

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR



-PROYECTO FIN DE CARRERA-

DESARROLLO DE UN SISTEMA DE
RECONOCIMIENTO DE HABLA
NATURAL PARA TRANSCRIBIR
CONTENIDOS DE AUDIO EN
INTERNET

Juan Manuel Perero Codosero

Marzo 2015

DESARROLLO DE UN SISTEMA DE RECONOCIMIENTO DE HABLA NATURAL PARA TRANSCRIBIR CONTENIDOS DE AUDIO EN INTERNET

AUTOR: Juan Manuel Perero Codosero

TUTOR: Daniel Tapias Merino

PONENTE: Doroteo Torre Toledano



Área de Tratamiento de Voz y Señales
Dpto. Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Marzo 2015

Resumen

El objetivo de este proyecto es desarrollar un sistema de reconocimiento de habla natural con el fin de transcribir contenidos de audio de Internet.

En primer lugar, se realizará un estudio del estado del arte para conocer la arquitectura típica de los sistemas de reconocimiento de voz y el uso de Modelos Ocultos de Markov para esta tarea.

Tras la implementación de un sistema de referencia, el ajuste de parámetros y su posterior evaluación, se llevará a cabo una optimización del sistema usando modelos del lenguaje adaptados al tópico de los contenidos a reconocer.

Por último, para demostrar su funcionamiento, se aplicará este sistema optimizado a una solución comercial, permitiendo así ampliar su funcionalidad.

En esta memoria se recogen los resultados de todas las pruebas y las conclusiones obtenidas.

Palabras Clave

Reconocimiento de voz, Modelos Ocultos de Markov, modelo del lenguaje, corpus, tópico, contenidos audiovisuales.

Abstract

The aim of this project is to develop a speech recognition system in order to transcribe Internet audiovisual content.

First of all, a state of the art research will be carried out to determine the common architecture of speech recognition systems and the use of Hidden Markov Models for this task.

After the implementation of a reference system, parameter adjustment and subsequent evaluation, a system optimization will be performed using topic-based language models for the content to recognize.

Finally, in order to demonstrate its operation, the optimized system will be applied to a commercial solution enabling the functionality to be extended.

The results of each test are captured in this document, in addition to obtained conclusions.

Keywords

Speech recognition, Hidden Markov Models, language models, corpus, topic, audiovisual content.

Agradecimientos

En primer lugar, me gustaría agradecer a mi tutor, Daniel Tapias, la oportunidad que me ha brindado de iniciarme en el mundo profesional, realizando este proyecto. Su atención y apoyo constante me han servido de guía en la consecución del mismo.

Quería también agradecer a todos los miembros de Sigma Technologies y de Tax Planning, la acogida y el cariño que me han mostrado. Especialmente a Jorge Rico, que ha presenciado en primera persona como avanzaba el proyecto, aportando siempre que era necesario, sus conocimientos y experiencia. Y a mi compañero Javier Antón, con el que he trabajado codo con codo y aprendido multitud de cosas.

El mayor de todos los agradecimientos va dirigido a mis padres Juan Manuel y Rosa María, a quienes debo todo lo que soy. Siempre han dado todo por mí, y su confianza y apoyo han sido imprescindibles. Y por supuesto a mi abuela Encarna, incondicionalmente orgullosa de mí.

Me gustaría hacer una mención especial a mi compañero Pencho, con el que he compartido “unos cuantos” laboratorios y bibliotecas. Su personalidad y generosidad han permitido forjar nuestra amistad.

Además, quería agradecer a mis compañeros y amigos, Raúl, Mario, Jorge, Guille, David y Ángel, todas las experiencias compartidas. Cada uno de ellos sabe por qué es especial para mí. A Sara, que contribuyó al arranque de esta memoria, y al resto de compañeros con los que he compartido esta etapa.

No quisiera olvidarme de mis amigos de toda la vida y del “Alcobendas United”, especialmente Marco, Isma, Robert, Nando, Jony, Álvaro, Rubén, Nacho y Sergio por todos estos años de amistad. Hemos crecido juntos y siempre han estado ahí.

Por último, quiero dar las gracias a todas las personas que se han cruzado conmigo a lo largo de estos años y, de una forma u otra, han contribuido a que llegase donde ahora estoy.

Gracias a todos de corazón.

Juan Manuel Perero Codosero.

Madrid, Marzo de 2015.

Quiero expresar el más sincero agradecimiento al Banco Santander por la concesión de la “Beca de Prácticas Santander CRUE-CEPYME”, que ha servido como complemento a mi formación y contribuido a mi acercamiento al ámbito profesional.



También quiero mostrar el agradecimiento a Telefónica I+D por la cesión de los derechos de las bases de datos que se han empleado en el desarrollo del sistema realizado como objetivo de este proyecto.



Además, quiero dar las gracias al Biometric Recognition Group - ATVS por facilitar la oferta de este proyecto fin de carrera y haberme concedido la oportunidad de realizarlo.



Por último, y no menos importante, agradecer a Sigma Technologies la oportunidad brindada de realizar este proyecto, y por supuesto, de la acogida por parte de todos sus miembros y el conocimiento adquirido de cada uno de ellos.



Índice general

Resumen	v
Abstract	vii
Agradecimientos	ix
Acrónimos	xix
1. Introducción	1
1.1. Motivación	1
1.2. Objetivos	1
1.3. Estructura de la memoria	2
2. Estado del Arte	3
2.1. Introducción general del reconocimiento de voz	3
2.2. Arquitectura de un reconocedor automático de habla	5
2.2.1. Preprocesado	6
2.2.2. Reconocimiento de patrones	7
2.2.3. Decisión	15
2.3. Reconocimiento con HMMs	15
2.3.1. Definición y caracterización	15
2.3.2. Los tres problemas básicos de los HMMs	18
3. Base de Datos	27
3.1. Preparación de los datos	27
4. Generación del Sistema de Referencia	31
4.1. Introducción	31
4.1.1. Software utilizado	32
4.2. Proceso de entrenamiento	34
4.2.1. Discriminación de los archivos de audio	35
4.2.2. Generación de archivos “.lab”	35
4.2.3. Generación del diccionario	35
4.2.4. Generación de los archivos de transcripciones y listas	36
4.2.5. Generación del modelo de lenguaje	37
4.2.6. Extracción de características	38

4.2.7. Generación de los HMM	38
4.3. Reconocimiento de verificación	43
5. Optimización del Sistema y Resultados	49
5.1. Descripción del problema	49
5.2. Escenario y generación de modelos	50
5.2.1. Tópicos	50
5.2.2. Corpus de entrenamiento de los modelos	51
5.2.3. Diccionarios específicos	52
5.3. Base de Datos de Test	54
5.4. Experimentos	54
5.4.1. Condiciones iniciales: Características del sistema de referencia .	55
5.4.2. Descripción general de los experimentos	56
5.4.3. Experimento 1: Modelos del lenguaje acotados en cada tópico .	59
5.4.4. Experimento 2: Modelo del lenguaje genérico para todos los tópicos	60
5.4.5. Experimento 3: Modelos del lenguaje mezclando corpus	63
5.4.6. Experimento 4: Diccionario genérico	64
5.4.7. Experimento 5: Diccionario genérico con vocabulario específico	65
5.4.8. Experimento 6: Modelos de lenguaje con mezclas ponderadas .	66
5.4.9. Experimento 7: Modelos de lenguaje con mezclas ponderadas (2 ^a pasada, 3-gramas)	68
5.4.10. Experimento 8: Modelos de lenguaje con mezclas ponderadas (2 ^a pasada, 5-gramas)	69
5.4.11. Experimento 9: Modelo del lenguaje global con mezclas ponde- radas	71
5.4.12. Experimento 10: Comparativa con los líderes del mercado	72
6. Aplicación del Sistema	75
6.1. Demostración	75
7. Conclusiones y Trabajo Futuro	79
7.1. Conclusiones	79
7.2. Trabajo futuro	81
Bibliografía	81
A. Elección número de gaussianas	85
B. Elección número ficheros de entrenamiento	93
C. Resultados exhaustivos de la optimización	97
D. Presupuesto	109
E. Pliego de condiciones	111

Índice de figuras

2.1.	Esquema básico de la arquitectura de un RAH.	6
2.2.	Función de alineamiento de DTW.	9
2.3.	Espacio bidimensional dividido en regiones (VQ).	10
2.4.	Definición de gramática.	12
2.5.	Diagrama de estados de la gramática.	12
2.6.	Cálculo de probabilidades de un modelo del lenguaje (bigramas).	14
2.7.	Representación gráfica de un HMM de 6 estados.	16
2.8.	Esquema del algoritmo de Viterbi.	23
3.1.	Archivos necesarios para HTK.	28
3.2.	Histograma distribución SNR de la base de datos.	30
3.3.	Histograma distribución <i>pitch</i> promedio de la base de datos.	30
4.1.	Logotipo de HTK3.	33
4.2.	Logotipo del software Julius.	34
4.3.	Entrenamiento de HMMs de fonemas.	39
4.4.	Ejemplo de archivo de 'macros' y de 'hmmdefs'.	39
4.5.	Mejora del modelo de silencio.	40
4.6.	Ejemplo de modelo de estados ligados.	42
4.7.	Ejemplo de suma de múltiples gaussianas.	43
4.8.	Gráfica selección de gaussianas.	46
4.9.	Gráfica selección número ficheros entrenamiento.	47
4.10.	Gráfica selección número ficheros entrenamiento (semilogarítmico).	47
5.1.	Tendencia tasa de acierto de palabra en función del <i>pitch</i> (Moda).	61
5.2.	Variación corpus entrenamiento LM genérico para cada tópico.	61
5.3.	Variación corpus entrenamiento LM genérico para Auto-test y Test.	62
5.4.	Variación corpus de entrenamiento LM genérico y específico.	63
5.5.	Variación diccionario genérico con LM genérico fijo.	65
5.6.	Variación ponderaciones de la mezcla de LMs (genérico y específico)	67
5.7.	Comparativa reconocimiento con N-gramas (Economía).	69
5.8.	Comparativa reconocimiento con N-gramas (Deporte).	70
5.9.	Comparativa reconocimiento con N-gramas (Moda).	70
5.10.	Ejemplo salida reconocimiento del sistema desarrollado.	73
5.11.	Ejemplo salida reconocimiento de la Web Speech API de Google.	73

