones definidas por la aplicación. En este caso, por un lado, se buscó conseguir la mayor calidad (naturalidad) de la síntesis posible, con una velocidad de síntesis adecuada, y por otro, se adaptó el CTH para trabajar con textos meteorológicos generados de forma automática a partir de un conjunto finito de componentes de dominio limitado. Con estos objetivos en mente, se ha diseñado un sistema de CTH basado en selección de unidades que utiliza *i)* un corpus de voz rico prosódicamente, *ii)* recupera la prosodia del corpus —es decir, no se incluye predicción prosódica—, *iii)* utiliza una estrategia de síntesis basada en la concatenación de locuciones, por lo que se simplifica la función de coste, *iv)* se trabaja con un corpus multidominio, pero en este caso sin módulo de clasificación de textos, ya que los textos están etiquetado (controlados por la aplicación) y, finalmente, *v)* resulta necesario dotar de información temporal a la aplicación multimodal para conseguir la sincronización audiovisual entre el mensaje oral y la información visual correspondiente (p.ej. movimientos labiales). Por otro lado, debido a las particularidades del diseño realizado, resulta necesario afrontar situaciones distintas a las habituales, como por ejemplo, la unión de fragmentos de voz de tamaño considerable con valores prosódicos distantes en los puntos de concatenación. A continuación se describen todos los elementos del sistema de síntesis desarrollado utilizados para dar respuesta a las necesidades planteadas por la aplicación del *Meteorólogo virtual*.

Introducción

Esta sección del trabajo de investigación presenta un sistema de conversión de texto en habla (CTH) de dominio limitado pero capaz de sintetizar cualquier texto de entrada (Black y Lenzo, 2000; Schweitzer et al., 2003). El objetivo del CTH desarrollado es conseguir una señal sintética de alta naturalidad. Por un lado, es un CTH de dominio restringido ya que está orientado a una aplicación meteorológica, siguiendo una aproximación de concatenación de locuciones o *phrase-splicing* (ver sección 3.2.1). Por otro lado, es un CTH genérico ya que el corpus de voz contiene todos los difonemas (más algunos trifonemas) del castellano (en la línea de lo descrito en el apartado 3.4.1). Así pues, aunque el diseño y los contenidos del corpus de voz son totalmente orientados a aplicación, el sistema de CTH es capaz de sintetizar *cualquier* texto de entrada (con la consiguiente pérdida de calidad respecto a los textos orientados a la aplicación). Además, esta aplicación de síntesis meteorológica sigue la filosofía de la síntesis multidominio descrita en el capítulo 3 del presente trabajo de investigación, pues el corpus de voz está estructurado en distintos subcorpus, en este caso, para aumentar la velocidad del proceso de síntesis —no es necesario incorporar el algoritmo de clasificación automática de textos para seleccionar el dominio más adecuado para la síntesis ya que el CTH recibe como entrada los textos a sintetizar etiquetados, como se describe más adelante. En este caso, el corpus se divide en tres dominios: *i) bienvenida*, *ii) previsión* y *iii) despedida*, que son los bloques a partir de los que se construye el mensaje sintético.

La figura D.2 muestra de forma esquemática el diagrama de bloques del proyecto *Personajes Virtuales*, que está constituida principalmente por tres módulos: *i)* el Generador del Guión, que define la planificación de escenas, los movimientos y expresiones del personaje, los textos a pronunciar y las animaciones adicionales (desarrollado por la CCRTV y el GTI-UPF); *ii)* el Conversor de Texto en Habla, que además de sintetizar el mensaje añade la información temporal de los eventos relacionados con la parte gráfica y la sincronización labial; y *iii)* el Generador de las escenas 3D animadas, que genera el video final adaptado al dispositivo de salida seleccionado (desarrollado por el GTI).

En este contexto, el CTH desarrollado tiene que ajustarse a la arquitectura y las especificaciones de la aplicación diseñada. Para cumplir estos requisitos, se han incorporado dos interfaces que controlan el flujo de datos entre el CTH y los módulos con los que se comunica (ver apartado de "*Interfaces de entrada y salida*"). Además, el enfoque clásico de sistema de síntesis concatenativa basada en corpus se ha ajustado para conseguir voz sintética de alta calidad, como se describe en el apartado "*Motor de síntesis*". Finalmente, el sistema de CTH es evaluado objetiva y subjetivamente en los experimentos desarrollados.

La figura D.3 muestra el diagrama de bloques esquemático del sistema de CTH diseñado. Como se puede observar, éste está formado por: *i)* una interfaz de entrada constituida por un *parser* XML, que se encarga de procesar el texto de entrada que está etiquetado con una versión modificada de SSML (Speech Synthesis Markup Language) (ver apartado D.1), *ii)* el motor de síntesis, *iii)* el corpus de voz en castellano, y *iv)* la interfaz de salida, que incorpora la información temporal para la sincronización de la escena.

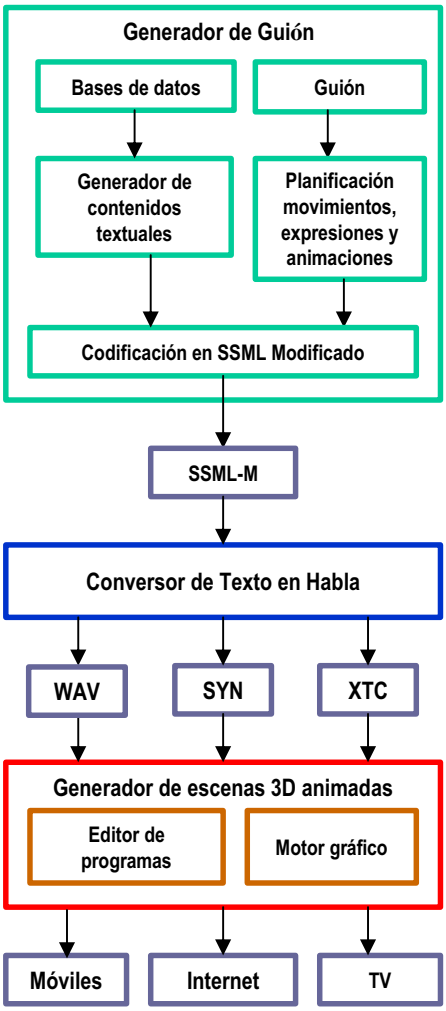


Figura D.2: Diagrama de bloques del proyecto *Personajes Virtuales*.

Como resultado, el CTH genera tres datos de salida: el fichero WAV, con las muestras de voz sintética, el fichero XTC (*XML Time Code*), con la información necesaria para sincronizar temporalmente los eventos (movimientos de cámara, gesticulaciones del locutor, etc.) con el mensaje oral (ver apartado D.1, y el fichero SYN, que contiene la información de sincronización labial (ver apartado D.1). Estos ficheros se traspasan al bloque de generación de escenas 3D para la obtención de los vídeos finales, tal y como se indica en la figura D.2.

Como muestra la figura D.3 el proceso de conversión de texto en habla del sistema empieza con el típico proceso de trascripción de texto a fonema basado en reglas (ver sección 2.1). A continuación el módulo de selección de unidades busca el conjunto de unidades óptimo para la síntesis (la unidad mínima de búsqueda es el difonema, acompañado por algún trifonema) (ver apartado D.1). Después de la selección de unidades, se recupera

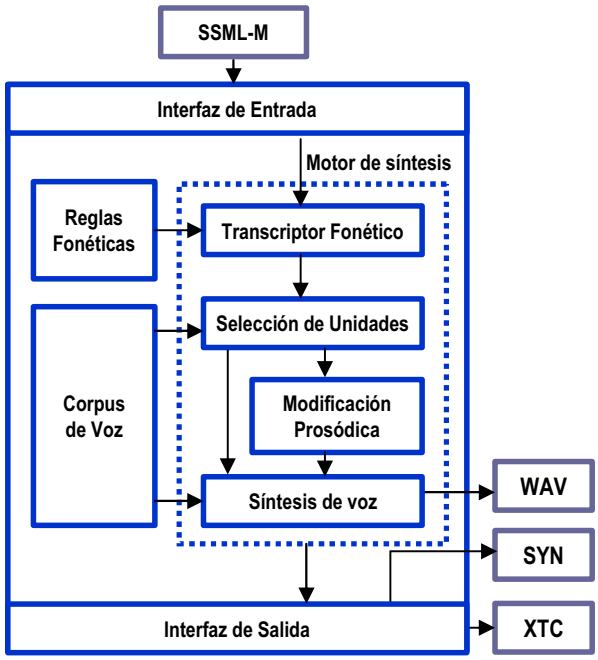


Figura D.3: Diagrama de bloques del CTH diseñado para la aplicación del *Meteorólogo virtual*.

la prosodia de las unidades escogidas como prosodia objetivo, ajustando las transiciones bruscas presentes en los puntos de concatenación mediante el proceso descrito en el apartado D.1. Finalmente, la señal de voz se genera utilizando un algoritmo tipo TD-PSOLA, como se describe en el apartado D.1.

Interfaces de entrada y salida

En este apartado se describen las interfaces de entrada y salida que se han incorporado a la arquitectura del CTH para interactuar con los bloques de Generación de guión y de Generación de escenas 3D, respectivamente (ver figuras D.2 y D.3).

Entrada de texto etiquetado mediante SSML modificado

Para satisfacer las necesidades de la aplicación del *Meteorólogo virtual*, se ha diseñado un lenguaje de marcas basado en SSML[3] con el objetivo de intercambiar datos entre los tres bloques principales de la aplicación (ver figura D.2). Este lenguaje SSML modificado corresponde al acrónimo SSML-M.

[3]Speech Synthesis Markup Language, Version 1.0 del W3C, http://www.w3.org/TR/speech-synthesis

Una de las particularidades de SSML-M es su capacidad de tratar con caracteres acentuados y especiales, cuestión esencial para el castellano —en contraposición con la versión SSML del inglés. Para evitar problemas de compatibilidad de formatos de codificación de caracteres, el analizador del documento SSML-M (*parser*) traducirá el código indicado en la cabecera del documento a un código interno del SSML-M, que, en este caso, será UTF-8. Asimismo, como valor añadido, esta estrategia asegura la portabilidad del sistema de CTH hacia distintos sistemas operativos.

El documento SSML-M está formado por dos partes: la cabecera y los bloques. La cabecera contiene la definición de las variables globales del video que especifican los parámetros del Generador de escenas 3D: el actor principal, el nombre de la ciudad o los símbolos de la temperatura, entre otros. Los bloques definen el contenido de las distintas componentes del video. Cada bloque está constituido por uno o más elementos <speak>. Estos elementos incluyen distintos atributos SSML así como *tokens* para extender la funcionalidad de SSML para describir los datos. Uno de los atributos incorporados a SSML es *corpus*, que se utiliza para indicar el subcorpus escogido del corpus multidominio para llevar a cabo la selección de unidades y la posterior síntesis. El resto de *tokens* incorporados en el documento SSML-M se utiliza para sincronizar los eventos visuales de escena, como por ejemplo, tomas de cámara o movimientos del personaje, con la señal de voz. Asimismo, se incorpora el atributo *cache* que, si está activo, permite al CTH recuperar un fichero de voz previamente generado, en lugar de tener que resintetizarlo de nuevo. El objetivo es acelerar el proceso de generación de la previsión meteorológica. Finalmente, y siguiendo el formato estándar de SSML, se puede modificar la prosodia de la señal de voz mediante los atributos *pitch*, *speed* y *volume*, dentro del elemento <prosody>. Concretamente, la información prosódica de las unidades puede ajustarse de forma relativa (p.ej. <prosody pitch="high"speed="slow"volume="high">).

Sincronización de eventos

Debido a que el sistema de CTH forma parte de una aplicación multimodal, es necesario sincronizar los eventos visuales con la señal oral sintética. Esta sincronización se obtiene mediante el código de tiempo descrito en el fichero XTC. Este fichero es simplemente una versión ampliada del fichero SSML-M de entrada, incluyendo la información para sincronizar temporalmente los elementos audiovisuales de la escena. Para ello, se incorporan los atributos *begin* y *end* dentro de los elementos <speak>, así como, a todos los elementos inferiores a <speak> dentro de la jerarquía XML (p.ej. los eventos audiovisuales incrustados en su interior). Además, el fichero XTC incorpora la ubicación completa de los ficheros de audio y sus correspondientes ficheros de sincronización labial como elementos adicionales dentro de los atributos <speak> (nótese como cada <speak> está asociado a una pareja de ficheros WAV y SYN). Así pues, cada fichero SSML-M está relacionado con un fichero XTC (ver ejemplo de la tabla D.1).

Tabla D.1: Ejemplo de una sección del contenido del fichero XTC para una previsión meteorológica concreta.

```
<speak actor = "main_actor"  voice_id="SPANISH"
cache="true"  corpus="welcome"
wav_file="f:/out/ORAL_TV_DAY_BARCELONA01.wav"
lypsync_file="f:/out/ORAL_TV_DAY_BARCELONA01.syn"
begin="0.000"end="4.941">
```

Sincronización labial

El personaje virtual debe sincronizar los movimientos labiales con el habla que produce para dar realismo al mensaje audiovisual generado. La información de sincronización labial se almacena en un fichero independiente del XTC, denominado SYN, para evitar sobrecargar el fichero XTC con demasiada información. El fichero SYN incorpora la información fonética de los sonidos producidos mediante el formato SAMPA, a la que acompaña la duración del fonema, siguiendo la misma estructura (*begin*/*end*) utilizada para el fichero XTC (ver ejemplo de la tabla D.2).

Corpus de voz

A continuación se describe el corpus de voz diseñado para este CTH, enfatizando sus particularidades respecto a otros corpus de dominio limitado. Concretamente, el diseño del contenido del corpus se ha llevado a cabo a partir de la base de datos que utiliza el Generador de Guión (ver figura D.2 para construir los contenidos textuales de las previsiones.

Un aspecto a tener en cuenta en el diseño del corpus es que el espacio de textos a sintetizar está dividido en tres bloques que corresponden a los diferentes elementos que constituyen el mensaje meteorológico: *i) bienvenida*, que hace referencia a la ciudad escogida; *ii) previsión*, que consta de la previsión del día actual y del día siguiente, pudiéndose dividir cada una de ellas en mañana, tarde y noche; y *iii) despedida*, que además puede incluir algún mensaje promocional. Para ello, se han grabado todas las expresiones utilizadas por

Tabla D.2: Ejemplo del contenido del fichero SYN para el inicio de una previsión meteorológica, en este caso, iniciada por la palabra "*Hola*".

```
< lip_sync>
<phoneme id="_"  begin="0.000"end="0.142/>
<phoneme id="o"  begin="0.142"end="0.254/>
<phoneme id="l"  begin="0.254"end="0.294/>
<phoneme id="a"  begin="0.294"end="0.526/>
<phoneme id="_"  begin="0.526"end="0.700/>
</ lip_sync>
```

el Generador de contenidos textuales automático con cierta variabilidad prosódica. Además, el corpus incluye todos los difonemas del castellano (aumentados con algunos trifonemas) para poder sintetizar *nuevas* palabras (p.ej. nombres de ciudades) o transiciones de unidades (no presentes, de entrada, en las combinaciones de expresiones grabadas).

Diseño del corpus

El diseño del corpus de voz es uno de los elementos clave en la construcción de un sistema de CTH (ver apartado 2.1.2). En este caso, se ha basado en conseguir la máxima cobertura fonética y léxica dentro del dominio de la aplicación, es decir, sobre las expresiones utilizadas por el "*Generador de contenidos*". Para ello, se parte de 850 previsiones generadas de forma automática, junto a la base de datos de componentes textuales que se utiliza para la generación automática de los textos, asegurando que cada componente clave aparezca al menos una vez en el corpus grabado.