5.12. Ejemplo salida reconocimiento del Dictado Automático de Apple.	74
6.1. Imagen de la aplicación. Búsqueda concreta.	76
6.2. Imagen de la aplicación. Diferentes apariciones en el vídeo.	76
B.1. Gráfica selección número ficheros entrenamiento.	93
B.2. Gráfica selección número ficheros entrenamiento (semilogarítmico).	94

Índice de tablas

3.1. Características de la base de datos.	29
4.1. Grupos de archivos de prueba.	45
4.2. Resultados pruebas test para 16 gaussianas.	48
5.1. Número de palabras específicas de cada tópico.	53
5.2. Características de los contenidos audiovisuales del entorno de test.	54
5.3. Condiciones establecidas en el Experimento 1.	59
5.4. Tasa de acierto de palabra con modelos del lenguaje acotados.	59
5.5. Relación directa tasa de acierto de palabra y <i>pitch</i> promedio (Moda).	60
5.6. Condiciones establecidas en el Experimento 2.	61
5.7. Condiciones establecidas en el Experimento 3.	63
5.8. Condiciones establecidas en el Experimento 4.	64
5.9. Condiciones establecidas en el Experimento 5.	65
5.10. Influencia sustracción y adición de palabras en tasa de acierto.	65
5.11. Condiciones establecidas en el Experimento 6.	66
5.12. Condiciones establecidas en el Experimento 7.	68
5.13. Comparativa tasa de acierto entre 1 ^a y 2 ^a pasada (3-gram).	68
5.14. Condiciones establecidas en el Experimento 8.	69
5.15. Condiciones establecidas en el Experimento 9.	71
5.16. Tasa de acierto de palabra según mezcla global ponderada.	71
5.17. Comparativa tasa de acierto con sistemas de Google y Apple.	74
A.1. Tasa de acierto frente al número de gaussianas (Auto-test 10-20 dB).	86
A.2. Tasa de acierto frente al número de gaussianas (Auto-test 20-30 dB).	87
A.3. Tasa de acierto frente al número de gaussianas (Auto-test 30-40 dB).	88
A.4. Tasa de acierto frente al número de gaussianas (Auto-test 40-99 dB).	89
A.5. Tasa de acierto frente al número de gaussianas (Test 30-40 dB).	90
A.6. Tasa de acierto frente al número de gaussianas (Test ruidoso).	91
B.1. Tasa de acierto frente al número de ficheros de entrenamiento.	95
C.1. Resultados completos Experimento 1.	97
C.2. Resultados completos Experimento 2. Tópicos.	98
C.3. Resultados completos Experimento 2. Auto-test y Test.	99
C.4. Resultados completos Experimento 3. Corpus específico (200K palabras).	100

C.5. Resultados completos Experimento 3. Corpus específico (20M palabras).	101
C.6. Resultados completos Experimento 4.	101
C.7. Resultados completos Experimento 5.	102
C.8. Resultados completos Experimento 6.	103
C.9. Resultados completos Experimento 7.	104
C.10 Resultados completos Experimento 8.	105
C.11 Resultados completos Experimento 9.	106
C.12 Resultados completos Experimento 10.	107

Acrónimos

ARPA	Advanced Research Projects Agency
ASR	Automatic Speech Recognition
CMS	Cepstral Mean Substraction
CMU	Carnegie Mellon University
CVN	Cepstral Variance Normalization
DCT	Discrete Cosine Transformate
DFT	Discrete Fourier Transformate
DNN	Deep Neural Networks
DTW	Dynamic Time Warping
FFT	Fast Fourier Transformate
HMM	Hidden Markov Model (Modelos Ocultos de Markov)
HTK	HMM ToolKit
LPC	Linear Predictive Coding
LVCSR	Large Vocabulary Continuous Speech Recognition
MFCC	Mel Frequency Cepstral Coefficient
PLP	Perceptual Linear Prediction
RAH	Reconocimiento de habla automático
SNR	Signal to Noise Ratio

Capítulo 1

Introducción

1.1. Motivación

El reconocimiento automático de voz ha tenido una gran evolución gracias a su amplia utilidad y a las facilidades que ha introducido en el desarrollo de diversas tareas. El hecho de liberar las manos de cualquier teclado o dispositivo de entrada en el control de procesos, supone una gran ventaja a la hora de introducir estos sistemas en nuestra vida diaria.

La posibilidad de obtener una representación simbólica discreta de una señal de voz continua, permite obtener en formato de texto la información vocal pronunciada por el hablante. Esta tarea es implementada en los llamados sistemas de reconocimiento automático de habla (*Automatic Speech Recognition - ASR*). Cada vez más, en este ámbito se tiende al aumento de la complejidad de los modelos, con el propósito de mejorar la precisión para distintas condiciones acústicas y vocabularios extensos.

Al tratarse de habla continua, la definición de un modelo estadístico de lenguaje desempeña un papel fundamental. Estos pueden adaptarse a la temática de los contenidos para obtener un mejor resultado en el reconocimiento [1].

Por ello, la motivación de este proyecto será principalmente estudiar el efecto de aplicación del modelado del lenguaje adaptado a tópicos determinados y su influencia en el reconocimiento de contenidos audiovisuales de Internet.

1.2. Objetivos

El objetivo fundamental de este proyecto es el desarrollo de un sistema de reconocimiento de habla natural en español, cuya evaluación se llevará a cabo con contenidos multimedia de Internet. Para alcanzar la meta establecida se fijarán una serie de ob-

jetivos parciales, los cuales se enunciarán a continuación:

En los comienzos, se hará uso de la base de datos disponible, tras su adaptación pertinente, con el fin de generar los modelos básicos que constituyan el sistema de referencia basado en Modelos Ocultos de Markov (*Hidden Markov Models - HMM*), gracias a las herramientas de HTK (HMM ToolKit) [2] para el proceso de entrenamiento, y a Julius [3] para la etapa de reconocimiento.

Seguidamente, se estudiarán cuales son los tamaños de corpus y de vocabulario óptimos para el entrenamiento de los modelos estadísticos del lenguaje. Y a partir de ellos, las mejoras obtenidas al adaptar los modelos al tópico o tarea en cuestión, respecto al empleo de un modelo del lenguaje de vocabulario genérico.

El banco de pruebas principal, fundamentado en la motivación del proyecto, hará uso de contenidos multimedia de Internet. Finalmente, evaluando los resultados y como muestra de su utilidad, se implementará alguna aplicación que emplee este sistema de reconocimiento generado.

1.3. Estructura de la memoria

Esta memoria de proyecto esta dividida en los siguientes capítulos:

- Capítulo 1. Introducción: motivación y objetivos del proyecto.
- Capítulo 2. Estado del Arte: sistemas de reconocimiento de voz, arquitectura, reconocimiento con HMMs.
- Capítulo 3. Base de Datos: descripción y caracterización de la base de datos.
- Capítulo 4. Generación del Sistema de Referencia: metodología de entrenamiento, reconocimiento y resultados.
- Capítulo 5. Optimización del Sistema y Resultados: mejora del sistema con modelos de lenguaje adaptados al tópico y resultados.
- Capítulo 6. Aplicación del Sistema: demostración de la aplicación del sistema implementado a una solución comercial.
- Capítulo 7. Conclusiones y Trabajo Futuro.
- Referencias y anexos.

Capítulo 2

Estado del Arte

En este capítulo se proporcionará una visión general de la evolución de los trabajos realizados en el área del reconocimiento de voz hasta nuestros días (sección 2.1). En las siguientes secciones se explicarán la arquitectura de este tipo de sistemas de reconocimiento y sus bloques (sección 2.2), así como una visión detallada de la generación de modelos estadísticos (sección 2.3).

2.1. Introducción general del reconocimiento de voz

El reconocimiento de habla natural ha experimentado un intenso y gran desarrollo gracias a los avances que han tenido lugar en el procesamiento de señal, algoritmos, arquitecturas y plataformas de computación.

Desde 1940, los laboratorios de AT&T y Bell se encargaron de desarrollar un dispositivo rudimentario para reconocer voz, fundamentándose en los principios de la fonética acústica, teniendo presente que el éxito de esta tecnología, dependería de su habilidad para percibir la información verbal compleja con alta precisión.

En la década de los 50, el sistema anterior conseguido permitía identificación de dígitos monolocutor, basada en medidas de resonancias espectrales del tracto vocal para cada dígito. Siguiendo esta línea, RCA Labs trabajó en el reconocimiento de 10 sílabas. Y es a finales de la década, cuando tanto la University College de Londres como el MIT Lincoln Lab, trataron de desarrollar un sistema de reconocimiento limitado de vocales y consonantes. Esta tarea parecía novedosa por el uso de información estadística y cuyo objetivo era una mejora del rendimiento en palabras de dos o más fonemas.

Fue por la década de los 60, cuando los sistemas electrónicos utilizados hasta el momento, sirven de pasarela a los sistemas con hardware específico, en los NEC Labs

de Japón. En esta etapa, cabe destacar tres proyectos notables en la investigación de esta disciplina:

- RCA Labs tenían como objetivo un desarrollo de soluciones realistas para los problemas en la falta de uniformidad de las escalas de tiempo en el habla. Para ello, diseñaron un conjunto de métodos de normalización en el dominio temporal, detectando fiablemente el inicio y fin de discurso.
- En la Unión Soviética, T. K. Vintsyuk, propone el empleo de métodos de programación dinámica para conseguir el alineamiento temporal de parejas de realizaciones. Surge de aquí la técnica *DTW* (*Dynamic Time Warping*).
- Por último, en el campo del reconocimiento de habla continua, D. R. Reddy de la Universidad de Stanford, desarrolla el seguimiento dinámico de fonemas, concluyendo su trabajo en un reconocedor de oraciones de amplio vocabulario.

Allá por los años 70, se originan críticas acerca de la viabilidad y utilidad del reconocimiento automático de habla. A pesar de esto, dicha disciplina se adentra en el mundo probabilístico, donde los principales campos de estudio son los siguientes: El reconocimiento de palabras aisladas estuvo fundamentado en el procedimiento de ajuste de patrones, programación dinámica, y más adelante, técnicas *LPC* (*Linear Predictive Coding*). Esta última se empleó exitosamente en la codificación y compresión de la voz, a través del uso de medidas de distancias sobre el conjunto de parámetros LPC. Los primeros intentos de reconocedores de habla continua y grandes vocabularios los llevaron a cabo IBM, con el dictado automático de voz, ARPA Speech Understanding Research, y la Universidad de Carnegie Mellon, con el exitoso sistema Hearsay I. Finalmente, en los AT&T Labs, se investigó en la dirección de los reconocedores independientes del locutor para aplicaciones telefónicas, finalizando este periodo con la implementación de sistemas *ASR* (*Automatic Speech Recognition*), favorecida por tarjetas microprocesador.

La década de los 80 se inicia con una base muy asentada en la construcción de sistemas de reconocimiento, a diferencia de los anteriores que sólo reconocía vocablos aislados, ahora tienen la capacidad de tratar con palabras encadenadas fluidamente. Uno de los avances más importante es el paso de métodos basados en comparación de plantillas a otros basados en modelos estadísticos, extendiéndose el uso de los Modelos Ocultos de Markov o HMMs. Estos experimentaron numerosas mejoras y se situaron como los mejores modelos que capturaban y modelaban la variabilidad del habla.

Las redes neuronales empezaron a tomar peso en este ámbito, y gracias al desarrollo de algoritmos de aprendizaje más eficaces, aparecieron modelos como el *perceptrón*.

Además se llevan a cabo una serie de avances:

- El diseño de unidades de decodificación fonética a partir de la experiencia de fonetistas en tareas de interpretación de espectogramas.
- La grabación de grandes bases de datos como TIMIT, que permite la comparación de resultados entre diferentes grupos de trabajo.
- El programa DARPA (Defence Advance Research Agency) contribuyó en Estados Unidos, al impulso del desarrollo de sistemas de reconocimiento para habla continua y vocabularios de gran tamaño con independencia del locutor.
- El desarrollo por parte de la CMU de su sistema SPHINX [4].

En los años 90, continuando con los objetivos ya propuestos anteriormente, se ampliaban los tamaños de vocabularios y se diversifican los campos de aplicación. Teniendo gran importancia su aplicación sobre línea telefónica, así como los resultados de este reconocimiento en entornos con condiciones adversas y ruido.

Los avances producidos en el ámbito de las tecnologías del habla cada día son más significativos. En el campo del reconocimiento automático de voz, los reconocedores actuales manejan cada vez vocabularios más grandes y reducen las tasas de error gracias al uso de algoritmos más eficientes, al uso de equipos más potentes y al aumento de complejidad de estos sistemas, con modelados más sofisticados. El amplio grado de aplicación en función de los usuarios y los distintos entornos, hacen que no haya un sistema de reconocimiento de voz universal y sea necesaria su adaptación a las condiciones de funcionamiento y al tipo de aplicación que se requiera.

2.2. Arquitectura de un reconocedor automático de habla

Los sistemas de reconocimiento automático de habla (RAH) han sido abordados desde diferentes enfoques como se ha comentado anteriormente (sección 2.1), siendo los probabilísticos, que emplean la “Teoría de la Decisión de Bayes”, la “Teoría de la Información” y las “Técnicas de Comparación de Patrones”, los que han aportado los mejores resultados.

Un sistema de estas características (Figura 2.1), tiene la finalidad de extraer de la información acústica contenida en la señal de voz, una representación de todo el conjunto de sonidos pronunciados en formato texto. Para llevar a cabo esta decodificación, existen diferentes técnicas partiendo de un conjunto de patrones que sean comparables con el mensaje de entrada, y devolviendo al final una secuencia de aquellos patrones que con mayor probabilidad representan dicho mensaje.

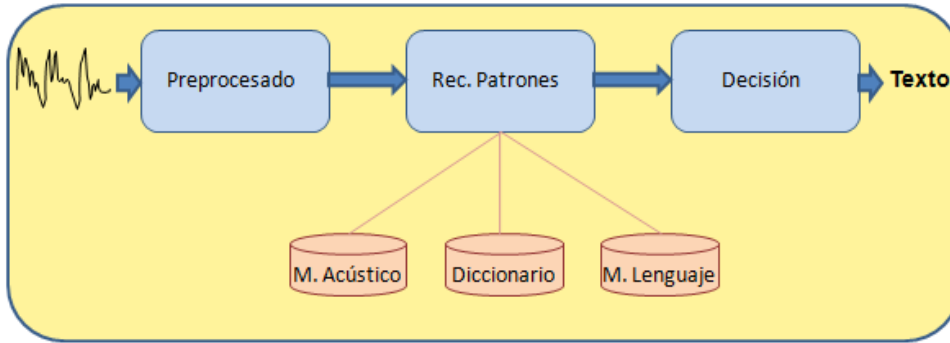


Figura 2.1: Esquema básico de la arquitectura de un RAH.

Una vez dada una visión general del propósito y funcionalidad de este tipo de sistemas de reconocimiento, se centrará la atención de este apartado en los sistemas de reconocimiento de habla continua y de vocabulario extenso (*Large Vocabulary Continuous Speech Recognition - LVCSR*), cuya arquitectura esta compuesta por los siguientes bloques:

2.2.1. Preprocesado

Este bloque engloba la extracción de características y transformación, procesamiento de características de robustez al ruido y la estimación de características adaptativas y discriminativas [5].

Extracción de características: El papel que desempeña el módulo de preprocesado es extraer a partir de la señal de voz una secuencia de vectores de características acústicas. Esto es realizado gracias a la transformada rápida de Fourier (*Fast Fourier Transform - FFT*) de la señal de voz dentro de una ventana de análisis, la cual se desplaza un intervalo de tiempo fijo. Las energías de las frecuencias vecinas dentro de cada trama son desechadas mediante un banco de filtros en la escala Mel, siendo las características de estos inspiradas en el proceso auditivo humano. A la salida de los filtros se aplica un logaritmo y los coeficientes son decorrelados a partir de la transformada discreta del coseno, dando lugar a un vector de coeficientes cepstrales de frecuencia Mel (*Mel Frequency Cepstral Coeficiente - MFCC*).

Posteriormente, estos coeficientes han sido reemplazados teniendo presente una mayor robustez al ruido, basado en coeficientes perceptuales de predicción lineal (*Perceptual Linear Prediction - PLP*).

En este contexto, la extracción de características ha beneficiado dos importantes técnicas: la primera de ellas, el uso de la media basada en el locutor, y la normalización de la varianza de los coeficientes cepstrales. Mientras que la sustracción de la

media cepstral basada en la pronunciación (*Cepstral Mean Subtraction - CMS*) es una técnica muy conocida, la normalización de la varianza cepstral (*Cepstral Variance Normalization - CVN*) se ha introducido recientemente. La segunda es la incorporación de contexto temporal entre las tramas, computando los coeficientes dinámicos o de velocidad y aceleración, también llamados coeficientes delta o delta-delta respectivamente). Estos son calculados a partir de las tramas próximas dentro de una ventana de aproximadamente unas 4 tramas de media. Estos coeficientes dinámicos se añaden a los estáticos formando así un vector final.

Características robustas al ruido: El ruido ambiente suele contaminar la señal de voz que obviamente afectará posteriormente al reconocimiento, de ahí que se trate este efecto en la etapa de preprocesado. El algoritmo SPLICE (“Stereo-based piecewise linear compensation for environments”) fue propuesto para entornos de ruido no estacionario, consistente en la eliminación del ruido por medio de la diferencia entre voz limpia y voz corrupta, asociada a la región más probable del espacio acústico. Otro algoritmo *QE* (*Quantile-based histogram equalization*) fue desarrollado para compensar distribuciones desalineadas de los datos de entrenamiento y de test.

Ambos fueron evaluados empleando un corpus de The Wall Street Journal, modificando el tipo y los niveles de ruido, pudiendo comprobarse mejoras en entornos experimentales limpios y multicondición.

Estimación de características adaptativas y discriminativas: La variación de las características acústicas puede ser observada entre los diferentes locutores o en un mismo locutor. Por ello, existen técnicas para generar un espacio de características canónicas, eliminando esta variabilidad mencionada en la medida de lo posible. Algunos ejemplos de ellas: normalización de la longitud del tracto vocal (*Vocal Tract Length Normalization - VTLN*), transformación de las características maximizando la verosimilitud bajo un modelo actual (*feature-space Maximum Likelihood Linear Regression - fMLLR*), transformación no lineal de la distribución de los datos de adaptación que será alineada con una distribución normal de referencia.

En la estimación de características discriminativas se usan técnicas como la transformación que permite obtener *offsets* dependientes del tiempo, a partir de una proyección lineal de un espacio de gaussianas posteriores de gran dimensión (*feature-space minimum phone error - fMPE*).

2.2.2. Reconocimiento de patrones

El principal motivo de emplear esta técnica es la consistencia de las representaciones de los patrones al definirse claramente un modelo matemático. Estas pueden servir de referencia a la hora de realizar comparaciones con alto grado de confianza;

para ello, serán precisos un conjunto de muestras etiquetadas y una metodología de entrenamiento [6].

2.2.2.1. Etapas

La representación de los patrones puede ser una plantilla (*template*) o un modelo estadístico (HMM), y se aplicará a un sonido, una palabra o una frase. Esta técnica puede dividirse en dos etapas: entrenamiento y comparación.

- **Entrenamiento:** Esta etapa consiste en la construcción de un patrón de referencia asociado a cada palabra o sub-unidad de palabra que se quiere reconocer, basándose en los vectores de características de aquellas unidades empleadas en el proceso de entrenamiento. Existen varias formas de llevar a cabo este proceso:
 - ✧ Entrenamiento casual: Se asigna un único patrón de sonido en la generación del patrón de referencia o un modelo estadístico aproximado.
 - ✧ Entrenamiento robusto: Se emplean varias versiones de cada unidad a reconocer (provenientes de un mismo locutor) generando así un patrón de referencia promedio o modelo estadístico promedio.
 - ✧ Entrenamiento por *clustering*: Se emplean gran volumen de datos, disponiendo de varias versiones de cada unidad (procedentes de un gran número de locutores) y así construir patrones de referencia o modelos estadísticos con alto grado de confianza.