Debido a que el mensaje de previsión de textos se divide en tres bloques independientes —bienvenida, previsión y despedida, a los que se añade un conjunto de palabras genérico (contiene los difonemas y trifonemas)—, el proceso de diseño del corpus para la aplicación del *Meteorólogo virtual* se desglosa en tres fases. La primera se encarga de escoger las frases que constituirán el subcorpus de bienvenida. Para ello, se toma como requisito que, además de incluir todas las componentes del corpus textual correspondientes a esta parte del mensaje, aparecieran todas las ciudades consideradas tanto dentro de una frase enunciativa como interrogativa —ya que éstas corresponden al elemento final de la bienvenida, por lo que se ven directamente afectadas por el tipo de entonación utilizado. A continuación, se recopilaron todas las frases correspondientes a la despedida por su reducido tamaño, grabándose varias veces para que la síntesis dispusiera del mismo mensaje pero con distintas entonaciones (evitando la típica monotonía de la locución de los sistemas de CTH). Finalmente, debido a la mayor variabilidad del subcorpus de previsiones (280 componentes clave a combinar), se aplica un algoritmo de *greedy* modificado (ver anexo C.1). Este algoritmo se puede ajustar para premiar la presencia de previsiones largas o cortas en la selección de los textos del corpus. En este caso, los resultados de cobertura fonética (difonemas y trifonemas) fueron muy parecidos tanto para las predicciones cortas como largas, por lo que se optó por escoger el conjunto de previsiones cortas (1381 frases, en total) para formar parte del subcorpus de previsiones, reduciendo así el tiempo de grabación necesario. Además, este análisis permitió llegar a otra conclusión: la baja cobertura fonética (difonemas y trifonemas) de las componentes del corpus. Así pues, para garantizar la correcta concatenación de componentes, y dar la posibilidad de sintetizar cualquier texto, se añadió a la lista de frases a grabar una lista de palabras que contenían la lista completa de difonemas y trifonemas (subcorpus *genérico*). La distribución de frases de cada uno de los subcorpus se presenta en la tabla D.3.

Tabla D.3: Distribución de frases y tamaño de cada subcorpus del corpus meteorológico.

Subcorpus	**Frases**	**Duración**
Bienvenida	960	36' 18"
Previsión	1381	74' 54"
Despedida	129	6' 42"
Genérico	1203	22' 06"
Total	3673	2h 30'

Construcción del corpus

Una vez diseñado el contenido textual del corpus, un locutor profesional de la CCRTV procedió a grabar el mismo en distintas sesiones —con una duración total de $2.5h$ (ver tabla D.3). Este locutor, escogido después de un proceso de elección previa de locutores[4], ajustó su estilo de locución al deseado, con el objetivo de conseguir una mayor naturalidad (estilo de locución *meteorológico*) y variabilidad (variaciones prosódicas, p.ej. énfasis en puntos distintos) de la síntesis.

A continuación, el corpus es etiquetado utilizando la Interfaz de Tratamiento del Habla (ver sección C.1, que incorpora el algoritmo de filtrado robusto de marcas (PMFA) descrito en el capítulo 4 de este trabajo. Finalmente, estas etiquetas (marcas de segmentación y marcas de *pitch*) son revisadas manualmente con la ayuda de la herramienta *CorpusTester* (ver sección C.1) para disponer de unos datos fiables tanto para la extracción de la prosodia de las unidades (se utiliza una estrategia *copy-prosody*) como para la síntesis del habla (se utiliza una estrategia de síntesis que necesita de las marcas de *pitch*).

Motor de síntesis

El sistema de CTH que constituye el motor de síntesis de la aplicación ha sido diseñado teniendo en cuenta que, por un lado, los mensajes que debe sintetizar se construyen a partir de un grupo limitado de componentes textuales, y por otro, la naturalidad de la síntesis es un elemento *clave* de las especificaciones. Por estos motivos, se optó por diseñar un CTH basado en selección de unidades utilizando la estrategia de concatenación de locuciones o *phrase-splicing* descrita en la sección 3.2.1. A grandes rasgos, esta técnica consigue sintetizar los mensajes de entrada a partir de la concatenación de segmentos de voz de tamaño considerable (en este caso, las componentes utilizadas por el sistema de generación automática de previsiones y grabadas en el corpus de voz). Este enfoque afecta a los distintos módulos que constituyen el motor de síntesis. El módulo de selección de unidades se ajustará para encontrar el mayor segmento posible de entre las componentes *clave* del

[4]En la fase de preselección se analizaron las características prosódicas de la señal de voz del grupo de locutores candidatos, evaluando la calidad de la síntesis obtenida sobre un experimento de resíntesis de las locuciones modificando el 20 %de su tono. A continuación, después de un experimento subjetivo, se ordenó a los candidatos según la naturalidad de la señal de voz obtenida después de ser modificada prosódicamente.

corpus. Se elimina el módulo de predicción prosódica típico de los CTH, y se sustituye por la prosodia del segmento seleccionado del corpus para preservar la variabilidad prosódica utilizada durante la fase de grabación del corpus (se evitan los errores de predicción y se consigue mayor realismo del habla del personaje). Este enfoque sólo es viable si se dispone de etiquetas fiables del corpus. Para ello, una vez obtenidas las marcas de segmentación (límites temporales de las unidades) y las marcas de *pitch* (utilizando el algoritmo descrito en el capítulo 4) se revisaron para evitar la presencia de errores de marcado utilizando la herramienta *CorpusTester* (ver sección C.1). Finalmente, el módulo de modificación de la señal simplemente resintetizará las unidades, ajustando su prosodia en tiempo de síntesis (deberá evitar discontinuidades prosódicas).

A continuación se describen estos procesos con más detalle.

Módulo de selección de unidades

El diseño e implementación del módulo de selección de unidades del CTH ha sido ajustado para seleccionar el conjunto de unidades que presente el mínimo número de puntos de concatenación, con el objetivo de recuperar las componentes clave completas. La elección de las unidades del corpus se realiza mediante una búsqueda dinámica exhaustiva (no se aplica preselección ni agrupación de unidades) teniendo en cuenta, dentro de la función de coste: las unidades básicas que constituyen el mensaje (difonemas o trifonemas con información de acento) y el tipo entonativo (enunciativo, interrogativo o exclamativo), como costes de unidad, y la secuencialidad de las unidades, como coste de concatenación. Así pues, la función de coste se ve simplificada respecto a la diseñada en el apartado 2.1.3 del presente trabajo de investigación, ya que, por el momento, no incorpora ningún elemento prosódico para discriminar entre las unidades candidatas —como se ha indicado anteriormente, la prosodia del mensaje se recupera de las unidades del corpus, en lugar de predecirla mediante el correspondiente módulo prosódico. Concretamente, el coste de unidad sólo toma en consideración la similitud entonativa de las unidades, según la matriz indicada en la tabla D.4. De este modo, se penalizará la degradación que las unidades sufrirán al tener que adaptar su prosodia desde un tipo entonativo a otro. Además, se utiliza un coste de concatenación binario ($C^c = 0$, para unidades consecutivas en el corpus y $C^c = 1$, de lo contrario) para evaluar la continuidad de las unidades seleccionadas. De este modo, se pretende premiar la recuperación de segmentos de componentes completos del corpus. Comentar que el proceso de selección de unidades se lleva a cabo en uno de los tres subcorpus que forman el corpus de voz (bienvenida, previsión y despedida), además de incorporar en todos los casos el subcorpus genérico para dar robustez a la síntesis. Es decir, se sigue una estrategia de síntesis en corpus independientes o *blending* para la CTH (ver sección 3.2.2).

Después de distintos experimentos preliminares, se pudo observar como, a menudo, varios grupos de secuencias de unidades candidatas presentaban el mismo valor acumulado de la función de coste. Este comportamiento, en parte atípico, está provocado, fundamentalmente, por dos razones: *i)* la presencia de distintas realizaciones de la misma componente clave en el corpus (variantes prosódicas o estilos de locución) con el objetivo de enriquecer la variabilidad del corpus, y *ii)* la simplicidad de la función de coste utilizada (no incorpora

Tabla D.4: Matriz de coste de unidad. Los acrónimos ENU, INT y EXC corresponden a las frases enunciativas, interrogativas y exclamativas, respectivamente.

C^t	**Objetivo**		
Selección	**ENU**	**INT**	**EXC**
ENU	0	0.05	0.05
INT	0.1	0	0.1
EXC	0.05	0.1	0

información prosódica). Como primer paso para solucionar esta cuestión, evitando elegir siempre los mismos segmentos de voz —cuestión diametralmente opuesta al objetivo de disponer una elevada variabilidad y naturalidad de la síntesis—, se incorpora un factor aleatorio en la selección de las secuencias de unidades idénticas en cuanto al coste acumulado. De este modo, se consigue que consultas sucesivas idénticas (para la misma ciudad y el mismo período de tiempo) puedan producir síntesis distintas, es decir, con distintos estilos de locución.

Ajuste prosódico

Una vez seleccionados los segmentos de voz del corpus que formarán el mensaje oral, se procede a recuperar su prosodia (F_0, duración y energía), que será utilizada como prosodia objetivo del módulo de síntesis de voz (estrategia *copy-prosody*). A menos que se indique en el fichero SSML-M, se mantiene la prosodia de las unidades para preservar el estilo de locución utilizado durante su grabación[5]. No obstante, debido a que la función de coste no incorpora ningún factor prosódico, se pueden producir discontinuidades tonales importantes en los puntos de concatenación de las unidades —sobretodo, en este caso, donde se trabaja con un corpus rico prosódicamente. Por este motivo, así como para poder incorporar modificaciones entonativas (p.ej. paso de un final de frase enunciativo a interrogativo), se incorpora a la arquitectura del CTH un Módulo de Ajuste del *Pitch* (MAP), encargado de ajustar de forma apropiada la curva de *pitch*, en estos casos.

Interpolación del tono en el punto de concatenación: La curva de *pitch* recuperada de las unidades seleccionadas constituye el tono objetivo que utilizará el módulo de síntesis de la señal de voz para construir el mensaje oral final. Aunque los valores prosódicos recuperados son fiables, como se ha comentado, pueden aparecer saltos en la curva de *pitch* en los puntos de concatenación de las unidades no consecutivas en el corpus. Para evitar una concatenación directa entre unidades con valores de F_0 muy distintos (procedentes de segmentos largos de voz), es necesario redefinir los valores de F_0 de los fonemas próximos al punto de concatenación. Para ello, se incorpora un proceso iterativo de ajuste progresivo de la curva de *pitch*, que aunque simple, es efectivo para evitar la existencia de disconti-

[5]Esta prestación, que permitía modificar la prosodia de la señal sintética a través de *tags* SSML, aunque soportada por el sistema de CTH, no se utiliza, por el momento, en la aplicación final

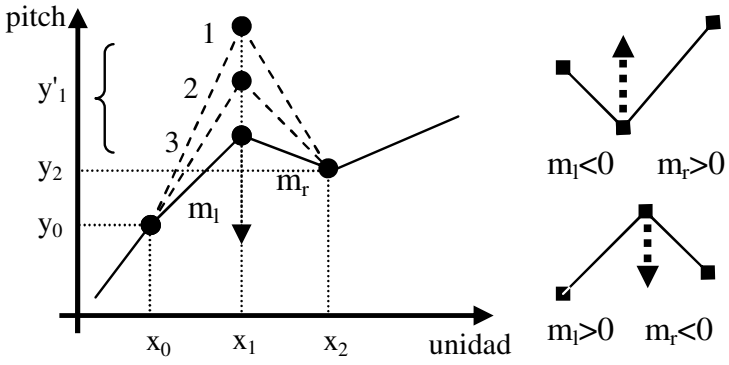


Figura D.4: Suavizado iterativo de un pico en la curva de *pitch* alrededor de un punto de concatenación de dos segmentos distintos. El ajuste depende de la concavidad del punto.

nuidades tonales en la síntesis. El ajuste de la F_0 de las unidades se lleva a cabo en una ventana de $\pm n$ unidades alrededor del punto de unión —en este trabajo $n = 3$, ajustado experimentalmente. El nuevo valor de F_0 se ajusta de forma iterativa (ver figura D.4) hasta conseguir que la diferencia entre la pendiente izquierda ($i - n$ valores) y derecha ($i + n$ valores) respecto al punto de concatenación (m_l y m_r, respectivamente, en la figura D.4) sea menor que un determinado umbral (empíricamente definido, en este caso $\Delta m_{max} = 0.2$):

$$|m_l - m_r| < \Delta m_{max} \tag{D.1}$$

Como se puede observar de la figura D.4, la variación del valor de la ordenada de la curva de *pitch* para el punto en cuestión dependerá de su concavidad. Si se trata de un punto cóncavo, el valor de partida será incrementando (ver ecuación (D.2)), pero si el punto es convexo, su valor se verá decrementando (ver ecuación (D.3)) hasta conseguir que la discontinuidad queda suavizada (diferencia entre pendientes menor que el umbral, ecuación (D.1)).

$$y_1^{'} < \frac{\Delta m_{max} + y_0 + y_2}{2} \tag{D.2}$$

$$y_1^{'} > \frac{-\Delta m_{max} + y_0 + y_2}{2} \tag{D.3}$$

Modificación del tono por variación de la entonación: Otra de las situaciones donde resulta necesario modificar la información de F_0 recuperada de las unidades se produce cuando la entonación del segmento de voz requerida es distinta de la que ha sido grabado, por lo que la curva de *pitch* debe se ajustada al nuevo patrón entonativo (por ejemplo, se solicita una final de frase interrogativo para una ciudad que sólo ha sido grabada dentro

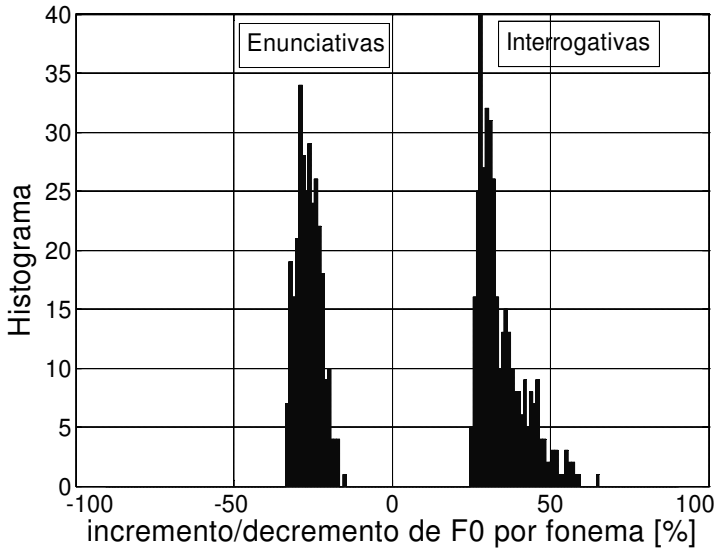


Figura D.5: Histograma de la variación relativa de F_0 para 400 frases enunciativas y 400 interrogativas.

una frase enunciativa). Concretamente, en el marco de la aplicación del *Meteorólogo Virtual* este proceso se realiza cuando es necesario aplicar una variación entonativa: *i)* a final de frase (p.ej. conversión de final de frase enunciativo a interrogativo, o viceversa), y *ii)* por la incorporación o eliminación de una pausa —una coma, generalmente. Ambas situaciones son debidas a que la componente clave recuperada del corpus no se ajusta perfectamente al patrón entonativo deseado, ya que disponer de todas las variantes posibles (entonación, variabilidad prosódica, posición en mensaje, pausado, etc.) de todas las componentes provocaría aumentar de forma considerable el tamaño del corpus. Por ello, se opta por abordar este problema mediante la modificación prosódica correspondiente, que aunque degrade, en parte, la señal sintética, permite dotar al CTH de cierta flexibilidad sin funcionar como un mero reproductor de mensajes pregrabados.