- **Comparación:** Esta etapa está fundamentada, como su propio nombre indica, en la comparación directa entre el vector característico asociado a la señal de voz (a reconocer) y todos los posibles patrones entrenados, con el fin de determinar el mejor ajuste de acuerdo a un criterio establecido. Se definirá una medida de similaridad (distancia) entre vectores característicos a partir de la cual obtener el patrón de referencia mejor ajustado a la señal a reconocer.

2.2.2.2. Modelo acústico

Uno de los elementos fundamentales en la técnica de reconocimiento de patrones es el modelo acústico, que es un conjunto de representaciones estadísticas de los diferentes sonidos del espacio acústico con el que se está trabajando. Su elaboración se lleva a cabo a partir de un volumen de datos de entrenamiento, consistentes en datos de voz con su correspondiente etiquetado (transcripciones), haciendo posible una asignación de cada sonido a su representación o carácter gráfico.

A continuación, se mencionarán brevemente las técnicas más importantes empleadas para generar estos modelos:

Hidden Markov Model (HMM)

Los HMMs son modelos estadísticos empleados en la representación de secuencias de datos espaciales o temporales, este último es el caso de la señal de voz. Estos modelos son la base tecnológica de los sistemas de reconocimiento de voz, sustituyendo desde los años 80 a las técnicas de comparación de patrones como los DTW, que modelaban la voz de forma determinista.

Se considera que el sistema a modelar es un proceso de Markov de parámetros desconocidos, los cuáles serán calculados a partir de los parámetros observables. Este procedimiento ha sido el empleado en este proyecto, por ello, se dedicará una explicación más amplia en la sección 2.3.

Dinamic Temporal Warping (DTW)

Consiste en el alineamiento temporal de los parámetros de la locución de test y los parámetros del patrón, como resultado se obtiene la función de menor coste, que alinea ambas locuciones. El amplio abanico existente entre todos los posibles caminos de alineamiento, se verá reducido por un conjunto de límites locales y una serie de limitaciones.

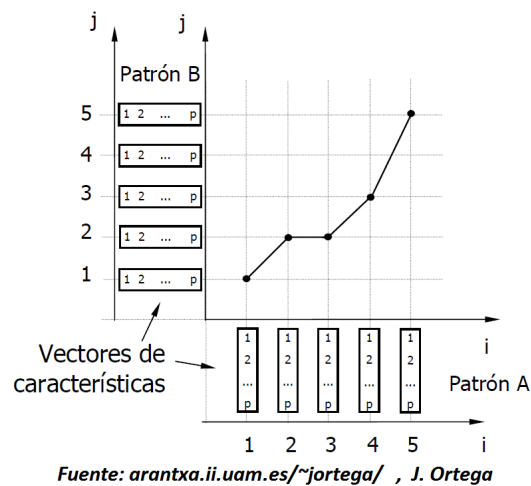


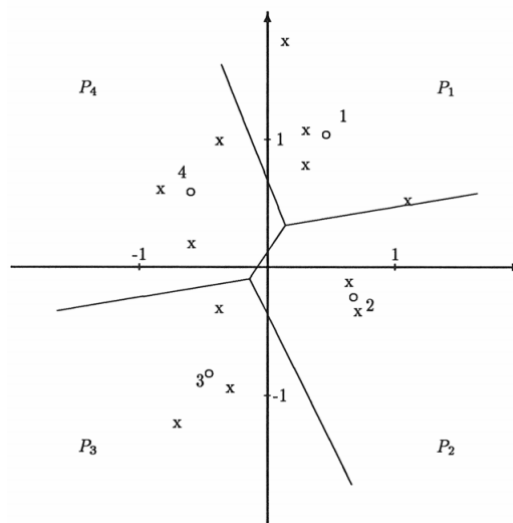
Figura 2.2: Función de alineamiento de DTW.

Vectorial Quantization (VQ)

Consiste en la representación de las características de las unidades como un espacio

vectorial, el cual cuenta con un conjunto infinito de patrones posibles (espacio de características). En este espacio se pretende asignar un conjunto de patrones desconocidos (test) a un conjunto finito de patrones de referencia; de manera que al vector a reconocer se le asigna un vector patrón cuya distancia a él sea mínima.

El espacio representativo quedará dividido en zonas o regiones, donde al vector representativo de esa región se denominará “*codeword*” (centroide), de forma que los vectores que caigan en dicha región se asignarán a dicho centroide. El conjunto de todos los centroides se denomina “*codebook*” (muestuario).



Fuente: arantxa.ii.uam.es/~jortega/ , J. Ortega

Figura 2.3: Espacio bidimensional dividido en regiones (VQ).

Como se puede observar en la Figura 2.3, correspondiente a un espacio bidimensional, se asignan aleatoriamente los centroides, representados con un 'o'; a su vez, los vectores de test son representados con una 'x' y serán asignados al centroide más cercano, mientras que cada una de las regiones podrían corresponderse con cada uno de los fonemas.

Deep Neural Networks (DNN)

Las redes neuronales profundas [7] son una forma alternativa de aprendizaje y de procesamiento automático, basado en el funcionamiento del sistema nervioso. Emplea una red neuronal basada en feed-forward que toma varias tramas de coeficientes como entrada y produce probabilidades a posteriori sobre los estados de HMM como salida. Se caracterizan por poseer un gran número de capas ocultas y son entrenadas usando nuevos métodos que mejoran otros procedimientos aquí ya mencionados.

2.2. ARQUITECTURA DE UN RECONOCEDOR AUTOMÁTICO DE HABLA 11

La conectividad completa entre capas adyacentes, se trata con la asignación de pesos iniciales de baja magnitud y aleatorios, y así se evita que todas las unidades ocultas en una capa tengan exactamente los mismos valores en cuanto a los pesos. En DNNs con gran número de capas y de elementos en cada capa, son modelos más flexibles que son capaces de modelar relaciones altamente no lineales entre entradas y salidas.

2.2.2.3. Diccionario

El diccionario llamado también lexicon, juega el papel de nexo entre la representación del nivel acústico y la secuencia de palabras a la salida del reconocedor. Consiste en un bloque que especifica tanto las palabras conocidas por el sistema, como los significados que construyen los modelos acústicos para cada entrada. Para LVCSR, normalmente el vocabulario es elegido con el objetivo de maximizar la cobertura para un tamaño de diccionario dado, pudiendo contener palabras iguales con más de una pronunciación. Además, la generación del diccionario se puede ver influida por aspectos como el tipo de habla, leída o espontánea [8], siendo recomendable tratar esta variabilidad de las pronunciaciones, y así obtener el mayor rendimiento posible al sistema.

2.2.2.4. Modelo del lenguaje

El modelo del lenguaje permite definir una estructura del lenguaje, es decir, restringir correctamente las secuencias de las unidades lingüísticas más probables. Son empleados en sistemas que hagan uso de una sintaxis y semántica compleja. Su funcionalidad debería consistir en aceptar (con alta probabilidad) frases correctas y rechazar (o asignar baja probabilidad) secuencias de palabras incorrectas.

Se pueden tener dos tipos de modelos: gramática cerrada de estados finitos y modelo de N-gramas.

Gramática cerrada de estados finitos

Este tipo de modelo representa restricciones del lenguaje de manera natural, permitiendo modelar dependencias tan largas como se quiera. Su aplicación conlleva una gran dificultad para tareas que hagan uso de lenguajes próximos a lenguajes naturales.

Partiendo de vocabulario finito compuesto por símbolos básicos (fonemas, letras, palabras, etc.) y una serie de reglas que restrinjan la construcción de una cadena o *string*, mediante la concatenación de elementos del vocabulario. La cadena o *string* será lo que se conoce como oración y los elementos del vocabulario serán palabras; todo esto se ilustrará con unos ejemplos a continuación:

- En el caso que se presenta en las Figuras 2.4 y 2.5, la tarea a abordar es muy apropiada para el uso de este tipo de gramáticas: sistema de marcación rápida por voz. La estructura de cada oración estará compuesta por un verbo que implica la acción de llamar/telefonar y un nombre de un contacto, o la acción de marcar más una secuencia de números. En este caso, todos los estados son equiprobables a partir de la acción pronunciada (marcar o llamar) y cuenta con estados de inicio y fin.

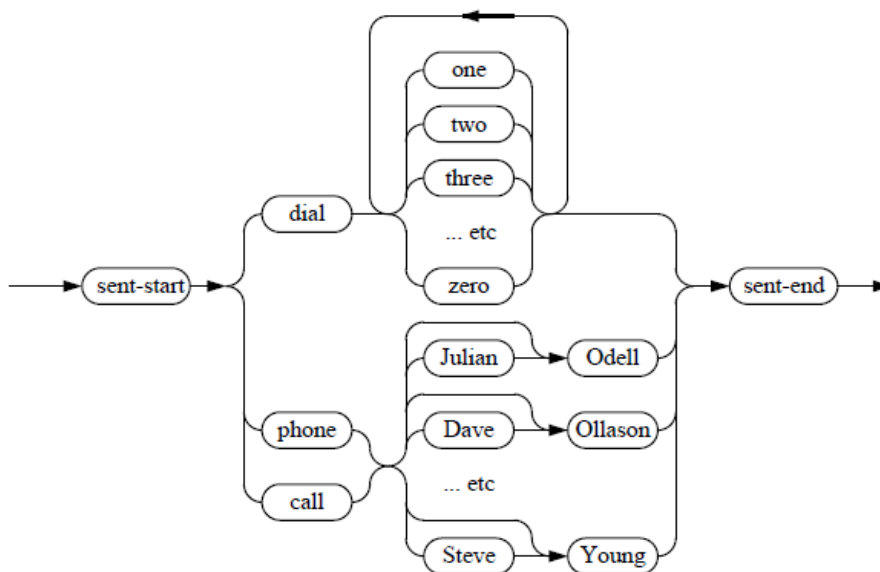
```

$digit = ONE | TWO | THREE | FOUR | FIVE |
        SIX | SEVEN | EIGHT | NINE | OH | ZERO;
$name   = [ JOOP ] JANSEN |
          [ JULIAN ] ODELL |
          [ DAVE ] OLLASON |
          [ PHIL ] WOODLAND |
          [ STEVE ] YOUNG;
( SENT-START ( DIAL <$digit> | (PHONE|CALL) $name) SENT-END )

```

Fuente: HTK Book 3.4

Figura 2.4: Definición de gramática.



Fuente: HTK Book 3.4

Figura 2.5: Diagrama de estados de la gramática.

- Por otro lado, el empleo de una de estas gramáticas en la tarea que acomete, de reconocer habla natural, supone un gran esfuerzo sin conseguir grandes resultados. La mecánica sería similar; si se considerase una estructura oracional

básica con sujeto (conjunto de nombres propios y comunes), verbo (todas las posibles acciones) y complementos, sería una buena aproximación, salvo por la excepción del idioma español que permite varios posibles órdenes sintácticos e incluso la omisión de algunas de sus partes. Además el diagrama de estados resultante sería de complejidad muy alta, con todas las posibles palabras del idioma y las distintas formas verbales. Por este motivo, para proceder al reconocimiento de habla natural, son empleados modelos de lenguaje estadísticos de N-gramas, obteniendo grandes resultados.

Modelo de N-gramas

Los modelos estadísticos de lenguaje de N-gramas [9], pretenden predecir la palabra siguiente de manera que se reduzca el espacio de búsqueda sólo a los candidatos más probables. El empleo de la información contextual, permite mejorar aplicaciones ahorrando medios. En idiomas con palabras que tienen la misma pronunciación u homónimas, se precisa claramente de la información contextual, y así mejorar la precisión del sistema.

Entrando en el fundamento matemático, este tipo de modelo corresponde a una cadena de Markov de orden $N - 1$. La probabilidad $P(w)$ de una secuencia de palabras $w = w_1, w_2, \dots, w_T$ de longitud T , es en primer lugar descompuesta en un producto de probabilidades condicionales, según la regla de Bayes:

$$P(w) = P(w_1) P(w_2 | w_1) \cdots P(w_T | w_1, \dots, w_{T-1}) = \prod_{t=1}^T P(w_t | w_1, \dots, w_{t-1}) \quad (2.1)$$

Un incremento de la longitud de la secuencia de palabras, producirá que la factorización anterior requiera de probabilidades con largas dependencias arbitrarias; por lo tanto, para aplicaciones prácticas la máxima longitud del contexto se limitará a $N - 1$ elementos predecesores. Por ello, la principal dificultad de estimar este modelo y aplicarlo, es la cantidad de recursos requeridos con el aumento del tamaño del contexto. Así pues, las variantes de modelos de N-gramas más utilizadas son bigramas y trigramas, mientras que modelos de 4-gramas apenas se emplean en sistemas de reconocimiento.

Una cuestión importante es la capacidad de estos modelos para describir bien los datos disponibles, mediante la probabilidad de la secuencia de palabras o las medidas teóricas de la información derivada de ella. La calidad del modelo puede evaluarse con la fiabilidad para describir los datos, por ejemplo la asignación de párrafos de texto a ciertas categorías o tópicos específicos, con los efectos que esto supone en

el reconocimiento de voz como se explicó en la sección 1.1, siendo esta una de las motivaciones de este proyecto.

El proceso de generación cuenta las frecuencias absolutas $c(w_1, w_2, \dots, w_N)$ de todos los conjuntos de símbolos, de todos los posibles contextos w_1, w_2, \dots, w_{N-1} . Las probabilidades condicionales $P(w_N | w_1, w_2, \dots, w_{N-1})$ pueden ser definidas por las frecuencias relativas $f(w_N | w_1, w_2, \dots, w_{N-1})$, dando lugar a la ecuación 2.2

$$P(w_N | w_1, w_2, \dots, w_{N-1}) := f(w_N | w_1, w_2, \dots, w_{N-1}) = \frac{c(w_1, w_2, \dots, w_N)}{c(w_1, \dots, w_{N-1})} \quad (2.2)$$

Los N-gramas que no han sido observados se denominan eventos ocultos, todas las probabilidades condicionales se definirán como cero, significando que para cada secuencia existirá un único evento oculto, esto no es deseable por la falta de fiabilidad al tener un conjunto limitado de datos. Entonces, el modelo estará sujeto a un post-procesado, desde técnicas de *smoothing* o suavizado, asignando parte de la probabilidad total a las palabras o N-gramas ocultos, hasta modelos más sofisticados como técnicas de descuento como *Good-Turing*, o los modelos de *back-off*.

A continuación, en la Figura 2.6 se muestra un ejemplo del cálculo de probabilidades en un modelo de lenguaje de bigramas. Será necesario realizar una serie de aclaraciones:

- Los elementos *!ENTER* y *!EXIT* actúan como marcadores de inicio y fin de la oración.
- En un modelo de trigramas, las probabilidades condicionales de cada elemento, en lugar de tener una dependencia del unigrama anterior, dependerá del bigrama que le precede. Tomando de partida el ejemplo de la figura, la probabilidad condicional de *black*, dependería del bigrama *“saw a”*, siendo más determinantes al emplear probabilidades de subconjuntos de palabras de mayor tamaño.

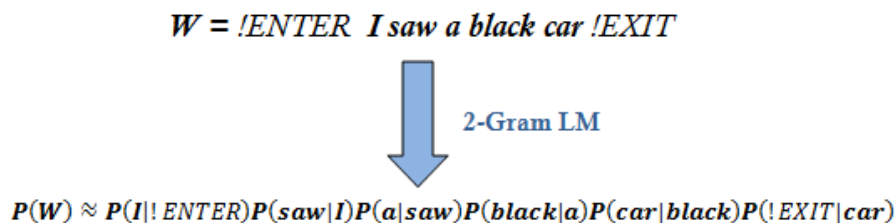


Figura 2.6: Cálculo de probabilidades de un modelo del lenguaje (bigramas).

2.2.3. Decisión

Esta última etapa consiste en la toma de decisión a la hora de asignar un patrón de los que se han generado en la fase de entrenamiento del sistema. Para ello, se hará uso de las medidas realizadas en la fase de comparación; es decir, los cálculos de parecido o similitud entre la realización acústica de entrada y el conjunto de modelos conocidos por el sistema. A partir de los valores de similitud obtenidos, el reconocedor debe tomar una decisión acerca de los sonidos que ha generado la señal de voz de entrada.

El teorema de decisión de Bayes expresa la probabilidad condicional de que los sonidos de entrada o combinaciones de ellos (trifonemas) pudiesen ser generados por alguno de los estados que modelan todas las posibles unidades del espacio acústico. Por otro lado, las distribuciones probabilísticas para el modelado del lenguaje, reflejan la frecuencia de aparición de las cadenas de palabras. La decisión estará basada en la mayor verosimilitud obtenida tanto del modelo acústico como del modelo del lenguaje.

Este bloque es uno de los que más relevancia tiene para el diseñador de la arquitectura del sistema de reconocimiento, ya que es la única salida observable por el usuario.

2.3. Reconocimiento con HMMs

Como se explicó brevemente en la sección 2.2, los HMMs se han convertido en la aproximación predominante en el reconocimiento de habla por su algorítmica y los resultados que se obtienen con ellos.

2.3.1. Definición y caracterización

Un modelo oculto de Markov es la representación de un proceso estocástico que se compone de dos elementos: una cadena de Markov de primer orden, con un número finito de estados, y un conjunto de funciones aleatorias asociadas a cada uno de los estados. En un instante concreto de tiempo el proceso está en un estado determinado y genera una observación mediante la función aleatoria asociada. En el instante siguiente, la cadena de Markov cambia de estado o permanece en el mismo siguiendo su matriz de probabilidades de transición entre estados, generando una nueva observación mediante la función aleatoria correspondiente. El observador externo solamente verá la salida de las funciones asociadas a cada estado, sin observar directamente la secuencia de estados de la cadena de Markov.

Tomando por mayor simplicidad la notación de modelos discretos, un HMM está caracterizado por:

- El número de estados en el modelo, N . Se denota cada estado como S_i , un estado en el instante t como q_t ; por lo tanto si el sistema se encuentra en el estado S_i en el instante t , $q_t = S_i$.
- El número de símbolos observables, M . Se denota a cada símbolo observable como v_j , la observación en el instante t como O_t , y si la observación en el instante t es v_j , se tomará $O_t = v_j$.
- La matriz de probabilidades de transición se define como $A = \{a_{i,j}\}$, siendo $a_{i,j} = P[q_{t+1} = S_j | q_t = S_i]$ y cumpliéndose que $1 \leq i, j \leq N$
- La distribución de probabilidad de observación en cada estado j como $B = \{b_j(k)\}$, siendo $b_j(k) = P[v_k(t) | q_t = S_j]$ y para $1 \leq j \leq N$ y $1 \leq k \leq M$.
- La probabilidad inicial de ocupación de cada estado como $\pi = \{\pi_i\}$, donde $\pi_i = P[q_1 = S_i]$ y para $1 \leq i \leq N$.

Por convenio, un modelo HMM con todos los parámetros será denotado de la siguiente forma: $\lambda = (A, B, \pi)$.

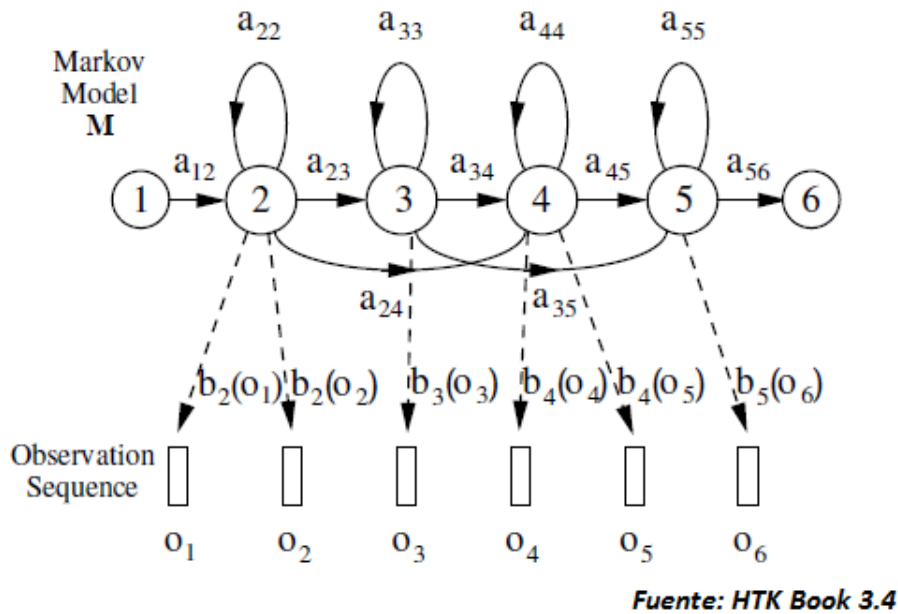


Figura 2.7: Representación gráfica de un HMM de 6 estados.