Del análisis del corpus, se observa, al comparar frases de contenido similar pero entonación distinta, que las variaciones tonales más significativas se producen a partir de la última vocal acentuada antes del punto donde se produce la variación entonativa ($p(x)$ en la ecuación (D.4)). El valor de F_0 de este punto (vocal acentuada) se utiliza como referencia para calcular el nuevo valor de F_0 asignado a las unidades consecutivas ($p(i)$, donde $i = x + 1, x + 2, \ldots$) para conseguir la entonación deseada (ver ecuación (D.4). Debido a que el incremento del tono en un final interrogativo es mucho mayor que el decremento del tono en un final enunciativo, se opta por modificar exponencialmente la curva de *pitch*, en el primer caso, mientras se aplica un decremento lineal en el segundo —ambos según un factor de ponderación $\beta \in [0, 1)$— respecto a la distancia de cada punto hasta la vocal acentuada $(i - x)$.

$$p(i) = \begin{cases} p(x) \cdot (1+\beta)^{(i-x)} & \text{if } p(i) > p(x) \\ p(x) \cdot \Big(1 - \beta(i-x)\Big) & \text{if } p(i) < p(x) \end{cases} \tag{D.4}$$

Para poder ajustar correctamente el valor de ponderación β, se analizó la prosodia de 800 frases del corpus, divididas en 400 interrogativas y 400 enunciativas. Como resultado, se ajustó el factor de variación relativa de F_0 a $\beta = 0.3$, para variaciones de entonación producidas a final de frase[6] (ver figura D.5). Por otro lado, después de un estudio similar, se ajustó el factor $\beta = 0.2$ para modificaciones de la entonación debidas a la incorporación o eliminación de una pausa. Se trata de un valor inferior al anterior, debido a que los cambios de entonación son menores en este caso. Finalmente, la figura D.6 presenta un ejemplo del efecto de este proceso, donde se muestra la curva de *pitch* original, obtenida a partir de los valores de F_0 recuperados de las unidades seleccionadas del corpus, acompañada de la curva resultante una vez adaptados los cambios de entonación.

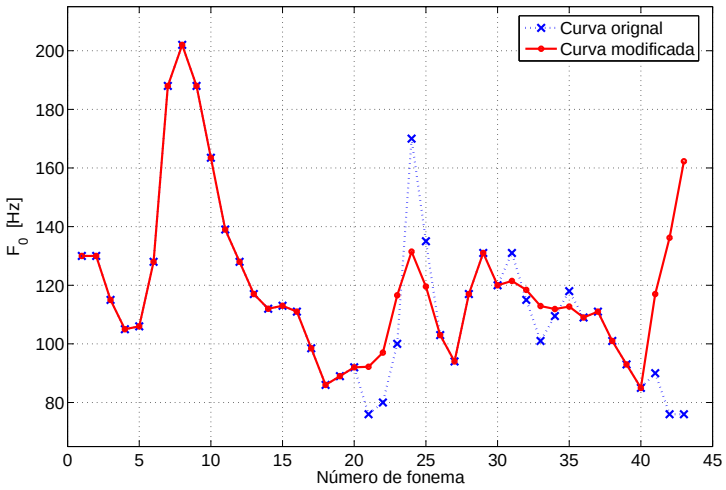


Figura D.6: Ajuste y suavizado de una curva de *pitch* en un punto de concatenación (unidad 23), por la incorporación de una pausa (unidad 31) y debido a una conversión enunciativa a interrogativa a final de frase (inicio en la unidad 40).

Módulo de síntesis

Una vez obtenida la lista definitiva de unidades con su prosodia asociada (modificada o no por el MAP), se procede a construir la señal sintética siguiendo un proceso idéntico al

[6]Nótese que en la presente aplicación, las frases interrogativas sólo corresponden a preguntas con respuesta Sí/No. Para otro tipo de preguntas, se tendría que repetir el estudio y reformular la transformación definida.

descrito en (Iriondo et al., 2003). En este caso, se trata de un algoritmo tipo TD-PSOLA (Moulines y Charpentier, 1990) que se encarga de sintetizar trama a trama todas las unidades. Concretamente, el ajuste de duración de la señal se basa en una interpolación de las tramas existentes (algoritmo $N2M$ (Iriondo et al., 2003)) y las pequeñas discontinuidades en energía presentes se minimizan mediante un suavizado trama a trama. Se optó por resintetizar toda la señal de voz en lugar de concatenar los fragmentos de voz natural (estrategia *phrase splicing* estricta) por distintas razones. Entre ellas: *i)* conseguir una calidad sintética sin fluctuaciones, es decir, evitar que el usuario perciba distintas calidades a lo largo de la señal sintética (p.ej. señal natural *vs.* señal modificada prosódicamente); *ii)* permitir la modificación prosódica de las componentes gravadas para ajustarse a patrones entonativos distintos a los propios (p.ej. cambio de entonación a final de frase); y *iii)* permitir la modificación de la prosodia intrínseca de las frases mediante etiquetado SSML (uno de los requisitos del proyecto, que finalmente no fue utilizado en la aplicación final).

Experimentos

A continuación se describen los experimentos realizados para evaluar el funcionamiento del CTH desarrollado. Primero se evalúa su funcionamiento en términos de las prestaciones objetivas conseguidas, y seguidamente, una vez ratificada informalmente la elevada calidad de la señal sintética generada, se evalúa la dependencia de este resultado respecto a los distintos elementos particulares del motor de síntesis, concretamente, el módulo de ajuste del *pitch* (MAP) y el impacto de la simplificación del coste de unidad (C^t) dentro de la función de coste. Para estos experimentos, la CCRTV proporcionó 900 ficheros de texto correspondientes a previsiones del tiempo de 175 ciudades distintas, que fueron sintetizadas para, además de validar la fiabilidad del sistema, realizar las pruebas que se describen a continuación.

Rendimiento del sistema

Debido a que el CTH descrito debía ser una aplicación que funcionara de forma automática 24h al día, se evaluó su rendimiento en términos de coste computacional sobre las 900 previsiones subministradas. Este experimento se realizó sobre un PC (PIV 3GHz - 1GB RAM) con sistema operativo Windows XP© y compilador Visual .NET 2003©.

A nivel de coste computacional, las previsiones sintetizadas presentaron una duración de $39.1 \pm 5.7s$ y fueron sintetizadas en $16.3 \pm 2.75s$ ($0.418\times$TR, donde TR indica tiempo real), validando, por un lado, la fiabilidad del CTH desarrollado (después de sintetizar las 900 previsiones sin problemas), y por otro, cumpliendo con las especificaciones del proyecto en términos de coste computacional.

Otro de los factores claves de la propuesta se basa en la recuperación de fragmentos lo más largos posible del corpus, es decir, se pretende recuperar las componentes claves grabadas. Para ello, en la función de coste, se ha definido un coste de concatenación binario para seleccionar unidades consecutivas en el corpus. A partir del resultado de las 900 previsiones

sintetizadas, se calcula el número medio de concatenaciones presentes en los ficheros, para evaluar así la capacidad del sistema de selección de unidades de recuperar componentes lo más largas posible. Concretamente, se obtiene un número de 0.55 ± 0.15 concatenaciones por frase, para 43.02 ± 5.23 unidades (difonemas y trifonemas) por previsión (con un promedio de 11.35 frases en cada una). Estos resultados (menos de una concatenación por frase) muestran el buen comportamiento del módulo de selección de unidades para cumplir con las especificaciones indicadas.

Finalmente, se analiza el impacto, a nivel de coste computacional, de trabajar con el corpus multidominio. Para ello, se comprueba, a partir de 505 previsiones completas, el aumento de la velocidad de la síntesis al trabajar con el corpus multidominio (el subcorpus genérico se añade a cada subcorpus para garantizar la cobertura total de las unidades de síntesis). Este estudio, demuestra que por el hecho de seleccionar las unidades sobre cada uno de los subcorpus en los que se divide el corpus de voz (bienvenida, previsión y despedida), se consigue una reducción media del 40 % del tiempo de ejecución si éste se compara con la selección de unidades sobre el corpus total (ver figura D.7) —resultado muy similar al obtenido también en los experimentos preliminares de CTH-MD (ver sección 3.4.1). Además, debido a la estrategia de síntesis utilizada —se trata de un CTH basado en concatenación de locuciones totalmente orientado a la aplicación en la que se enmarca—, se consigue la misma calidad —es decir, el mismo número medio de concatenaciones por previsión— si se trabaja sobre el corpus multidominio que si se trabaja con todo el corpus. Por lo tanto, la estrategia multidominio queda totalmente justificada dentro de la aplicación del *Meteorólogo virtual*.

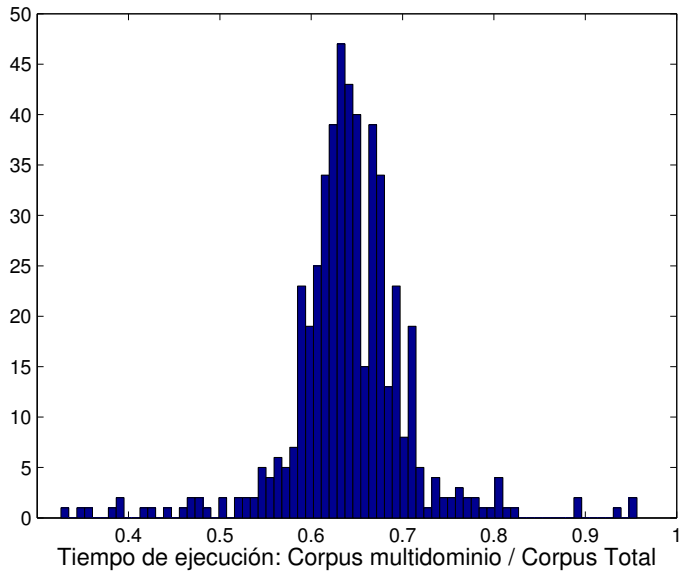


Figura D.7: Histograma de la relación entre el tiempo de ejecución sobre el corpus multidominio respecto al corpus total para 505 previsiones meteorológicas.

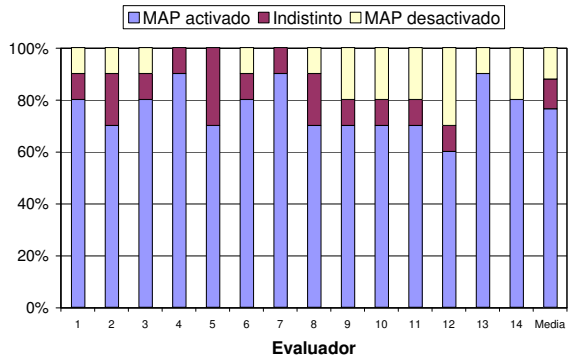
Pruebas subjetivas

A continuación, se presentan los resultados obtenidos de la evaluación de dos de los elementos particulares del CTH desarrollado relacionados directamente con la calidad sintética obtenida: el módulo de ajuste de *pitch* (MAP) y la matriz de costes de unidad (C^t). Para ello, se realizan dos tests de preferencia para evaluar el impacto de MAP y C^t sobre la calidad de la síntesis, formados por 10 y 12 pares de frases, respectivamente. Cada par estará constituido por una frase sintetizada con el módulo activo y el otro con el módulo desactivado, construidas a partir de textos de bienvenida y previsión. El evaluador puede escuchar las versiones de la misma frase tantas veces como crea necesario. Debe escoger la versión que le parezca más natural de entre las dos, pudiendo indicar que no tiene preferencia por ninguna de ellas (indistinto). Ambos tests fueron realizados por 14 miembros del Departamento de Comunicaciones y Teoría de la Señal de Enginyeria i Arquitectura La Salle.

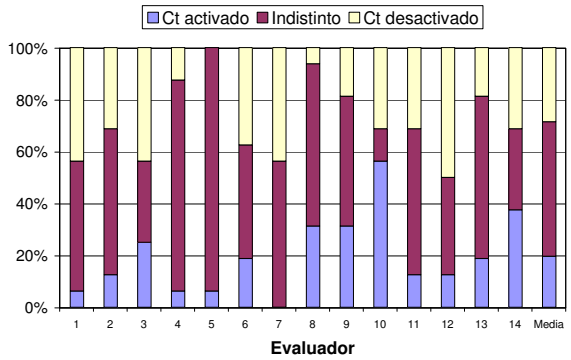
El resultado de las pruebas subjetivas se presenta en la figuras D.8. Como se puede observar de la figura, la activación de MAP en el proceso de síntesis es mucho más crítica que incorporar C^t en la función de coste de la selección. En el primer experimento (figura D.8(a)), los evaluadores presentan una clara preferencia (76 % de media más 11 % de indiferentes) a la necesidad de incorporar este módulo dentro del motor de síntesis —con una clara robustez estadística en términos de ANOVA ($F(2, 39) = 259.13, p < 0.000$). Mientras que en el segundo experimento (figura D.8(b)), los evaluadores presentan una cierta tendencia hacia la indiferencia en las comparativas entre los resultados obtenidos de incorporar o no la ponderación indicada en la matriz de la tabla D.4 para seleccionar las unidades con la entonación más adecuada ($F(2, 39) = 12.91, p < 0.000$). En este caso, se trata de una prueba más complicada que la anterior, ya que el efecto de esta selección queda enmascarado por el MAP, que modifica finalmente la prosodia para conseguir la más adecuada en términos de entonación. De algún modo, la prueba pretendía demostrar que cuanto más cercana la unidad se encuentre a la entonación deseada, menor modificación de la señal será necesaria por lo que la calidad sintética será mayor. No obstante, los resultados demuestran que éste es un factor con menos impacto en la calidad final respecto a la activación o no del MAP.

Discusión

El sistema de CTH de dominio limitado, pero a la vez capaz de sintetizar cualquier texto gracias a la inclusión de todos los difonemas y trifonemas del castellano, presenta un buen comportamiento tanto en términos de la calidad sintética alcanzada así como del coste computacional. No obstante, la calidad sintética decrece de forma considerable si se introducen textos de contenido alejado al dominio de la aplicación —al igual que sucede en (Hamza y Pitrelli, 2005), la calidad *in-script* > calidad *in-domain* > calidad *out-of-domain*). Esto se debe, fundamentalmente, a que el CTH no dispone de módulo de modelado prosódico, ya que se prima recuperar la prosodia real de los segmentos de voz para dotar de mayor naturalidad y realismo al personaje virtual. La inclusión de un módulo prosódico, la optimización del proceso de selección de unidades (p.ej. agrupando las unidades



(a) Módulo de ajuste de *pitch* (MAP) activado o desactivado.



(b) Matriz de C^t activada o desactivada.

Figura D.8: Preferencias de los evaluadores sobre las parejas sintéticas: 10 pares para MAP y 12 pares para C^t.

en lugar de realizar una búsqueda exhaustiva) o la incorporación de un proceso de selección de los segmentos con coste idéntico más inteligente (p.ej. seleccionar aleatoriamente uno de los segmentos y luego escoger la secuencia de componentes más próxima prosódicamente al seleccionado) son algunas de las líneas de trabajo que quedan abiertas para un futuro, en el desarrollo de aplicaciones similares a la descrita en esta sección anexa al trabajo de investigación.