Dependiendo de la naturaleza de la matriz de distribución de probabilidades de salida B , los HMMs se pueden clasificar en varios tipos: modelos discretos, modelos continuos y modelos semicontinuos.

- **Modelos discretos:** En este tipo de modelos, las observaciones son vectores compuestos por símbolos de un alfabeto finito de N elementos distintos. Para cada elemento del vector de símbolos se define una densidad discreta y la probabilidad del vector se calcula multiplicando las probabilidades de cada componente siendo éstos independientes entre sí.

- **Modelos continuos:** Definen las distribuciones de probabilidad en espacios de observaciones continuos, muy conveniente en el ámbito de este proyecto ya que se trata de la señal de voz que es propiamente continua. Se suelen restringir el número de parámetros del sistema para conseguir una mayor manejabilidad de éste y consistencia de las re-estimaciones; para ello, se emplean mezclas de distribuciones paramétricas como gaussianas, para definir las transiciones. Cada estado x_i tendrá un conjunto específico $V(x_i, \lambda)$ de funciones densidad de probabilidad. Llamando v_k a cada de las funciones de densidad de probabilidad, las probabilidades de las salidas se pueden expresar como:

$$b_i(y) = p(y | x_i, \lambda) = \sum_{v_k \in V(x_i, \lambda)} p(y | v_k, x_i, \lambda) P(v_k | x_i, \lambda) \quad (2.3)$$

- **Modelos semicontinuos:** En ellos se modelan distribuciones complejas con un elevado número de mezclas de funciones paramétricas y un gran corpus de entrenamiento. Se compartirán las mismas distribuciones de probabilidad con pesos distintos, entre todos los estados del modelo.

Como se acaba de comentar, la dependencia temporal de la señal de voz permite que los HMMS se adapten muy bien en sistemas de reconocimiento, además del cálculo de probabilidades acústicas gracias a la capacidad de esta técnica a la hora de modelar estadísticamente la generación de voz. Para su uso se tendrán en cuenta dos hipótesis:

1. El análisis localizado de la señal de voz, permitirá la división de ésta en fragmentos, estados, en los que se puede considerar su pseudoestacionariedad [10]. Esto es gracias, a que en la ventana de análisis, la señal mantiene su periodicidad, teniendo presente las transiciones.

2. La hipótesis de independencia de Markov, que enuncia que la probabilidad de observación de que se genere un vector de características, depende únicamente del estado actual y no de elementos anteriores [11].

2.3.2. Los tres problemas básicos de los HMMs

Existen tres problemas fundamentales de los Modelos Ocultos de Markov, cuya solución hace de ellos, una técnica de la robustez y utilidad ya mencionada en aplicaciones reales:

- **Evaluación o de puntuación:** Dada una secuencia de observaciones acústicas y un modelo oculto de Markov, ¿cómo calcular la probabilidad de que dicho modelo genere la secuencia de observación vista? Esta probabilidad, $P(O|\lambda)$, se determina a partir del algoritmo de forward-backward [12].
- **Decodificación o reconocimiento de estados:** Dada una secuencia de observaciones acústicas y un modelo oculto de Markov, ¿cuál es la secuencia de estados óptima que explique dichas observaciones? La secuencia de estados óptima se consigue gracias a un alineamiento de la secuencia de observación con los estados, mediante el algoritmo de Viterbi [13].
- **Entrenamiento:** Dado un conjunto de observaciones de entrenamiento y un modelo oculto de Markov, ¿cómo se ajustan los parámetros del modelo para maximizar la probabilidad de observar el conjunto de entrenamiento? Este ajuste paramétrico será solucionado con el algoritmo de Baum-Welch [14].

2.3.2.1. Problema de evaluación. Algoritmo de Forward-Backward

Partiendo de un modelo HMM definido por $\lambda = (A, B, \pi)$ y la secuencia de observaciones $O = O_1 O_2 \cdots O_T$, suponiendo que la secuencia de estados es $Q = q_1 q_2 \cdots q_T$, la probabilidad de la secuencia de observaciones dada una secuencia de estados, viene dada por:

$$P(O|Q, \lambda) = \prod_{t=1}^T P(O_t|q_t, \lambda) \quad (2.4)$$

Con la asunción de la independencia estadística de las observaciones, tenemos:

$$P(O|Q, \lambda) = b_{q_1}(O_1) b_{q_2}(O_2) \cdots b_{q_T}(O_T) \quad (2.5)$$

Por otro lado, la probabilidad conjunta de O y de Q será:

$$P(O, Q|\lambda) = P(O|Q, \lambda) P(Q, \lambda) \quad (2.6)$$

Sabiendo que la probabilidad de la secuencia de estados Q , puede expresarse como el producto de la probabilidad del estado inicial y de las probabilidades de transición

entre estados :

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} \pi_{q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T} \quad (2.7)$$

La probabilidad buscada se calculará sumando las probabilidades anteriormente definidas para todos los caminos posibles o secuencias de estados:

$$P(O|\lambda) = \sum_{\forall Q} P(O|Q, \lambda) P(Q|\lambda) = \sum_{q_1, q_2, \dots, q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T) \quad (2.8)$$

De la expresión anterior se puede deducir que en el instante inicial $t = 1$, se tiene presencia en el estado q_1 con una probabilidad inicial π_{q_1} y se genera la observación O_1 con probabilidad $b_{q_1}(O_1)$. En un instante de tiempo siguiente $t = 2$, la transición del estado q_1 a q_2 se producirá con una probabilidad $a_{q_1 q_2}$ y se generará la observación O_2 con probabilidad $b_{q_2}(O_2)$. Este proceso se realizará de la misma forma hasta la última transición, es decir hasta el estado final q_T .

El problema de este cálculo directo es que requiere un número muy elevado de operaciones, del $O(2TN^T)$, siendo inviable. Esto se soluciona con el algoritmo de forward-backward que realiza cálculos intermedios que emplea a posteriori y que supone una reducción del coste computacional del $O(TN^2)$.

Se llevará a cabo el siguiente procedimiento:

- **Inicialización** de la variable forward, que representa la probabilidad de observar la secuencia parcial hasta el instante t y estar en el estado S_i en dicho instante (ecuación 2.9); y que en el instante $t = 1$, será como muestra la ecuación 2.10.

$$\alpha_t(i) = P(O_1 O_2 \cdots O_t, q_t = S_i | \lambda) \quad (2.9)$$

$$\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N \quad (2.10)$$

- A través del **método inductivo**, se calculan las variables forward en el instante $t + 1$ a partir de las variables forward en el instante t , de las probabilidades de transición y probabilidades de observación. Se realizarán un total de $N + 1$ productos y $N - 1$ sumas.

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(O_{t+1}) \quad (2.11)$$

- El último paso es de **finalización**. La probabilidad deseada se calcula como

suma de las probabilidades hacia delante en el último instante posible, T :

$$P(O|\lambda) = \sum_{i=1}^N \alpha_T(i) \quad (2.12)$$

Otro algoritmo es el de backward, que tiene el mismo fundamento pero la probabilidad de observación de una secuencia i y se modela como:

$$\beta_t(i) = P(O_{t+1}O_{t+2}\cdots O_T | q_t = S_i, \lambda) \quad (2.13)$$

donde $\beta_t(i)$ representa la probabilidad de observar la secuencia parcial $O_{t+1}\cdots O_T$ desde el instante $t+1$ y estar en el estado S_i en el instante t . Se puede calcular la probabilidad empleando tanto el método de forward como de backward, o ambos a la vez que implican una resolución fácil del problema.

Se llevará a cabo el siguiente procedimiento:

- **Inicialización** de la variable backward, teniendo presente que todos los estados son equiprobables, se obtiene la ecuación 2.14; y para el instante $t = T$, quedará como se muestra a continuación (ecuación 2.15).

$$\beta_t(i) = \frac{1}{N} \quad (2.14)$$

$$\beta_T(i) = 1, \quad 1 \leq i \leq N \quad (2.15)$$

- A través del **método inductivo**, se calculan las variables backward, de derecha a izquierda recursivamente, donde se tiene el mismo coste computacional que en el caso anterior.

$$\beta_t(i) = \sum_{j=1}^N \alpha_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \quad (2.16)$$

2.3.2.2. Problema de decodificación. Algoritmo de Viterbi

Existe la necesidad de encontrar la secuencia de estados que explique la secuencia de observaciones dada. Este proceso de decodificación, puede solucionarse de acuerdo a varios criterios.

El primero de ellos puede ser la elección del estado más probable, en cada instante de tiempo. Para ello se tendría que maximizar en cada instante de tiempo la siguiente variable, la cual representa la probabilidad de ocupación de cada estado en el instante t :

$$\gamma_t(i) = P(q_t = S_i | O, \lambda) \quad (2.17)$$

Una forma probable sería calcular las variables forward y backward para calcular la variable anterior:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^N \alpha_t(i)\beta_t(i)} \quad (2.18)$$

Y por último se tomaría el estado más probable:

$$q_t = \arg \max_{1 \leq i \leq N} [\gamma_t(i)], \quad 1 \leq i \leq N \quad (2.19)$$

Este criterio no parece ser el más adecuado debido a que no tiene presente la probabilidad de ocurrencia de la secuencia de estados y, por ejemplo, al tener una probabilidad de transición entre estados nula, $a_{ij} = 0$, podría dar como resultado una secuencia de estados (secuencia de fonemas) que no tuviese sentido.