Por otro lado, los resultados conseguidos con este trabajo han sido uno de los factores decisivos en la consecución del primer proyecto financiado por la Comunidad Europea en el que participa el grupo. Este proyecto, denominado SALERO [7], tiene como objetivo el diseño e implementación de herramientas que permitan la reusabilidad de objetos audiovisuales (imagen, vídeo, voz, efectos de sonido, etc.) en el ámbito del entretenimiento (cine, video juegos, etc.). En este proyecto participan 11 entidades de 5 países distintos, y se ha iniciado en Enero de 2006, con finalización prevista para Diciembre de 2009.

D.2. Locutor Virtual

Este proyecto nació de la colaboración entre los grupos de Visión por Computador y Tecnologías del Habla del Departamento de Comunicaciones y Teoría de la Señal de Enginyeria i Arquitectura La Salle. Las líneas de investigación a nivel del procesamiento de la señal de voz y del tratamiento de la imagen convergieron en el concepto de *síntesis multimodal*, que recientemente ha dado lugar al Grupo de Investigación en Procesamiento Multimodal [8].

[7] *Semantic AudiovisuaL Entertainment Reusable Objects*, proyecto IST-FP6-027122

[8] N de expediente: 2005SGR00806

Por un lado, durante el último trimestre de 1999 y todo el 2000, el grupo de Tecnologías del Habla llevó a cabo el proyecto "*Aplicación del conversor de texto en habla EMOVS sobre SAPI de Microsoft*" para Televisión de Cataluña, S.A. Este proyecto consistió en la migración del conversor de texto en habla (CTH) en catalán basado en difonemas a la arquitectura SAPI (*Speech Application Interface*[9]) de Microsoft ©. SAPI nació con el objetivo de ser un marco de normalización o estandarización de la comunicación entre las aplicaciones que incorporaran procesado de la señal de voz (síntesis o reconocimiento del habla). Por otro lado, el Grupo de Visión por Computador llevaba tiempo trabajando con herramientas de análisis y seguimiento de objetos con unos resultados muy interesantes.

Como resultado de la colaboración entre ambos grupos se desarrolló la primera versión de un sistema de síntesis multimodal basado en una cabeza parlante de apariencia fotorrealista, denominado *Locutor Virtual*. Este sistema, integró las tecnologías de procesamiento de la imagen para el modelado y síntesis de la apariencia visual del personaje virtual y las tecnologías de síntesis de voz desarrolladas hasta el momento (en este caso, integradas dentro de SAPI).

Una vez constatada la viabilidad de la propuesta, se solicitó financiación dentro del *Plan de Investigación Científica, Desarrollo e Innovación Tecnológica de la Investigación Técnica (PROFIT)*[10]. Mediante este proyecto se pudo llevar a cabo el primer prototipo de interfaz persona-máquina multimodal, basada en un locutor de apariencia fotorrealista (Melenchón, Alías y Iriondo, 2002; Melenchón et al., 2003). El locutor puede reproducir audiovisualmente *cualquier* texto de forma automática, sincronizando la voz con la gesticulación facial, para conseguir una sensación de gran naturalidad de la interfaz (Melenchón et al., 2003). El sistema es capaz de generar distintas expresiones faciales —alegre, neutra y triste— acompañadas de la señal de voz correspondiente, mediante la selección manual de la emoción utilizando una interfaz de usuario (corresponde a un ejemplo de selección manual del estilo de locución en el contexto de la CTH multidominio descrita en el capítulo 3 del este trabajo).

A continuación se describen los rasgos fundamentales del módulo de síntesis de imagen integrado en la aplicación del *Locutor Virtual*, así como su interacción con el módulo de síntesis de voz.

Módulo de procesamiento digital de la imagen

La componente visual del *Locutor Virtual* parte de un modelo facial en dos dimensiones (2D) parametrizado basado en imágenes reales y personalizable (Melenchón, Iriondo y Alías, 2002). Es un modelo que permite la incorporación de expresiones faciales y de ciertos movimientos de la cabeza, características que consiguen mejorar la naturalidad de la interfaz. El proceso de entrenamiento del modelo se fundamenta en técnicas de seguimiento robusto que posibilitan la personalización específica del locutor para cada *usuario* de forma bastante sencilla.

[9] http://www.microsoft.com/speech/
[10] Proyecto FIT-150500-2002-410

Módulo de síntesis del habla en SAPI 4.0 y SAPI 5.1

La aplicación del *Locutor Virtual* inicialmente se diseñó sólo para funcionar con motores SAPI 4.0. Posteriormente, se adaptó para permitir la coexistencia de motores basados en SAPI 4.0 y SAPI 5.1. De este modo, se aprovechan las ventajas incorporadas por la nueva versión de la interfaz (p.e. trabajar con *tags* en formato XML) y se mantiene la compatibilidad hacia atrás. Asimismo, el motor de síntesis (CTH) en catalán implementado inicialmente en SAPI 4.0 fue migrado a SAPI 5.1. Tanto la aplicación como el motor de síntesis fueron reprogramados con el fin de incorporar los nuevos objetos e interfaces de SAPI 5.1.

Como conclusión de este proceso, se dispone de una aplicación muy flexible y usable, ya que no sólo permite incorporar cualquier motor de síntesis de voz compatible con SAPI (el *Locutor Virtual* puede utilizar cualquier idioma, mientras el motor de síntesis escogido notifique correctamente los sonidos generados a la aplicación), sino que lo puede hacer con motores de diversas versiones de SAPI.

Arquitectura de la aplicación

El *Locutor Virtual* se ha diseñado para ser compatible con la interfaz SAPI. De este modo, se facilita la incorporación de nuevas *voces* y nuevos *idiomas* a la aplicación multimodal, ya que, como refleja la la figura D.9, la interfaz SAPI actúa de enlace entre la aplicación y el motor de síntesis (el sistema de conversión de texto en habla, CTH). El *Locutor Virtual* podrá hacer uso de cualquier CTH compatible con SAPI (en inglés, *SAPI-compliant*). El CTH será el encargado de notificar a la aplicación qué fonema está sonando en cada instante de tiempo, para que el bloque de síntesis visual de la aplicación genere la apariencia facial (*visema*) más adecuada. Así, se consigue la sincronización audiovisual del mensaje, dando realismo al mensaje audiovisual generado.

En este apartado se ha descrito el proceso de normalización del CTH del área de Tecnologías del Habla mediante la interfaz SAPI de Microsoft ®, (en una primera fase para SAPI 4.0 y posteriormente para SAPI 5.1). Así pues, se consigue que el motor de síntesis sea utilizable desde cualquier aplicación compatible con SAPI, interpretando los comandos (*tags*) que la aplicación inserta en el texto a sintetizar. Además, se ha intervenido en el proceso de creación de la aplicación *Locutor Virtual*, incorporando el motor de síntesis de voz en catalán a la aplicación e interviniendo en el proceso de diseño de la aplicación orientándola a la compatibilidad con SAPI. Asimismo, en una segunda versión de la aplicación, ésta ha sido adaptada para la gestión simultánea de motores SAPI 4.0 y SAPI 5.1. De este modo, se consigue que el CTH no sea una herramienta cerrada y sólo utilizable por una sola aplicación, sino que pueda ser integrada en cualquier aplicación compatible con SAPI.

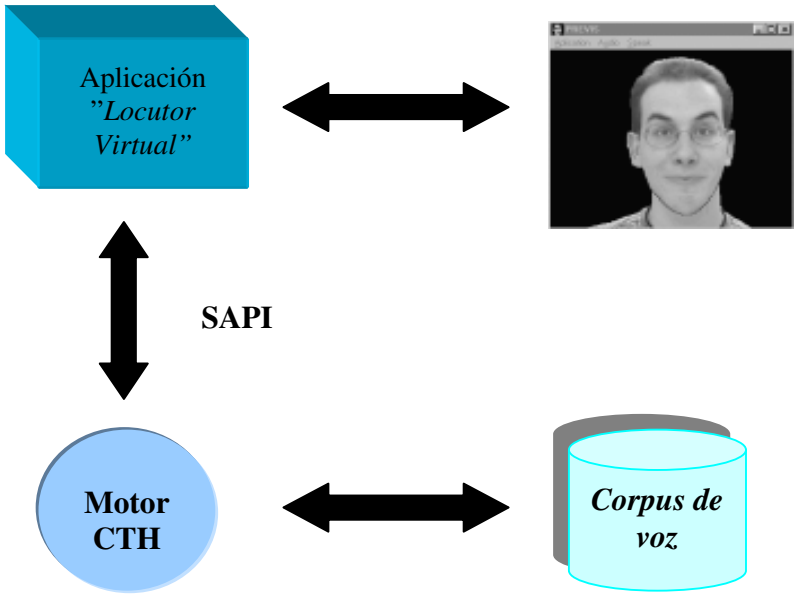


Figura D.9: Diagrama de bloques del CTH interactuando con el *Locutor Virtual*.

Bibliografía

Aas, K. y L. Eikvil. 1999. Text categorisation: A survey. Informe Técnico 941, Norwegian Computing Center.

Achan, K., S. Roweis, A. Hertzmann, y B. Frey. 2005. A segment-based probabilistic generative model of speech. En *Proceedings of ICASSP*, volumen 5, páginas 221–224, Philadephia, USA.

Aiikita, Y. y T. Kawahara. 2004. Language Model Adaptation based on PLSA of Topics and Speakers. En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 1045–1048, Jeju Island, Corea del Sur.

Alías, F. y I. Iriondo. 2001a. Asignación automática de marcas de pitch basada en programación dinámica. *Procesamiento del Lenguaje Natural*, 27:225–231, Septiembre.

Alías, F. y I. Iriondo. 2001b. Segmentador de fonemas en catalán basado en DHMM. En *Actas del XVI Simposium Nacional de la Unión Científica Internacional de Radio (URSI)*, páginas 149–150, Madrid.

Alías, F. y I. Iriondo. 2002. La evolución de la Síntesis del Habla en Ingeniería La Salle. En Red Temática en Tecnologías del Habla (RTTH), editor, *II Jornadas en Tecnología del Habla*, Granada.

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

Alías, F., I. Iriondo, L. Formiga, X. Gonzalvo, C. Monzo, y X. Sevillano. 2005. High quality Spanish restricted-domain TTS oriented to a weather forecast application. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 2573–2576, Lisboa, Portugal.

Alías, F. y X. Llorà. 2003. Evolutionary weight tuning based on diphone pairs for unit selection speech synthesis. En *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech)*, páginas 1333–1336, Geneve, Suiza.

Alías, F., X. Llorà, I. Iriondo, y L. Formiga. 2003. Ajuste subjetivo de pesos para selección de unidades a través de algoritmos genéticos interactivos. *Procesamiento del Lenguaje Natural*, 31:75–82, Septiembre.

Alías, F., X. Llorà, I. Iriondo, X. Sevillano, L. Formiga, y J. C. Socoró. 2004a. Perception-Guided and Phonetic Clustering Weight Tuning Based on Diphone Pairs for Unit Selection TTS. En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 1221–1224, Jeju Island, Corea del Sur.

Alías, F., X. Sevillano, P. Barnola, L. Formiga, I. Iriondo, y Socoró. J. C. 2004b. Conversión de Texto en Habla Multidomino. En Red Temática en Tecnologías del Habla (RTTH), editor, *III Jornadas en Tecnología del Habla*, páginas 101–106, Valencia.

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

Alías, Francesc. 1999. *Segmentador Automático*. Proyecto final de carrera. Ingeniería superior en Electrónica (Especialidad en Imagen y Sonido). Enginyeria i Arquitectura La Salle. Universitat Ramon Llull.

Alías, F., X. Llorà, L. Formgia, K. Sastry, y D.E. Goldber. 2006. Efficient interactive weight tuning for TTS synthesis: reducing user fatigue by improving user consistency. En *Proceedings of ICASSP*, volumen I, páginas 865–868, Toulouse, Francia.

Alías, F., C. Monzo, y J.C. Socoró. 2006. A Pitch Marks Filtering Algorithm based on Restricted Dynamic Programming. En *Proc. of InterSpeech - International Conference on Spoken Language Processing (ICSLP)*, Pittsburg, USA, Septiembre.

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

Alías, F., J.C. Socoró, X. Sevillano, I. Iriondo, y X. Gonzalvo. 2006. Multi-domain Text-to-Speech Synthesis by Automatic Text Classification. En *Proc. of InterSpeech - International Conference on Spoken Language Processing (ICSLP)*, Pittsburg, USA, Septiembre.

Allan, J. 2001. Perspectives on Information Retrieval and Speech. *Lecture Notes in Computer Science (Workshop on Information Retrieval Techniques for Speech Applications)*, 2273:1 – 10.

Ananthapadmanabha, T.V. y B. Yegnanarayana. 1975. Epoch extraction of voiced speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(11):562–570, Diciembre.

Arnold, D. V. y H.-G. Beyer. 2003. Interactive evolutionary computation: fusion of the capabilities of the ec optimization and human evaluation. *Massachusetts Institute of Technology Evolutionary Computation*, 11(2):111–127.

Asami, K., T. Takezawa, y G. Kikui. 2002. Topic detection of an utterance for Speech Dialogue Processing. En *Proceedings of ICSLP*, páginas 1977–1980, Denver, USA.

Baeza-Yates, R. y B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press Books, New York.

Bagshaw, P. C., S. M. Hiller, y M. A. Jack. 1993. Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. En *Proceedings of EuroSpeech*, volumen 2, páginas 1003–1006, Berlín, Alemania.

Bagshaw, Paul Christopher. 1994. *Automatic Prosodic Analysis for Computer-Aided Pronunciation Teaching*. Tesis Doctoral, The University of Edinburgh, Setiembre.

Balestri, M., A. Paechiotti, S. Quazza, P. L. Salza, y S Sandri. 1999. Choose the best to modify the least: a new generation concatenative synthesis system. En *Proceedings of EuroSpeech*, volumen 5, páginas 2291–2294, Budapest, Hungría.

Baluja, S. y R. Caruana. 1995. Removing the Genetics from the Standard Genetic Algorithm. En A. Prieditis y S. Russel, editores, *Proceedings of The International Conference on Machine Learning*, páginas 38–46, San Mateo, CA. Morgan Kaufmann Publishers.

Barner, K. E. 1996. Nonlinear estimation of DEGG signals with applications to speech pitch detection. En *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, páginas 2243–2246, Philadelphia, USA.

Barros, M. J., R. Maia, K. Tokuda, F. G. Resende, y D. Freitas. 2005. HMM-based European Portuguese TTS System. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 2801–2804, Lisboa, Portugal.

Batůšek, R. 2001. An objective measure for assessment of the concatenative TTS segment inventories. En *Proceedings of EuroSpeech*, páginas 2099–2102, Aalborg, Dinamarca.

Batůšek, R. 2002. An analysis of Limited Domains for Speech Synthesis. En *Proceedings of TSD*, páginas 265–268, Brno, República Checa. Springer.

Bellegarda, J. 2004. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42(1):93108.

Beutnagel, M. y A. Conkie. 1999. Interaction of units in a unit selection database. En *Proceedings of EuroSpeech*, volumen 3, páginas 1063–1066, Budapest, Hungría.

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, Berlín, Alemania.

Beutnagel, M., A. Conkie, y A. Syrdal. 1998. Diphone synthesis using unit selection. En *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia.

Beutnagel, M., M. Mohri, y M. Riley. 1999. Rapid unit selection from a large speech corpus for concatenative speech synthesis. En *Proceedings of EuroSpeech*, volumen 2, páginas 607–610, Budapest, Hungría.

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

Black, A.W. 2003. Unit Selection and Emotional Speech. En *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech)*, páginas 1649–1652, Geneve, Suiza.

Black, A.W. y N. Campbell. 1995. Optimising slection of units from speech databases for concatenative synthesis. En *Proceedings of EuroSpeech*, volumen 1, páginas 581–584, Madrid.

Black, A.W. y A. Font Llitjós. 2002. Unit Selection without a phoneme set. En *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, páginas 207–210, Santa Monica, USA.

Black, A.W. y K. Lenzo. 2000. Limited Domain Synthesis. En *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volumen 2, páginas 411–414, Beijing, China.

Black, A.W. y K. Lenzo. 2001a. Flite: a small fast run-time synthesis engine. En *The 4th ISCA Workshop on Speech Synthesis*, páginas 157–162, Perthshire, Escocia.

Black, A.W. y K. Lenzo. 2001b. Optimal Data Selection for Unit Selection Synthesis. En *The 4th ISCA Workshop on Speech Synthesis*, Perthshire, Escocia.

Black, A.W. y K. Lenzo. 2004. Multilingual Text-to-Speech Synthesis. En *Proceedings of ICASSP*, volumen 3, páginas 761–764, Montreal, Canadá.

Black, A.W. y P. Taylor. 1994. CHATR: A generic speech synthesis system. En *Proceedings of COLING-94*, volumen II, páginas 983–986, Kyoto, Japón.

Black, A.W. y P. Taylor. 1997a. Automatically clustering similar units for unit selection in speech synthesis. En *Proceedings of EuroSpeech*, páginas 601–604, Rodas, Grecia.

Black, A.W. y P. Taylor. 1997b. The Festival Speech Synthesis System: System documentation. Informe Técnico HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Escocia, UK.

Black, A.W. y K. Tokuda. 2005. Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 77–80, Lisboa, Portugal.

Blouin, C., O. Rosec, P.C. Bagshaw, y C. d'Alessandro. 2002. Concatenation cost calculation and optimisation for unit selection in TTS. En *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, USA.

Bánhalmi, A., K. Kovács, A. Kocsor, y L. Tóth. 2005. Fundamental Frequency Estimation by Least-Squares Harmonic Model Fitting. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 305–308, Lisboa, Portugal.

Boëffard, O. y F. Emerard. 1997. Application-dependent prosodic models for Text-to-Speech synthesis and automatic design of learning database corpus using genetic algorithm. En *EuroSpeech*, volumen IV, páginas 2507–2510, Rodas, Greece.

Bonafonte, A., I. Esquerra, A. Febrer, J. A. Fonollosa, y F. Vallverdú. 1998. The UPC Text-to-Speech System for Spanish and Catalan. En *Proceedings of ICSLP*, volumen 5, páginas 1667–1670, Sydney, Australia.

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*, páginas 201–206, Jenolan Caves, Australia.

Breiman, L., J. H. Friedman, R.A. Olshen, y Stone C. J. 1984. *Classification and Regression Trees*. The Wadsworth & Brooks/Cole Advanced & Books Software.

Breuer, R. y J. Abresch. 2004. Phoxsy: Multi-Phone Segments for Unit Selection Speech Synthesis . En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 1217–1220, Jeju Island, Corea del Sur.

Bulut, M., S.S. Narayanan, y A.K. Syrdal. 2002. Expressive Speech Synthesis Using a Concatenative Synthesizer. En *Proceedings of ICSLP*, páginas 1265–1268, Denver, USA.

Caldwell, C. y V. S. Johnston. 1991. Tracking a criminal suspect through face-space with a genetic algorithm. En *Proceedings of the Fourth International Conference on Genetic Algorithms*, páginas 416–421. Morgan Kaufmann.

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

Campbell, N. 2005. Developments in Corpus-Based Speech Synthesis: Approaching Natural Conversational Speech. *IEICE Transactions on Information and Systems (Invited paper)*, E88D(3):376–383, Marzo.

Campbell, N. y P. Mokhtari. 2003. Voice Quality: the 4th Prosodic Dimension. En *The 15th International Congress of Phonetic Sciences (ICPhS)*, páginas 2417–2420, Barcelona.

Campillo, F., J.L. Alba, y E. Rodríguez Banga. 2005. A Neural Network Approach for the Design of the Target Cost Function in Unit-Selection Speech Synthesis. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 2533–2536, Lisboa, Portugal.

Campillo, F. y E. Rodríguez Banga. 2002. Combined prosody and candidate unit selections for corpus-based Text-to-Speech systems. En *Proceedings of ICSLP*, volumen 1, páginas 141–144, Denver, USA.

Campillo, F. y E. Rodríguez Banga. 2003. On the design of cost functions for unit-selection speech synthesis. En *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech)*, páginas 289–292, Geneve, Suiza.

Campillo, Francisco León. 2005. *Síntesis de voz basada en selección de unidades acústicas y prosódicas*. Tesis Doctoral, Escola Técnica Superior de Enxeñeiros de Telecomunicación. Universidad de Vigo.

Camps, J., G. Bailly, y J. Martí. 1992. Synthèse à partir du texte pour le catalan. En *Proc. 19èmes Journées d'Études sur la Parole*, páginas 329–333, Bruxelles, Francia.

Cavnar, W.B. y J.M. Trenkle. 1994. N-Gram-Based Text Categorization. En *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, páginas 161–175, Las Vegas, USA.

Cernak, M. y M. Rusko. 2005. An evaluation of a synthetic speech using the PESQ measure. En *Proceedings of Forum Acusticum 2005*, Budapest, Hungría.

Chen, J-H. y Y-A. Kao. 2001. Pitch Marking Based on an Adaptable Filter and a Peak-Valley Estimation Method. *Computational Linguistics and Chinese Language Processing*, 6(2):1–12, Febrero.

Chen, S.-H., S.-J. Chen, y C.-C. Kao. 2006. Perceptual distortion analysis and qualitative estimation of prosody-modified speech for TD-PSOLA. En *Proceedings of ICASSP*, volumen I, páginas 861–864, Toulouse, Francia.

Cheng, Y.M. y D. O'Shaughnessy. 1989. Automatic and reliable estimation of glottal closure instant and period. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(12):1805–1815, Diciembre.

Chu, M., C. Li, P. Hu, y E. Cahng. 2002. Domain adaptation for TTS systems. En *Proceedings of ICASSP*, volumen 1, páginas 453–456, Orlando, USA.

Chu, M. y H. Peng. 2001. An Objective Measure for Estimating MOS of Synthesized Speech. En *Proceedings of EuroSpeech*, páginas 2087–2090, Aalborg, Dinamarca.

Chu, M., H. Peng, H.-Y. Yang, y E. Chang. 2001. Selecting non-uniform units from a very large corpus for concatenative speech synthesizer. En *Proceedings of ICASSP*, volumen 2, páginas 785–788, Salt Lake City, USA.

Chuang, Z.-J. y C.-H. Wu. 2002. Emotion recognition from textual input using an emotional semantic network. En *Proceedings of ICSLP*, páginas 2033–2036, Denver, USA.

Chung, G., S. Seneff, y L. Hetherington. 1999. Towards Multi-Domain Speech Understanding Using a Two-Stage Recognizer. En *Proceedings of EuroSpeech*, páginas 2655–2658, Budapest, Hungría.

Chung, Grace. 2001. *Towards multi-domain speech understanding with exible and dynamic vocabulary.* Tesis Doctoral, Massachusetts Institute of Technology, Junio.

Clark, R. A. J., K. Richmond, y S. King. 2004. Festival 2 - Build Your Own General Purpose Unit Selection Speech Synthesiser. En *Proceedings of the 5th ISCA Speech Synthesis Workshop*, páginas 173–178, Pittsburgh, USA.

Clark, R.A.J., K. Richmond, y S. King. 2005. Multisyn voices from ARCTIC data for the Blizzard challenge. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 101–104, Lisboa, Portugal.

Coello-Coello, Carlos A. December, 1998. An updated survey of GA-Based Multiobjective Optimization Techniques. Technical report lania-rd-09-08, Laboratorio Nacional de Informática Avanzada (LANIA), Xalapa, Veracruz, México.

Cohen, W.W. 1995. Fast Effective Rule Induction. En *Proc. of the 12th International Conference on Machine Learning*, páginas 115–123, Tahoe City, CA. Morgan Kaufmann.

Colotte, V. y Y. Laprie. 2002. Higher precision pitch marking for TD-PSOLA. En *Proceedings of the XI European Signal Processing Conference (EUSIPCO)*, volumen 1, páginas 419–422, Toulouse, Francia.

Conkie, A. 1999. Robust unit selection system for speech synthesis. En *Joint Meeting of ASA, EAA, and DAGA2*, Berlín, Alemania.

Conkie, A., M. C. Beutnagel, A. K. Sydal, y P. E. Brown. 2000. Preselection of candidate units in a unit selection-based Text-To-Speech Synthesis System. En *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volumen 3, páginas 314–317, Beijing, China.

Conkie, A. y S. Isard. 1996. Optimal coupling of diphones. En J. P. H. Santen R. W. Sproat J. P. Olive, y J. Hirschberg, editores, *Progress in Speech Synthesis*. Springer-Verlag, Berlín, Alemania.

Coorman, G., J. Fackrell, P. Rutten, y B. Van Coile. 2000. Segment selection in the L&H RealSpeak laboratory TTS system. En *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volumen 2, páginas 395–398, Beijing, China.

Cosi, P., F. Tesser, R. Gretter, y C. Avesani. 2001. Festival Speaks Italian! En *Proceedings of EuroSpeech*, páginas 509–512, Aalborg, Dinamarca.

Cowie, J. y W. Lehnert. 1996. Information extraction. *Communications of the ACM*, 39(1):80–91.

Cristianini, Nello y John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines*. Cambridge Press.

Darwin, Charles. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. Edición traducida de Planeta-Agostini (1992).

Davis, Lawrence. 1991. *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York.

de Cheveigné, A. y H. Kawahara. 2001. Comparative evaluation of F0 estimation algorithms. En *Proceedings of EuroSpeech*, páginas 2451–2454, Aalborg, Dinamarca.

de Cheveigné, A. y H. Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America (JASA)*, 111(4):1917–1930.

Deb, K., S. Agrawal, A. Pratab, y T. Meyarivan. 2000. A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II. KanGAL report 200001, Indian Institute of Technology.

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.

Diéguez, J., C. García, y A. Cardenal. 2005. Effective topic-tree based language model adaptation. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 1289–1292, Lisboa, Portugal.

Dikshit, P., S.A. Zahorian, y Nagulapati, S. 2005. A two-phase pitch marking method for TD-PSOLA synthesis. En *Proceedings of ICASSP*, volumen 1, páginas 233–236, Philadephia, USA.

Ding, W. y N. Campbell. 1997. Optimising unit selection with voice source and formants in the CHATR speech synthesis system. En *Proceedings of EuroSpeech*, páginas 537–540, Rodas, Grecia, Septiembre.

Donovan, R. 2000. Segment preselection in decision-tree based speech synthesis systems. En *Proceedings of ICASSP*, volumen 2, páginas 937–940, Istanbul, Turquía.

Donovan, R. E. 2001. A new distance measure for costing spectral discontinuities in concatenative speech synthesizers. En *The 4th ISCA Workshop on Speech Synthesis*, Perthshire, Escocia.

Donovan, R. E. y E. M. Eide. 1998. The IBM Trainable Speech Synthesis System. En *Proceedings of ICSLP*, volumen 5, páginas 1703–1706, Sydney, Australia.

Donovan, R. E., A. Ittycheriah, M. Franz, B. Ramabhadran, E. Eide, M. Viswanathan, R. Bakis, W. Hamza, M. Picheny, P. Gleason, T. Rutherfoord, P. Cox, D. Green, E. Janke, S. Revelin, C. Waast, B. Zeller, C. Guenther, y J. Kunzmann. 2001. Current Status of the IBM Trainable Speech Synthesis System. En *The 4th ISCA Workshop on Speech Synthesis*, Perthshire, Escocia.

Donovan, R.E., M. Franz, J.S. Sorensen, y S. Roukos. 1999. Phrase Splicing and Variable Substitution using the IBM Trainable Speech Synthesis System. En *Proceedings of ICASSP*, volumen 1, páginas 373–376, Phoenix, USA.

Droppo, J. y A. Acero. 1998. Maximum a posteriori pitch tracking. En *Proceedings of ICSLP*, páginas 943–946, Sydney, Australia.

Duda, R.O, P.E. Hart, y D. G. Stork. 2001. *Pattern Classification*. John Wiley & Soncs, Inc.

Durant, E., G. Wakefield, D. Van Tasell, y M. Rickert. 2004. Efficient Perceptual Tuning of Hearing Aids With Genetic Algorithms. *Trans. IEEE on Speech & Audio Processing*, 12(2):144–155.

Dutoit, T. y H. Leich. 1996. MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 13:435–440, Junio.

Dutoit, T. y Y. Stylianou, 1997. *Text-to-Speech Synthesis*, páginas 323–338. Handbook of Computational Linguistics. Oxford University Press.

Dutoit, Thierry. 1997. *An introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht.

Eide, E., A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith, y M. Viswanathan. 2003. Recent Improvements to the IBM Trainable Speech Synthesis System. En *Proceedings of ICASSP*, volumen 1, páginas 708–711, Hong Kong.

Esquerra, I., A. Febrer, y C.Ñadeu. 1998. Frequency analysis of phonetic units for concatenative synthesis in Catalan . En *Proceedings of ICSLP*, volumen 5, páginas 1959–1962, Sydney, Australia.

Febrer, A. y A. Bonafonte. 2000. Síntesis del habla por concatenación basada en selección. En *I Jornadas en Tecnolgías del Habla*, Sevilla.

Febrer, Albert. 2001. *Síntesi de la parla per concatenació basada en la selecció*. Tesis Doctoral, Departament de Teoria del Senyal i Comunicacions. Universitat Politècnica de Catalunya, Enero.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.

Ferencz, A., J. Kim, Y-B. Lee, y J-W. Lee. 2004. Automatic pitch marking and reconstruction of glottal closure instants from noisy and deformed electro-glotto-graph signals. En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 2437–2440, Jeju Island, Corea del Sur.

Fischer, V., J. Botella, y S. Kunzmann. 2004. Domain Adaptation Methods in The IBM trainable Text-To-Speech System. En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 1165–1168, Jeju Island, Corea del Sur.

Formiga, Lluís. 2003. *Ajuste de pesos de selección para síntesis del habla a través de algoritmos genéticos interactivos*. Trabajo final de carrera. Ingeniería técnica en Informática. Enginyeria i Arquitectura La Salle. Universitat Ramon Llull.

Formiga, Lluís. 2005. *Reducción de la fatiga y la ambigüedad en el ajuste subjetivo de pesos para síntesis del habla*. Proyecto final de carrera. Ingeniería superior en Informática. Enginyeria i Arquitectura La Salle. Universitat Ramon Llull.

Frakes, W.B. y R. Baeza-Yates. 1992. *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, New Jersey.

François, H. y O. Boëffard. 2002. The Greedy Algorithm and its Application to the Construction of a Continuous Speech Database. En *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, volumen 5, páginas 1420–1426, Las Palmas de Gran Canaria.

Freund, Y. y R.E. Schapire. 1996. Experiments with a new boosting algorithm. En *13th International Conference on Machine Learning*, páginas 148–156.

Friedman, J. H. 1994. Flexible metric nearest neighbor classification. Informe Técnico 113, Stanford University Statistics Department, Palo Alto, CA.

Galibert, O., G. Illouz, y S. Rosset. 2005. Ritel: An Open-Domain, Human-Computer Dialog System. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 909–912, Lisboa, Portugal.

García, M., M.T. Martín, y L.A. Ureña. 2001. Categorización de textos multilingües basada en Redes Neuronales. *Procesamiento del Lenguaje Natural*, 27:265–271, Septiembre.