El segundo criterio, a su vez el más adecuado, consiste en elegir el camino completo de estados con mayor probabilidad global (secuencia de fonemas válida). Se resuelve utilizando el algoritmo de Viterbi. Antes de explicar en qué consiste, se definirá la siguiente variable auxiliar:

$$\delta_t(i) = \max_{q_1 q_2 \dots q_{t-1}} P[q_1 q_2 \dots q_t = i, O_1 O_2 \dots O_t | \lambda] \quad (2.20)$$

Donde $\delta_t(i)$ representa la mejor puntuación (máxima probabilidad) obtenida a través de una secuencia única de estados ($q_1 q_2 \dots q_{t-1}$) hasta llegar en el instante t , al estado i . Una ventaja de este algoritmo es que si se conoce la anterior variable para todos los estados en el instante t , se pueden calcular, también para todos los estados, en el instante siguiente ($t + 1$):

$$\delta_{t+1}(i) = \left[\max_i \delta_t(i) a_{ij} \right] b_j(O_{t+1}) \quad (2.21)$$

Además de estas variables, se necesitará disponer también del estado i que maximiza el argumento de la ecuación arriba enunciada. Se llevará a cabo almacenando todos sus valores para cada instante t y cada estado j en otra variable $\varphi_t(j)$.

El proceso completo es el siguiente:

- **Inicialización** de la variable auxiliar en el instante inicial y de la variable de

almacenamiento de estados que proporcionan máximos.

$$\delta_1(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N \quad (2.22)$$

$$\varphi_1(i) = 0 \quad (2.23)$$

- Por **recursión**:

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), \quad 2 \leq t \leq T, 1 \leq j \leq N \quad (2.24)$$

Se guardan los valores máximos, que servirán posteriormente para obtener el camino óptimo:

$$\varphi_t(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}], \quad 2 \leq t \leq T, 1 \leq j \leq N \quad (2.25)$$

- La parte de **finalización**:

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)] \quad (2.26)$$

$$q_T^* = \arg \max_{1 \leq i \leq N} [\delta_T(i)] \quad (2.27)$$

- El **backtracking** consistirá en realizar el camino desde el instante final al inicial, adoptando aquellos valores que maximizan cada paso de la etapa de recursión, obteniendo así la secuencia de estados óptima:

$$q_t^* = \varphi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \dots, 1 \quad (2.28)$$

Según se muestra la imagen de la Figura 2.8, el algoritmo de Viterbi funciona similar al algoritmo forward empleado en la fase de evaluación, teniendo también el mismo coste computacional que éste, del $O(TN^2)$.

2.3.2.3. Problema de entrenamiento. Algoritmo de Baum-Welch

Para estimación de los parámetros del modelo $\lambda = (A, B, \pi)$ que maximizan la probabilidad de observación $P(O|\lambda)$, se utilizará el algoritmo de Baum-Welch, que consiste en un caso particular del algoritmo de Expectation-Maximization (EM) aplicado a los HMMs. Se pretende mediante una serie de iteraciones y bajo el criterio de máxima verosimilitud (Maximum Likelihood), ir encontrando máximos locales de la probabilidad de observación.

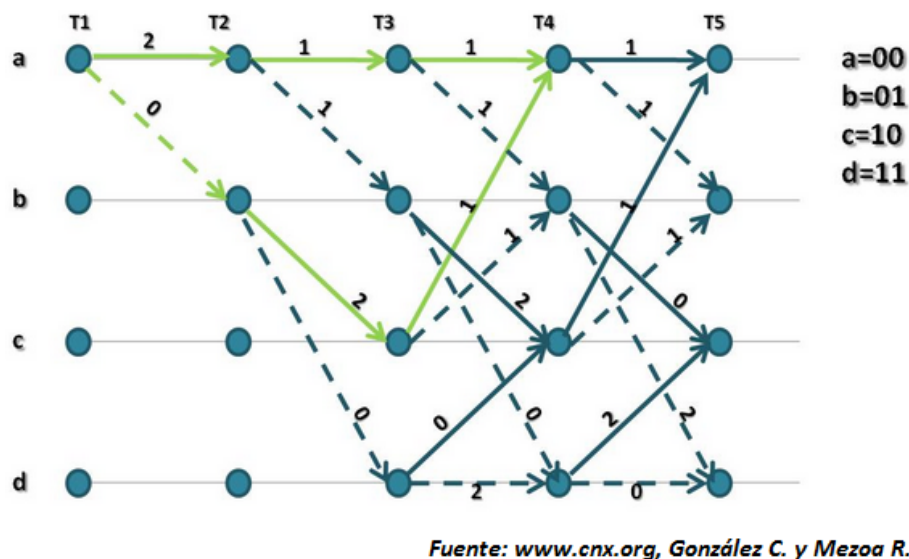


Figura 2.8: Esquema del algoritmo de Viterbi.

En primer lugar, se comenzará por definir una función auxiliar, que depende de los parámetros anteriores del modelo λ y de la nueva estimación de ellos $\bar{\lambda}$:

$$Q(\lambda | \bar{\lambda}) = \sum_Q P(Q | O, \lambda) \log [P(O, Q | \bar{\lambda})] \quad (2.29)$$

Según el algoritmo de EM, se garantiza que maximizando la función anterior respecto a los nuevos parámetros, se obtendrá una mayor verosimilitud en la siguiente iteración:

$$\max_{\bar{\lambda}} [Q(\lambda | \bar{\lambda})] \implies P(O | \bar{\lambda}) \geq P(O | \lambda) \quad (2.30)$$

Este proceso se repetirá para ir obteniendo nuevos parámetros en cada iteración, para seguir aumentando la verosimilitud hasta el punto en el que el algoritmo converja o el incremento de verosimilitud sea mínimo.

A continuación, se hará una distinción entre los dos pasos del algoritmo:

- En primer lugar, el **paso de Expectation**, tiene como misión calcular los elementos de la ecuación 2.29, que dependen del modelo anterior, principalmente el término $P(Q | O, \bar{\lambda})$, es decir las probabilidades de todas las secuencias de estados dados el modelo anterior y las observaciones.

Los parámetros que se van a estimar son:

- ✧ La probabilidad de estar en el estado S_i en el instante t , que vienen dada por la probabilidad de ocupación del estado anteriormente definido, que se podía calcular en función de las variables forward y backward (ecuaciones 2.20 y 2.21).
- ✧ El número esperado de transiciones desde el estado S_i , la cual puede obtenerse a partir de las variables anteriormente mencionadas, de la siguiente forma:

$$\sum_{t=1}^{T-1} \gamma_t(i) \quad (2.31)$$

- ✧ El número esperado de transiciones desde el estado S_i al estado S_j . Para ello, inicialmente se definirá la probabilidad de transición entre ambos estados, y posteriormente se obtendrá el número de transiciones sumando los valores obtenidos:

$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O | \lambda)} \quad (2.32)$$

$$\sum_{t=1}^{T-1} \xi_t(i, j) \quad (2.33)$$

- En segundo lugar, se tendrá el **paso de Maximization**, consistente en maximizar la función (ecuación 2.29) una vez ya calculados los parámetros en el paso anterior, y dará lugar a unos parámetros nuevos, siendo los siguientes:

$$\bar{\pi}_i = \text{frecuencia esperada en } S_i \text{ en el instante } (t = 1) = \gamma_1(i) \quad (2.34)$$

$$\bar{a}_{ij} = \frac{\text{número esperado de transiciones desde } S_i \text{ a } S_j}{\text{número esperado de transiciones desde } S_i} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (2.35)$$

$$\bar{b}_j(k) = \frac{\text{número esperado de veces en el estado } j \text{ observando } v_k}{\text{número esperado de veces en el estado } j} = \frac{\sum_{t=1, s.t. O_t=v_k}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \quad (2.36)$$

Los algoritmos descritos anteriormente, que solucionan los tres problemas enun-

ciados al principio de esta subsección, serán empleados en este proyecto, tanto en el proceso de entrenamiento como en el de evaluación del sistema de reconocimiento. La metodología seguida será descrita más adelante en la sección 4 y en ella serán empleadas herramientas de software que hacen uso de estos algoritmos.

Capítulo 3

Base de Datos

La base de datos de entrenamiento es uno de los elementos más importantes a la hora de generar un sistema de reconocimiento de habla. La calidad de la misma determina la viabilidad de un buen entrenamiento. En este caso se han utilizado diversas bases de datos cedidas por Telefónica I+D con las que se ha conformado la base de datos final de entrenamiento que se detalla a continuación.

3.1. Preparación de los datos

La base de datos consta de un listado de archivos de audio y de un listado de archivos de etiquetas con el mismo nombre pero distinta extensión. Dichos archivos de etiqueta contienen dos partes:

- La transcripción del archivo de audio que es imprescindible.
- Información adicional como puede ser: nombre de la base de datos, identificador de hablante, sexo, idioma, dialecto, SNR del archivo, *pitch* del hablante, fecha y lugar de grabación, eventos fonéticos (como clics, saturación, mala pronunciación...), tipo de ruido de fondo (si lo hubiera).

A partir de estas etiquetas generales se puede extraer el campo de transcripción y adaptarlo a lo que necesita cada programa de entrenamiento. Para el software HTK el archivo de etiquetas debe tener cada palabra de la transcripción en una línea distinta, y que la última línea sea un símbolo de punto.

Los archivos deben estar transcritos correctamente para evitar entrenar fonemas de forma errónea. Los archivos de audio deben ser fieles a dichas transcripciones, asegurándose de que no están vacíos ni intercambiados.

La riqueza de la base de datos se puede medir por diversos factores:

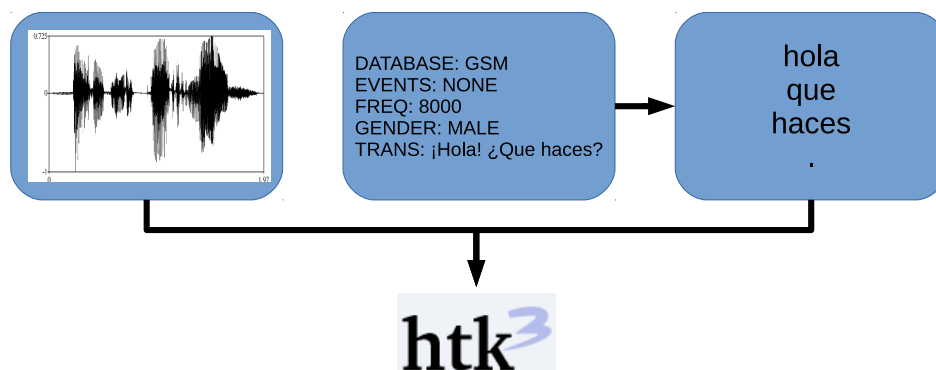


Figura 3.1: Archivos necesarios para HTK.