Gerhard, D. 2003. Pitch Extraction and Fundamental Frequency: History and Current Techniques. Informe Técnico TR-CS 2003-6, University of Regina, Regina, Saskatchewan, Canadá, Noviembre.

Goldberg, D. E., B. Korb, y K. Deb. 1989. Messy genetic algorithms: Motivation, analysis, and first results. *Complex Systems*, 3(5):493–530.

Goldberg, D.E. 1989. *Genetic Algorithms in Search Optimization and Machine Learning*. Addison-Wesley.

Goldberg, D.E. 2002. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Kluwer Academic Publishers.

Goncharoff, V. y P. Gries. 1998. An algorithm for accurately marking pitch pulses in speech signals. En *Proceedings of IASTED International Conference on Signal and Image Processing*, páginas 281–284, Las Vegas, USA.

Gorin, A.L., G. Riccardi, y J.H. Wright. 1997. How may I help you? *Speech Communication*, 23(1/2):113–127.

Guaus, R., F. Gudayol, y J. Martí. 1996. Conversión Texto-Voz mediante síntesis PSOLA. En *Jornadas Nacionales de Acústica*, páginas 355–358, Barcelona.

Guaus, R. y I. Iriondo. 2000a. Diphone-Based Unit Selection for Catalan TTS Synthesis. En *Proceedings of the International Conference on Text, Speech and Dialogue (TSD)*, Brno, República Checa. Springer.

Guaus, R. y I. Iriondo. 2000b. Unit Selection based on Diphones for Catalan Text-to-Speech Conversion. En *Workshop on developing language resources for minority languages*, Atenas, Grecia.

Guaus, R., J. Oliver, F. Gudayol, y J. Martí. 1997. Síntesis de voz utilizando difonemas: Uniones entre vocales. *Procesamiento del Lenguaje Natural*, 21:69–74, Julio.

Guaus, R., J. Oliver, H. Moure, I. Iriondo, y J Martí. 1998. Síntesis de voz por concatenación de unidades: Mejoras en la calidad segmental. En *TecniAcústica*, Lisboa, Portugal.

Haffner, P., G. Tur, y J.H. Wright. 2003. Optimizing SVMs for complex call classification. En *Proceedings of ICASSP*, volumen 1, páginas 632–635, Hong Kong.

Hagmüller, M. y G. Kubin. 2005. Poincaré Sections for Pitch Mark Determination. En *Proceedings of ISCA Tutorial and Research Workshop on Non-linear Speech Processing*, páginas 107–113, Barcelona.

Hamza, W., R. Bakis, E. M. Eide, M. A. Picheny, y J. F. Pitrelli. 2004. The IBM Expressive Speech Synthesis System. En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 2577–2580, Jeju Island, Corea del Sur.

Hamza, W. y R. Donovan. 2002. Data-driven segment preselection in the IBM trainable speech synthesis system. En *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP*, volumen 4, páginas 2609–2612, Denver, USA.

Hamza, W. y J.F. Pitrelli. 2005. Combining the flexibility of speech synthesis with the naturalness of pre-recorded audio: a comparison of two approaches to phrase-splicing TTS. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 2585–2588, Lisboa, Portugal.

Harbeck, S., A. Kießling, R. Kompe, H. Niemann, y E. Nöth. 1995. Robust pitch period detection using dynamic programming with an ANN cost function. En *Proceedings of EuroSpeech*, volumen 2, páginas 1337–1340, Madrid, Setiembre.

Harik, G.R., F. G. Lobo, y D.E. Goldberg. 1999. The Compact Genetic Algorithm. *IEEE Transactions on Evolutionary Computation*, 3(4):287–297, November.

Hazen, T.J., I.L. Hetherington, y A. Park. 2001. FST-Based Recognition Techniques for Multi-Lingual and Multi-Domain Spontaneous Speech. En *Proceedings of EuroSpeech*, volumen 2, páginas 1591–1594, Aalborg, Dinamarca.

Hersh, W., C. Buckley, T. Leone, y D. Hichman. 1994. OHSUMED: an interactive retrieval evaluation and new large text collection for research. En *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, página 192201, Dublín, Irlanda.

Hess, W. 1983. *Pitch Determination of Speech Signals*. Springer-Verlag, Berlin; New York.

Hofer, G., K. Richmond, y R.A.J. Clark. 2005. Informed blending of databases for emotional speech synthesis. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 501–504, Lisboa, Portugal.

Holland, J. H. 1975. *Adaptation in Natural and Artificial Systems*. Univesity of Michigan Press.

Holland, J. H. 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. MIT Press/ Bradford Books edition.

Hon, H., A. Acero, X. Huang, J. Liu, y M. Plumpe. 1998. Automatic Generation of synthesis units for trainable text-to-speech systems. En *Proceedings of ICASSP*, volumen 1, páginas 293–296, Seattle, USA.

Hori, T., D. Willett, y Y. Minami. 2003. Language model adaptation using WFST-based speaking-style translation. En *Proceedings of ICASSP*, volumen 1, páginas 228–231, Hong Kong.

Hosom, J-P. 2005. *F*0 Estimation for Adult and Childrens Speech. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 317–320, Lisboa, Portugal.

HTK. 2001. HTK v3.1.1. http://htk.eng.cam.ac.uk/index.shtml.

Huang, X., A. Acero, J Adcock, H. Hon, J. Goldsmith, J. Liu, y M. Plumpe. 1996. WHISTLER: a Trainable Text-to-Speech System. En *Proceedings of ICSLP*, páginas 2387–2390, Philadelphia, USA.

Huang, X., A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, y M. Plumpe. 1997. Recent Improvements on Microsoft's Trainable Text-To-Speech System - Whistler. En *Proceedings of ICASSP*, páginas 959–962, Munic, Alemania.

Huang, X., A. Acero, y H-W. Hon. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall.

Hunt, A. y A.W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. En *Proceedings of ICASSP*, volumen 1, páginas 373–376, Atlanta, USA.

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

Iida, A. y N. Campbell. 2001. A database design for a concatenative speech synthesis system for the disabled. En *The 4th ISCA Workshop on Speech Synthesis*, páginas 188–194, Perthshire, Escocia.

Iida, A., N. Campbell, F. Higuchi, y M. Yasumura. 2003. A corpus-based Speech Synthesis System with Emotion. *Speech Communication*, 40(1,2):161–187.

Iida, A., N. Campbell, S. Iga, F. Higuchi, y M. Yasumura. 2000. A Speech Synthesis System with Emotion for Assisting Communication. En *Proceedings of the ISCA Workshop on Speech and Emotion*, páginas 167–172, Newcastle, Irlanda del Norte.

Iriondo, I., F. Alías, y J. Melenchón. 2002. Un modelo híbrido orientado a la síntesis multimodal del habla. *Procesamiento del Lenguaje Natural*, 29:159–163, Septiembre.

Iriondo, I., F. Alías, J. Melenchón, y M. A. Llorca. 2004. Modeling and Synthesizing Emotional Speech for Catalan Text-to-Speech Synthesis. *Lecture Notes in Artificial Intelligence (Tutorial and Research Workshop on Affective Dialog Systems)*, (3068):197–208, Junio.

Iriondo, I., F. Alías, J. Sanchis, y J. Melenchón. 2003. A Hybrid Method Oriented to Concatenative Text-to-Speech Synthesis. En *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech)*, páginas 2953–2958, Geneve, Suiza.

Iriondo, I., R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. Blanco, D. Bernadas, J. Oliver, D. Tena, y L. Longhi. 2000. Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. En *Proceedings of the ISCA Workshop on Speech and Emotion*, páginas 161–166, Newcastle, Irlanda del Norte.

Iriondo, I., J. Martí, J. Oliver, R. Guaus, y H. Moure. 1999. Hacia una síntesis concatenativa de alta calidad para aplicaciones de conversión texto-habla. *Procesamiento del Lenguaje Natural*, 25:109–113, Septiembre.

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

Itou, K., A. Fujii, y T. Ishikawa. 2001. Language modeling for multi-domain speech-driven text retrieval. En *Proc of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, páginas 327 – 330.

Iyer, R., J. Ma, Gish H., y O. Kimball. 2002. Training Topic Classifiers for Conversational Speech with Limited Data. En *Proceedings of ICSLP*, volumen 3, páginas 1501–1504, Denver, USA.

Jiang, D., W. Zhang, L. Shen, y L. Cai. 2005. Prosody Analysis and Modeling for Emotional Speech Synthesis. En *Proceedings of ICASSP*, volumen 1, páginas 281 – 284, Philadephia, USA.

Jilka, M. y A. K. Syrdal. 2002. The AT&T German Text-to-Speech System: realistic linguistic description. En *Proceedings of ICSLP*, volumen 1, páginas 113–116, Denver, USA.

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. En *Proceedings of ECML-98, 10th European Conference on Machine Learning*, número 1398, páginas 137–142. Springer Verlag, Heidelberg, DE.

Joachims, T. 2000. SVMlight. http://ais.gmd.de/˜ thorsten/svm_light/.

Joachims, T. 2002. *Learning to Classify Text using Support Vector Machines*. Kluwer Academic Publishers.

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

Kaban, A. y M. Girolami. 2000. Unsupervised Topic Separation and Keyword Identification in Document Collections: A Projection Approach. Informe Técnico 10, Department of Computing and Information Systems, University of Paisley.

Kasi, K. y S.A. Zahorian. 2002. Yet another algorithm for pitch tracking. En *Proceedings of ICASSP*, volumen 1, páginas 361–364, Orlando, USA.

Kawai, H., T. Toda, J. Ni, M. Tsuzaki, y K. Tokuda. 2004. XIMERA: A New TTS from ATR Based on Corpus-Based Technologies. En *Proc. 5th ISCA Speech Synthesis Workshop*, páginas 179–184, Pittsburgh, USA.

Kim, S., Y. Lee, y K. Hirose. 2001. Pruning of Redundant Synthesis Instances Based on Weighted Vector Quantization. En *Proceedings of EuroSpeech*, volumen 3, páginas 2231–2234, Aalborg, Dinamarca.

Klabbers, E . y R. Veldhuis. 1998. On the Reduction of Concatenation Artefacts in Diphone Synthesis. En *Proceedings of ICSLP*, páginas 1983–1986, Sydney, Australia.

Klabbers, E. y R. Veldhuis. 2001. Reducing audible spectral discontinuities. *IEEE Transactions on Speech and Audio Processing*, 9(1):39–51.

Kobayashi, M., M. Sakamoto, T. Saito, Y. Hashimoto, M. Nishimura, y K. Suzuki. 1998. Wavelet analysis used in text-to-speech synthesis. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 45(8):1125 – 1129.

Kominek, J. y A.W. Black. 2003. The CMU ARCTIC databases for speech synthesis. Informe Técnico CMU-LTI-03-177, Language Technologies Institute, Carnegie Mellon University.

Kominek, J. y A.W. Black. 2004. The CMU ARCTIC speech databases. En *Proceedings of the 5th ISCA Speech Synthesis Workshop*, páginas 223–224, Pittsburg, USA.

Kounoudes, A., P.A. Naylor, y M. Brookes. 2002. The DYPSA algorithm for estimation of glottal closure instants in voiced speech. En *Proceedings of ICASSP*, volumen 1, páginas 349–352, Orlando, USA.

Krishnamurthy, A. y D. Childers. 1986. Two-Channel Speech Analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):730–743, Agosto.

Kumar, R. 2004. A Genetic Algorithm for Unit Selection based Speech Synthesis. En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 1233–1236, Jeju Island, Corea del Sur.

Lane, I.R., T. Kawahara, T. Matsui, y S.Ñakamura. 2004. Out-of-domain detection based on confidence measures from multiple topic classification. En *Proceedings of ICASSP*, volumen 1, páginas 757–760, Montreal, Canadá.

Lane, I.R., T. Kawahara, T. Matsui, y S.Ñakamura. 2005. Dialogue Speech Recognition by Combining Hierarchical Topic Classification and Language Model Switching. *IEICE Transactions on Information and Systems*, E88D(3):446–454, Marzo.

Laprie, Y. y V. Colotte. 1998. Automatic pitch marking for speech transformations via TD-PSOLA. En *Proceedings of the XI European Signal Processing Conference (EUSIPCO)*, Rodas, Grecia.

Larrañaga, P. y J. A. Lozano. 2002. *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*. Kluwer Academic Publishers.

Lauri, F., I. Illina, y F. Fohr, D. Korkmazsky. 2003. Using Genetic Algorithms for Rapid Speaker Adaptation. En *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech)*, páginas 1497–1500, Geneve, Suiza.

Lee, M., D. P. Lopresti, y J. P. Olive. 2001. A Text-to-Speech Platform for Variable Length Optimal Unit Searching Using Perceptual Cost Functions. En *The 4th ISCA Workshop on Speech Synthesis*, páginas 75–80, Perthshire, Escocia.

Lenzo, K. y A.W. Black. 2000. Diphone collection and synthesis. En *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volumen 3, páginas 306–309, Beijing, China.

Lenzo, K. y A.W. Black. 2002. Customized synthesis: blending and tiering. En *The 21st annual Applied Voice Input/Output Society conference (AVIOS)*, Santa Jose, USA.

Lewis, D. D. 1994. Colección de textos en Inglés Reuters-21578 . Disponible en: http://www.research.att.com/˜ lewis/reuters21578.html.

Li, X., J. Malkin, y J. Bilmes. 2004. Graphical model approach to pitch tracking. En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 1101–1104, Jeju Island, Corea del Sur.

Lin, C-Y. y J-S. R. Jang. 2004. A two-phase pitch marking method for TD-PSOLA synthesis. En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 1189–1192, Jeju Island, Corea del Sur.

Liu, H., H. Lieberman, y T. Selker. 2003. A model of textual affect sensing using real-world knowledge. En *Proceedings of ICSLP*, página 125132, Miami, USA.

Liu, J., T.F. Zheng, J. Deng, y W. Wu. 2005. Real-time Pitch Tracking Based on Combined SMDSF. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 301–304, Lisboa, Portugal.

Llisterri, J., N. Fernández, F. Gudayol, J.J. Poyatos, y J. Martí. 1993. Testing user's acceptance of Ciber232, a text to speech system used by blind persons. En B. Granström S. Hunnicutt, y K.-E. Spens, editores, *Proceedings of an ESCA Workshop on Speech and Language Technology for Disabled Persons*. Stockholm, Suecia, páginas 203–206.

Llorà, Francesc Xavier. 1996. *Optimización de funciones reales de variable real utilizando algoritmos genéticos bajo un entorno Windows*. Trabajo final de carrera. Ingeniería técnica en Informática. Enginyeria i Arquitectura La Salle. Universitat Ramon Llull.

Llorà, Francesc Xavier. 2001. *Aprenendizaje artificial evolutivo utilizando paralelismo de grano fino en el marco de la minería de datos*. Tesis Doctoral, Departamento de Informática de Enginyeria i Arquitectura La Salle. Universitat Ramon Llull, Noviembre.

Llorà, X., F. Alías, L. Formiga, K. Sastry, y D. E. Goldberg. 2005. Evaluation Consistency in iGAs: User Contradictions as Cycles in Partial-Ordering Graphs. IlliGAL Report No. 2005022, University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, Urbana, IL.

Llorà, X., K. Sastry, F. Alías, D. E. Goldberg, y M. Welge. 2006. Analyzing Active Interactive Genetic Algorithms using Visual Analytics. En *Proceedings of Genetic and Evolutionary Computation Conference 2006 (GECCO-2006)*, Seattle, USA. ACM Press. (También IlliGAL Report No. 2006004).

Llorà, X., K. Sastry, D. E. Goldberg, A. Gupta, y L. Lakshmi. 2005. Combating User Fatigue in iGAs: Partial Ordering, Support Vector Machines, and Synthetic Fitness. *Proceedings of Genetic and Evolutionary Computation Conference 2005 (GECCO-2005)*, páginas 1363–1371. (También IlliGAL Report No. 2005009).

Lo, W.K., T. Lee, y P.C. Ching. 1998. Development of Cantonese spoken language corpora for speech processing. En *Proceedings of 1998 International Symposium on Chinese Spoken Language Processing*, páginas 102–107, Singapore.

López-Cózar, R. 2003. Uso de Información Contextual en la Interfaz de Entrada de Un Sistema de Diálogo. *Procesamiento del Lenguaje Natural*, 31:251–258, Septiembre.

López-Cózar, R., A.J. Rubio, P. García, J.E. Díaz-Verdejo, y J.M López-Soler. 2000. NoVo: Sistema Automático Basado en Reconocimiento de Voz para el Acceso Remoto a Noticias. En Red Temática en Tecnologías del Habla (RTTH), editor, *I Jornadas en Tecnología del Habla*, Sevilla.

Macon, M. W., A. E. Cronk, y J. Wouters. 1998. Generalization and discrimination in tree-structured unit selection. En *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, páginas 195–200, Jenolan Caves, Australia.

Mann, Iain. 1999. *An Investigation of Nonlinear Speech Synthesis and Pitch Modification Techniques*. Tesis Doctoral, The University of Edinburgh, Setiembre.

Martí, J. 1985. *Estudi acústic del català i síntesi automàtica per ordinador*. Tesis Doctoral, Universitat de València.

Martí, J. 1987. Síntesis del habla: Evolución histórica y situación actual. En F. Casacuberta y E. Vidal, editores, *Reconocimiento automático del habla*. Marcombo, Boixareu, páginas 187–205.

Martí, J. 1990. Estado actual de la síntesis de voz. En *Estudios de Fonética Experimental*, número 4, páginas 147–168.

Martí, M.A., J.A. Alonso, T. Badia, J. Campàs, X. Gómez, J. Gonzalo, J. Llisterri, J. Rafel, H. Rodríguez, J. Soler, y M.F. Verdejo. 2003. *Tecnologías del lenguaje*. Editorial UOC.

Melenchón, J., F. Alías, y I. Iriondo. 2002. PREVIS: a Person-specific Realistic Virtual Speaker. En *Proceedings of the International Conference of Multimedia and Expo (ICME)*, Lausanne, Suiza.

Melenchón, J., F. de la Torre, I. Iriondo, F. Alías, E. Martínez, y Ll. Vicent. 2003. Text to visual synthesis with appearance models. En *IEEE International Conference on Image Processing (ICIP)*, páginas 237–240, Barcelona, Septiembre.

Melenchón, J., I. Iriondo, y F. Alías. 2002. Modelo 2d parametrizado basado en imágenes reales orientado a síntesis de cabezas parlantes. En *Actas del XVII Simposium Nacional de la Unión Científica Internacional de Radio (URSI)*, páginas 383–384, Alcalá de Henares.

Mendel, Gregor. 1965. *Experiments in Plant Hybridization*. Harvard University Press.

Meng, H.M., K.C. Keung, C.K. Siu, T.Y. Fung, y P.C. Ching. 2002. CU VOCAL: Corpus-Based Syllable Concatenation for Chinese Speech Synthesis Across Domains and Dialects. En *Proceedings of ICSLP*, páginas 2373–2376, Denver, USA.

Meron, Y. y K. Hirose. 1999. Efficient weight training for selection based synthesis. En *Proceedings of EuroSpeech*, volumen 5, páginas 2319–2322, Budapest, Hungría.

Michalewicz, Zbigniew. 1992. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag.

Miller, B. L. y D. E. Goldberg. 1995. Genetic algorithms, tournament selection, and the effects of noise. *Complex Systems*, 9(3):193–212. (Also IlliGAL Report No. 95006).

Miller, B. L. y D. E. Goldberg. 1997. Genetic algorithms, selection schemes, and the varying effects of noise. *Evolutionary Computation*, 4(2):113131.

Miralles, Pere. 2005. *Modelado de la prosodia mediante aprendizaje analógico aplicado a la síntesis del habla*. Proyecto final de carrera. Ingeniería técnica en Informática. Enginyeria i Arquitectura La Salle. Universitat Ramon Llull.

Mittendorf, E., B. Mateev, y P. Schäuble. 2000. Using the co-occurence of words for retrieval weighting. *Information Retrieval*, 3:243–251.

Miyanaga, K., T. Masuko, y T. Kobayashi. 2004. A style control technique for HMM-based speech synthesis. En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 1437–1440, Jeju Island, Corea del Sur.

Möbius, B. 2000. Corpus-based speech synthesis: methods and challenges. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)*, 6(4):87–116.

Möbius, B. 2001. Rare events and closed domains: two delicate concepts in Speech Synthesis. En *The 4th ISCA Workshop on Speech Synthesis*, páginas 41–46, Perthshire, Escocia.

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 *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volumen 1, páginas 621–624, Beijing, China.

Montero, J.M., J. Gutiérrez-Arriola, J. Colás, E. Enríquez, y J.M. Pardo. 1999. Analysis and Modelling of Emotional Speech in Spanish. En *The 14th International Congress of Phonetic Sciences (ICPhS)*, volumen 2, páginas 957–960, San Francisco, USA.

Montoya, N. 1999. *El uso de la voz en la publicidad audiovisual dirigida a los niños y su eficacia persuasiva*. Tesis Doctoral, Universitat Autònoma de Barcelona.

Moulines, E. y F. Charpentier. 1990. Pitch-Synchronous waveform processing techniques for Text-to-Speech synthesis using diphones. *Speech Communication*, (9):453–467.

Moulines, E., F. Emerard, D. Larreur, J. L. Le Saint Milon, L. Le Faucheur, F. Marty, F. Charpentier, y C. Sorin. 1990. A real-time french text-to-speech system generating high-quality synthetic speech. En *Proceedings of ICASSP*, volumen 1, páginas 309–312, Albuquerque, New Mexico.

Myers, K., M. Kearns, S. Singh, y M. Walker. 2000. A boosting approach to topic spotting on subdialogues. En *Proceedings of the 17th International Conference on Machine Learning (ICML)*, página 655662., Stanford, USA.

MySQL. 2003. http://www.mysql.com.

Nakata, T., T. Ikeda, S. Ando, y A. Okumura. 2002. Topic Detection Based on Dialogue History. En *Proceedings of the ACL Workshop on Speech-to-Speech Translation: Algorithms and Systems*, páginas 9–14, Philadelphia, USA.

Nakatani, C.H. y J. Chu-Carroll. 2000. Using dialogue representations for Concept-to-Speech generation. En *Proceedings of the ANLP/NAACL 2000 Workshop on Conversational Systems*, páginas 48–53, Seattle, USA. Association for Computational Linguistics.

Ney, H. 1982. A time warping approach to fundamental period estimation. *IEEE Transactions on Systems, Man and Cybernetics*, 12(3):383–388.

Ngoc, T. V. y C. d'Alessandro. 1999. Robust glottal closure detection using the wavelet transform. En *Proceedings of the European Conference on Speech Technology (EuroSpeech)*, páginas 2805–2808, Budapest, Hungría.

Oversdotter, C. y X. Llorà. 2006. Evolving Emotional Prosody. IlliGAL Report No. 2006018, University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, Urbana, IL.

Oversdotter, C. y R. Sproat. 2005. Perceptions of emotions in expressive storytelling. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 533–536, Lisboa, Portugal.

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

Pakucs, B. 2003. Towards Dynamic Multi-Domain Dialogue Processing. En *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech)*, páginas 741–744, Ginebra, Suiza.

Pakucs, B. 2004. Butler: A Universal Speech Interface for Mobile Environments. *Lecture Notes in Computer Science (International Symposium on Human Computer Interaction with Mobile Devices and Services)*, (3160):399–403, Setiembre.

Pantazis, Y., Y. Stylianou, y E. Klabbers. 2005. Discontinuity Detection in Concatenative Speech Synthesis based on Nonlinear Speech Analysis. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 2817–2820, Lisboa, Portugal.

Pareto, Vilfredo. 1896. *Cours d'Economie Politique, volume I and II*. F. Rouge, Lausanne.

Park, S.S., C.K. Kim, y N.S. Kim. 2003. Discriminative weight training for unit-selection based speech synthesis. En *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech)*, volumen 1, páginas 281–284, Ginebra, Suiza.

Peng, F. y D. Schuurmans. 2003. Combining Naive Bayes and n-Gram Language Models for Text Classification. *Lecture Notes in Computer Science (Advances in Information Retrieval: Proceedings of The 25th European Conference on Information Retrieval Research (ECIR03))*, 2633:335–350.

Peng, F., D. Schuurmans, y S. Wang. 2003. Language and Task Independent Text Categorization. En *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL'03)*, páginas 110–117, Edmonton, Canadá.

Peng, H., Y. Zhao, y M. Chu. 2002. Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation. En *Proceedings of ICSLP*, páginas 1341–1344, Denver, USA.

Plante, F., G. Meyer, y W.A. Ainsworth. 1995. A pitch extraction reference database. En *Proceedings of EuroSpeech*, páginas 837–840, Madrid.

Portele, T., S. Goronzy, M. Emele, A. Kellner, S. Torge, y J. te Vrugt. 2003. SmartKom-Home - An Advanced Multi-Modal Interface to Home Entertainment. En *Proceedings of EuroSpeech*, páginas 1897–1900, Geneve.

Porter, M.F. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Prasanna, S.R.M y B. Yegnanarayana. 2004. Extraction of pitch in adverse conditions. En *Proceedings of ICASSP*, volumen 1, páginas 109–112, Montreal, Canadá.

Pérez-Piñar, D. y C. García. 2005. Application of Confidence Measures for Dialogue Systems through the Use of Parallel Speech Recognizers. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 2785–2788, Lisboa, Portugal.

Quast, H., O. Schreiner, y M.R. Schroeder. 2002. Robust pitch tracking in the car environment. En *Proceedings of ICASSP*, volumen 1, páginas 353–356, Orlando, USA.

Quazza, S., L. Donetti, L. Moisa, y P. L. Salza. 2001. ACTOR©: a multilingual unit-selection speech synthesis system. En *The 4th ISCA Workshop on Speech Synthesis*, Perthshire, Escocia.

Quinlan, J. R. 1993. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, CA.

Rabiner, L. y B. Juang. 1993. *Fundamentals of Speech Recognition*. Prentice Hall.

Rabiner, L. R. y Schafer. 1978. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ.

Rabiner, L.R, M. J. Cheng, A. E. Rosenberg, y C. A. McGonegal. 1976. A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24(5):399–418, Octubre.

Raux, A., B. Langner, A.W. Black, y M. Eskenazi. 2005. Lets Go Public! Taking a Spoken Dialog System to the Real World. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 885–888, Lisboa, Portugal.

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.

Rodríguez, A., P. Lázaro, N. Montoya, J. Blanco, D. Bernadas, J. Oliver, y L. Longhi. 1999. Modelización acústica de la expresión emocional en el español. *Procesamiento del Lenguaje Natural*, 25:159–166, Septiembre.

Rodríguez Banga, E., F. Campillo, E. Fernández, y F. Méndez. 2002. Sistema de conversión texto-voz en lengua gallega basado en la selección combinada de unidades acústicas y prosódicas. *Procesamiento del Lenguaje Natural*, 29:153–158, Septiembre.

Rudnicky, A.I., C. Bennett, A.W. Black, A. Chotomongcol, K. Lenzo, A. Oh, y R. Singh. 2000. Task and Domain Specific Modelling in the Carnegie Mellon Communicator System. En *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volumen 2, páginas 130–134, Beijing, China.

Rüggenmann, Klaus y Iryna Gurevych. 2004. Assigning domains to speech recognition hypotheses. En Srinivas Bangalore y Hong-Kwang Jeff Kuo, editores, *Proceedings of HLT-NAACL Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing*, páginas 70–77, Boston, Massachusetts, USA. Association for Computational Linguistics.

Rutten, P., G. Coorman, J. Fackrell, , y B. Van Coile. 2000. Corpus based speech synthesis in the Lernout & Hauspie RealSpeak TTS system. En *Proceedings of IEEE symposium on State-of-the Art in Speech Synthesis*, páginas 16/1–16/7, Savoy Place, Londres.

Sagisaka, Y. 1988. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. En *Proceedings of ICASSP*, páginas 679–682, New York, USA.

Sagisaka, Y., N. Kaiki, N. Iwahashi, y K. Mimura. 1992. ATR - ν-TALK speech synthesis system. En *Proceedings of ICSLP*, volumen 1, páginas 483–486, Banff, Canadá.

Sagisaka, Y., T. Yamashita, y Y. Kokenawa. 2004. Speech Synthesis with Attitude. En *Proc. of Speech Prosody*, páginas 401–404, Nara, Japón.

Sagisaka, Y., T. Yamashita, y Y. Kokenawa. 2005. Generation and perception of F_0 markedness for communicative speech synthesis. *Speech Communication*, 46(1):376–384.

Saito, T. y M. Sakamoto. 2005. A VoiceFont Creation Framework for Generating Personalized Voices. *IEICE Transactions*, 88-D(3):525–534.

Sakamoto, M. y T. Saito. 2000. An Automatic Pitch-Marking Method using Wavelet Transform. En *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volumen 3, páginas 650–653, Beijing, China.

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

Salton, G. y C. Buckley. 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 25(4):513–523.

Salton, G. y M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

Sassano, M. 2003. Virtual Examples for Text Classification with Support Vector Machines. En *Proceedings of 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, páginas 208–215, Japón.

Sastry, K. y D. E. Goldberg. 2002. Genetic algorithms, efficiency enhancement, and deciding well with fitness functions with differing bias values. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2002)*, páginas 536–543. (Also IlliGAL Report No. 2002003).

Sato, Y. 1996. Voice conversation using evolutionary computation of prosodic control. En *12th Symposium on Human Interface*, páginas 469–475.

Sato, Y. 1997. Voice conversation using evolutionary computation of prosodic control. En *Intelligent Processing of Manufacturing of Materials '97*, páginas 342–348.

Sato, Y. 2005. Voice quality conversion using interactive evolution of prosodic control. *Applied Soft Computing Journal*, (5):181–192.

Schapire, R.E. y Y. Singer. 2000. BoosTexter: A Boosting-based System for Text Categorization. *Macine Learning*, 39(2/3):135–168.

Schröder, M. 2004. Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. *Lecture Notes in Artificial Intelligence (Tutorial and Research Workshop on Affective Dialog Systems)*, (3068):209220, Junio.

Schweitzer, A., N. Braunschweiler, T. Klankert, B. Säuberlich, y B. Möbius. 2003. Restricted Unlimited Domain Synthesis. En *Proceedings of EuroSpeech*, páginas 1321–1324, Geneve.

Sebag, M. y Q. Ducoulombier. 1998. Extending Population-Based Incremental Learning to Continuous Search Spaces. *Lecture Notes in Computer Science*, 1498:418–427.

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

Sebastiani, F. 2005. Text categorization. En Alessandro Zanasi, editor, *Text Mining and its Applications*. WIT Press, Southampton, UK, capítulo 4, páginas 109–129.

Secrest, B.G. y G.R. Doddington. 1982. Postprocessing techniques for voice pitch trackers. En *Proceedings of ICASSP*, volumen 7, páginas 172–175, París, Francia.

Seneff, S., E. Hurley, R. Lau, C. Pao, P. Schmid, y V. Zue. 1998. GALAXY-II: A reference architecture for conversational system development. En *Proceedings of ICSLP*, página 931934, Sydney, Australia.

Sethy, A., P.G. Georgiou, y S.Ñarayanan. 2005. Building Topic Specific Language Models From Webdata Using Competitive Models. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 1293–1296, Lisboa, Portugal.

Sevillano, X., F. Alías, y J.C. Socoró. 2004. ICA-Based Hierarchical Text Classification for Multi-domain Text-to-Speech Synthesis. En *Proceedings of ICASSP*, volumen 5, páginas 697–700, Montreal, Canadá.

Sevillano, X., G. Cobo, F. Alías, y J.C. Socoró. 2006a. Feature Diversity in Cluster Ensembles for Robust Document Clustering. En *The 29th Annual International ACM SIGIR*, Seattle, USA.

Sevillano, X., G. Cobo, F. Alías, y J.C. Socoró. 2006b. Robust Document Clustering by Exploiting Feature Diversity in Cluster Ensembles. *Procesamiento del Lenguaje Natural*, 37, Setiembre.

Sha, F., J.A. Burgoyne, y L.K. Saul. 2004. Multiband statistical learning for f_0 estimation in speech. En *Proceedings of ICASSP*, volumen 5, páginas 661–664, Montreal, Canadá.

Sha, F. y L. Saul. 2005. Real-time pitch determination of one or more voices by nonnegative matrix factorization. En Lawrence K. Saul Yair Weiss, y Léon Bottou, editores, *Advances in Neural Information Processing Systems*, número 17. MIT Press, Cambridge, MA, páginas 1233–1240.

Shawe-Taylor, J. y N. Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Sproat, R. 1997. Text Interpretation for TTS Synthesis. En R.A. Cole J. Mariani H. Uszkoreit A. Zaenen, y V. Zue, editores, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge, UK.

Stamatatos, E., G. Kokkinakis, y N. Fakotakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471 – 495.

Stöber, K., T. Portele, P. Wagner, y W. Hess. 1999. Synthesis by Word Concatenation. En *Proceedings of EuroSpeech*, volumen 2, páginas 619–622, Budapest, Hungría.

Stylianou, Y. 1996. *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modifications*. Tesis Doctoral, École Nationale des Télécommunications, París, Francia.

Stylianou, Y. 1998. Removing phase mismatches in concatenative speech synthesis. En *Proceedings of the Third ESCA Workshop in Speech Synthesis*, páginas 267–272, Jenolan Caves, Australia.

Stylianou, Y. 1999. Synchronization of Speech Frames Based on Phase Data with Applica-tion to Concatenative Speech Synthesis. En *Proceedings of EuroSpeech*, páginas 2343–2346, Budapest, Hungría.

Stylianou, Y. 2001. Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis. *IEEE Transactions on Speech and audio Processing*, 9(1):21–29, Enero.

Stylianou, Y., T. Dutoit, y J. Schroeter. 1997. Diphone Concatenation using a Harmonic plus Noise Model of Speech. *Proceedings of EuroSpeech*, páginas 613–616.

Stylianou, Y. y A. K. Syrdal. 2001. Perceptual and Objective Detection of Discontinuities in Concatenative Speech Synthesis. En *Proceedings of ICASSP*, volumen 2, páginas 987–990, Salt Lake City, USA.

Sugimoto, F., K. Yazu, M. Murakami, y M. Yoneyama. 2004. A method to classify emotional expressions of text and synthesize speech. En *Proc. of 1st Int. Symposium on Control, Communications and Signal Processing*, páginas 611 – 614, Hammamet, Túnez.

Sun, X. 2000. A pitch determination algorithm based on Subharmonic-to-Harmonic Ratio. En *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volumen 4, páginas 676–679, Beijing, China.

Sun, X. 2002. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. En *Proceedings of ICASSP*, volumen 1, páginas 333–336, Orlando, USA.

Sundaram, S. y S.Ñarayanan. 2003. An empirical text transformation method for spontaneous speech synthesizers. En *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech)*, volumen 2, páginas 1221–1224, Ginebra, Suiza.

Sundaraman, S. y S.Ñarayanan. 2002. Spoken language synthesis: Synthesis of spontaneous monolog speech. En *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, USA.

Syrdal, A., G. Moehler, K. Dusterhoff, Conkie A., y A.W. Black. 1998. Three Methods of Intonation Modeling. En *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, páginas 305–310, Jenolan Caves, Australia.

Syrdal, A. K., C. W. Wightman, A. Conkie, M. Stylianou, Y. Beutnagel, J. Schroeter, V. Strom, K. Lee, y M. Makashay. 2000. Corpus-based techniques in the AT&T NextGen synthesis system. En *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volumen 3, páginas 410–415, Beijing, China.

Syrdal, A.K. y A.D. Conkie. 2004. Data-driven Perceptually-based Join Costs. En *Proc. 5th ISCA Speech Synthesis Workshop*, páginas 49–54, Pittsburgh, USA.

Syrdal, A.K. y A.D. Conkie. 2005. Perceptually-based Data-driven Join Costs: Comparing Join Types. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 2813–2816, Lisboa, Portugal.

Tachibana, M., J. Yamagishi, K. Onishi, T. Masuko, y T. Kobayashi. 2004. HMM-Based Speech Synthesis with Various Speaking Styles Using Model Interpolation. En *Proceedings of Speech Prosody*, páginas 413–416, Nara, Japón.

Takagi, H. 2001. Interactive Evolutionary Computation: fusion of the capabilities of the EC Optimization and Human Evaluation. *Proceedings of the IEEE*, 89(9):1275–1296.

Takeda, K., A. Katsuo, y Y. Sagisaka. 1990. On unit selection algorithms and their evaluation in non-uniform speech synthesis. En *Proceedings of ESCA Workshop on Speech Synthesis*, páginas 35–38, Autrans, Francia.

Talkin, D. 1995. A Robust Algorithm for Pitch Tracking (RAPT). En W. B. Kleijn y K. K. Paliwal, editores, *Speech Coding and Synthesis*. Elsevier Science, Amsterdam, NL, capítulo 14, páginas 495–518.

Tao, J. y T. Tan. 2004. Emotional Chinese Talking Head System. En *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI)*, páginas 273 – 280, State College, PA, USA.

Taylor, P. 2000. Concept-to-Speech synthesis by phonological structure matching. *Philosophical Transactions of the Royal Society, Series A*, 356(1769):1403–1416.

Taylor, P. y A.W. Black. 1999. Speech synthesis by phonological structure matching. En *Proceedings of EuroSpeech*, volumen 4, páginas 1531–1534, Budapest, Hungría.

Taylor, P., A.W. Black, y R. Caley. 1998. The architecture of the festival speech synthesis system. En *Proceedings of the Third ESCA Workshop in Speech Synthesis*, páginas 147–151, Jenolan Caves, Australia.

Taylor, P., A.W. Black, y R. Caley. 2000-2003. Building Voices in the Festival Speech Synthesis System (DRAFT). En *http://festvox.org/bsv/*.

Teixeira, J.P., D. Freitas, D. Braga, M.J. Barros, y V. Latsch. 2001. Phonetic events from the labeling the European Portuguese database for speech synthesis, FEUP/IPBDB. En *Proceedings of EuroSpeech*, páginas 1707–1710, Aalborg, Dinamarca.

Tesser, F., P. Cosi, C. Drioli, y G. Tisato. 2005. Emotional Festival-MBROLA TTS Synthesis. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 505–508, Lisboa, Portugal.

Toda, T., H. Kawai, y M. Tsuzaki. 2003. Optimizing integrated cost function for segment selection in concatenative speech synthesis based on perceptual evaluations. En *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech)*, volumen 1, páginas 297–300, Ginebra, Suiza.

Toda, T., H. Kawai, y M. Tsuzaki. 2004. Optimizing Sub-Cost Functions for Segment Selection Based on Perceptual Evaluations in Concatenative Speech Synthesis. En *Proceedings of ICASSP*, páginas 657–660, Montreal, Canadá.

Toda, T., H. Kawai, M. Tsuzaki, y K. Shikano. 2002. Perceptual Evaluation of cost for Segment Selection in Concatenative Speech Synthesis. En *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, USA.

Toda, T. y K. Tokuda. 2005. Speech Parameter Generation Algorithm Considering Global Variance for HMM-based speech synthesis. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 2801–2804, Lisboa, Portugal.

Toda, Tomoki. 2003. *High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion*. Tesis Doctoral, Graduate School of Information Science, Nara Institute of Science and Technology, Marzo.

Todoroki, Y. y H. Takagi. 2000. User interface of an interactive evolutionary computation for speech processing. En *6th International Conference on Soft Computing (IIZUKA2000)*, páginas 112–118.

Tombros, Anastasios. 2002. *The effectiveness of query-based hierarchic clustering of documents for information retrieval*. Tesis Doctoral, Faculty of Computing Science, Mathematics and Statistics. University of Glasgow, Marzo.

Tsuzaki, M. y H. Hisashi. 2002. Feature extraction for unit selection in concatenative speech synthesis: comparision between AIM, LPC and MFCC. En *Proceedings of ICSLP*, volumen 1, páginas 137–140, Denver, USA.

Tur, G. 2004. Cost-Sensitive Call Classification. En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 1581–1584, Jeju Island, Corea del Sur.

Turk, O., M. Schröder, B. Bozkurt, y L.M. Arslan. 2005. Voice quality interpolation for emotional Text-to-Speech synthesis. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 797–800, Lisboa, Portugal.

van Santen, J. P. H. 1997. Combinatorial issues in text-to-speech synthesis. En *Proceedings of EuroSpeech*, volumen 5, páginas 2511–2514, Rodas, Grecia.

van Santen, Jan P. H. y Adam L. Buchsbaum. 1997. Methods for Optimal Text Selection. En *Proceedings of EuroSpeech*, páginas 553–556, Rodas, Grecia.

van Son, R. J. J. H., H. Binnenpoorte, D. van den Heuvel, y L. C. W. Pols. 2001. The IFA Corpus: a Phonemically Segmented Dutch "Open Source"Speech Database . En *Proceedings of EuroSpeech*, páginas 2051–2054, Aalborg, Dinamarca.

Veldhuis, R. 2000. Consistent Pitch Marking. En *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volumen 3, páginas 207–210, Beijing, China.

Veldhuis, R. y E. Klabbers. 2003. On the Computation of the Kullback-Leibler Measure for Spectral Distances. *IEEE Transactions on Speech and Audio Processing*, 11(1):100–103.

Vepa, J., J. Ayachitam, y K. V. K. Kalpana Reddy. 2002. A text-to-speech synthesis system for Telugu. En *Proceedings of ICSLP*, volumen 1, páginas 157–160, Denver, USA.

Vepa, J., S. King, y P. Taylor. 2002. Objective distance measures for spectral dicontinuities in concatenative speech synthesis . En *Proceedings of ICSLP*, volumen 4, páginas 2604–2608, Denver, USA.

Vilares, J., F.M. Barcala, y M.A. Alonso. 2001. Normalización de términos multipalabra mediante sintáctica. *Procesamiento del Lenguaje Natural*, 27:123–130, Septiembre.

Vincent, D., O. Rosec, y T. Chonavel. 2006. Glottal Closure Instant estimation using an appropriateness measure of the source and continuity constraints. En *Proceedings of ICASSP*, volumen I, páginas 381–384, Toulouse, Francia.

Viterbi, A. J. 1967. Error Bounds for Convolutional Codes and an Asymtotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, 13:260–267.

Wahlster, W., N. Reithinger, y A. Blocher. 2001. SmartKom: Multimodal communication with a life-like character. En *Proceedings of EuroSpeech*, volumen 2, página 15471550, Aalborg, Dinamarca.

Wang, C. y S. Seneff. 2000. Robust pitch tracking for prosodic modeling in telephone speech. En *Proceedings of ICASSP*, volumen 3, páginas 1343–1346, Istanbul, Turquía.

Watanabe, T. y H. Takagi. 1995. Recovering sytem of the distorted speech using interactive genetic algorithms. En *IEEE, International Conference on Systems, Man and Cybernetics (SMC'95)*, volumen 1, páginas 684–689.

Wells, J., W. Barry, M. Grice, A. Fourcin, y D. Gibbon. 1992. Standard Computer-Compatible Transcription. Final Report. ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation SAM-UCL-037, Phonetics and Linguistics Department, University College London, Londres.

Willet, P. 1983. Similarity coefficients and weighting functions for automatic document classification: an empirical comparison. *International Classification*, 10:138–142.

Witten, Ian H. y Eibe Frank. 2000. *Data Mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann.

Wouters, J. y M. Macon. 2001. Control of spectral dynamics in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(1):30–38.

Wu, C.-H. y J.-H. Chen. 2001. Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis. *Speech Communication*, 35:219–237.

Yamagishi, J., K. Onishi, T. Masuko, y T. Kobayashi. 2003. Modeling of Various Speaking Styles and Emotions for HMM-Based Speech Synthesis. En *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech)*, volumen III, páginas 2461–2464, Geneve, Suiza.

Yamagishi, J., K. Onishi, T. Masuko, y T. Kobayashi. 2005. Acoustic Modelling of Speaking Styles and Emotional Expressions in HMM-based Speech Synthesis. *IEICE Transactions on Information and Systems*, E88D(3):502–509.

Yamagishi, J., M. Tachibana, T. Masuko, y T. Kobayashi. 2004. Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis. En *Proceedings of ICASSP*, volumen 1, páginas 5–8, Montreal, Canadá.

Yang, Y. y X. Liu. 1999. A re-examination of text categorization methods. En *SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, páginas 42–49, Berkeley, USA.

Yang, Y. y J.O. Pedersen. 1997. A comparative study on feature selection in text categorization. En Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, páginas 412–420, Nashville, USA. Morgan Kaufmann Publishers, San Francisco, US.

Yi, J. y J. Glass. 1998. Natural-sounding speech synthesis using variable-length units. En *Proceedings of ICSLP*, páginas 1167–1170, Sydney, Australia.

Yi, J. y J. Glass. 2002. Information-theoretic criteria for unit selection synthesis. En *Proceedings of ICSLP*, volumen 4, páginas 2617–2620, Denver, USA.

Yi, Jon Rong-Wei. 2003. *Corpus-Based Unit Selection for Natural-Sounding Speech Synthesis*. Tesis Doctoral, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Ying, S.G., L.H. Jamieson, y C.D. Michell. 1996. A probabilistic approach to AMDF pitch detection. En *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, páginas 1201–1204, Philadelphia, USA.

Yu, A-T. y H-C Wang. 2004. New harmonicity measures for pitch estimation and voice activity detection. En *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, páginas 2429–243, Jeju Island, Corea del Sur.

Zen, H. y T. Toda. 2005. An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005. En *Proceedings of the 9th International Conference on Speech Communication and Technology (InterSpeech)*, páginas 93–96, Lisboa, Portugal.

Zhao, Y., P. Liu, Y. Li, Y. Chen, y M. Chu. 2006. Measuring Target Cost in Unit Selection with KL-Divergence between Context-Dependent HMMs. En *Proceedings of ICASSP*, volumen I, páginas 725–728, Toulouse, Francia.

Zhu, W., W. Zhang, Q. Shi, y F. Chen. 2002. Corpus Building for Data-Driven TTS System. En *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, USA.

Universitat Ramon Llull

Esta Tesis Doctoral ha sido defendida el día ____ de_____ de 2006 en el

Centro _____

de la Universitat Ramon Llull

delante del Tribunal formado por los Doctores abajo firmantes, habiendo obtenido la calificación:

Presidente/a

_____

Vocal

_____

Vocal

_____

Vocal

_____

Secretario/aria

_____

Doctorando/a

_____

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es