- El número de archivos con los que cuenta
- El número de frases y palabras distintas de entrenamiento
- El número de palabras y caracteres totales
- El promedio de caracteres por palabra
- El número de horas de audio y cuántas de ellas son de silencio y voz
- La cantidad de frases de hombres y mujeres
- El número de hablantes distintos
- La SNR promedio
- El *pitch* promedio

Para la generación de la base de datos de entrenamiento se ha partido de bases de datos más pequeñas que se han ido limpiando y uniendo, eliminando archivos con demasiado ruido, incompletos, SNR muy baja, etc. También se ha reducido la tasa binaria de los archivos de audio con velocidad superior para ajustarlos a la calidad telefónica.

Para el entorno de evaluación, de esta base de datos se han extraído 1000 archivos representativos del total de la misma para hacer pruebas de reconocimiento con archivos no entrenados con la SNR promedio de la base de datos. También se han extraído otros 1000 archivos con ruido para hacer pruebas con archivos no entrenados ruidosos. Cada conjunto suma 70 minutos de audio.

Las bases de datos de origen son de diversa índole, siendo aproximadamente el 60 % voz telefónica (tanto de GSM como línea fija) grabada en diversos ambientes (calle, bares, coches, hogares...) y un 40 % voz limpia grabada en estudio.

El idioma de la base de datos es el español de España, conteniendo muestras de casi todas las comunidades autónomas para los distintos acentos.

Así, la base de datos final está en formato WAV, a 8 kHz, 16 bits y 1 canal.

En la siguiente tabla se pueden ver el resto de parámetros de la base de datos utilizada para el entrenamiento:

Tabla 3.1: Características de la base de datos.

Características	Valor
Nº Archivos / Frases totales	244.132
Nº Frases distintas	19.997
Nº Palabras Totales	1.063.866
Nº Letras sin espacios	5.746.114
Promedio letras por palabra	5,40
Nº Palabras distintas	11.182
Nº de Horas de audio	275,64
Nº de horas sin silencios	124,95
Silencio (%)	54,67
Voz (%)	45,33
Frases hombres	141.391
Frases mujeres	102.700
Hombres (%)	57,92
Mujeres (%)	42,08
Hablantes distintos	32.309
Velocidad de muestreo (Hz)	8.000
Bits por muestra	16
Canales de audio	1
SNR Promedio	33,7
Pitch Promedio Hombres (Hz)	123
Pitch Promedio Mujeres (Hz)	202

Otros factores relevantes son:

- La distribución de SNRs, para caracterizar mejor la calidad de los archivos de la base de datos.
- La distribución del *pitch* de los hablantes, que nos da indicaciones sobre las frecuencias fundamentales de la voz presente en la base de datos. Es un factor más relevante que diferenciar solamente entre hombres y mujeres.

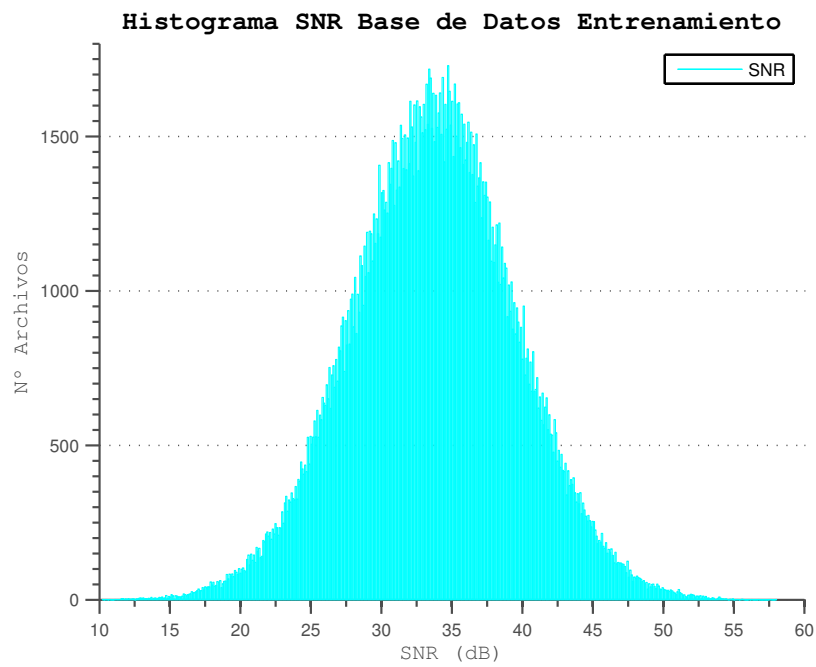


Figura 3.2: Histograma de la distribución de la SNR de la base de datos.

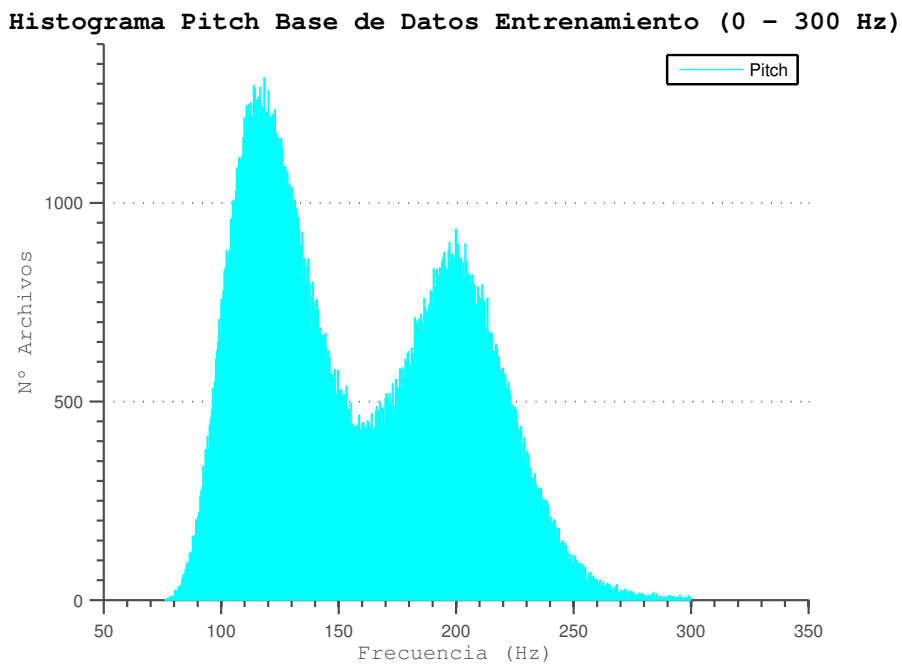


Figura 3.3: Histograma de la distribución del *pitch* promedio de la base de datos.

Capítulo 4

Generación del Sistema de Referencia

En este capítulo se detallan los pasos seguidos para la realización del sistema de referencia que se utilizará en las pruebas posteriores.

4.1. Introducción

El proyecto consta principalmente de tres partes:

- La primera es la configuración del sistema de entrenamiento y la definición de los procedimientos de entrenamiento que sirva para estimar los modelos acústicos y de lenguaje iniciales.
- La segunda es utilizar un sistema de reconocimiento con el fin de:
 - ✧ Calcular la tasa de aciertos de los archivos entrenados para verificar si el entrenamiento progresa adecuadamente y tener una referencia del máximo que se puede alcanzar con el sistema.
 - ✧ Reconocer el conjunto de los archivos de test para obtener tasas efectivas en un entorno no controlado.
 - ✧ Introducir cambios en el sistema de entrenamiento y verificar si las nuevas modificaciones introducidas provocan una mejora o empeoramiento en el sistema global.
 - ✧ Decidir a partir de los resultados experimentales cuál va a ser el sistema base de entrenamiento.

- Una vez obtenido un sistema base de entrenamiento fiable, efectuar los nuevos experimentos que tienen el objetivo de mejorar el sistema inicial.

Se ha generado desde cero el sistema de entrenamiento a utilizar partiendo solamente de las herramientas del software HTK. Su realización ha llevado en torno al 65 % de las horas dedicadas, sumando el tratamiento de la base de datos, generación de modelos de lenguaje iniciales y generación de modelos acústicos de referencia.

En las tareas de reconocimiento y pruebas efectuadas se han invertido el resto de horas del proyecto, junto con la redacción de este documento.

4.1.1. Software utilizado

- **Linux**

Se ha decidido efectuar este proyecto sobre un sistema operativo Ubuntu-Linux dada la facilidad de conseguir diversas librerías que son necesarias para la compilación y correcta ejecución del software HTK, Julius, y otros programas detallados más adelante.

Además se ha utilizado el lenguaje "Shell Script (Bash)" para realizar la mayoría de las tareas, ya que permite generar bucles de forma sencilla, encadenar y paralelizar las llamadas a las herramientas HTK y Julius. También se aprovecharon las herramientas de tratamiento de datos y audio que llevan incorporadas la mayoría de las distribuciones de Linux.

AWK, *sed* y *grep*, entre otras herramientas, han sido utilizadas para la limpieza de las etiquetas de la base de datos, adaptarlas al formato HTK.

- **Sox**

"Sound eXchange" es la "navaja suiza" del tratamiento de audio. Permite efectuar diversos tratamientos sobre un archivo de sonido.

En este caso se ha utilizado para reducir la frecuencia de muestreo de los archivos hasta 8 kHz, para contar las horas de voz (excluyendo el silencio) de la base de datos, y durante el filtrado de la misma para detectar y eliminar los archivos vacíos.

- **Praat**

Herramienta de uso libre de alta potencia en el tratamiento de señales de voz. Además de observar la forma de onda y el espectro de archivos de audio, se ha utilizado para calcular la SNR de los archivos de sonido, dado que permite separar la señal de voz de su silencio. También se ha utilizado en el cálculo del *pitch* (frecuencia fundamental) promedio de las grabaciones de la base de datos.

- **CMU-Cambridge SLM Toolkit**

Para la generación de parte de los modelos de lenguaje se ha utilizado la herramienta "CMU-Cambridge Statistical Language Modeling Toolkit v2.05".

Esta aplicación permite generar un modelo de lenguaje con cualquier combinación de N-gramas a partir de un texto y de un archivo de configuración que sirve para señalar los límites de las frases.

- **SRI Language Model**

Se trata de otra completa aplicación para la generación de modelos de lenguaje con diversos índices de N-grama. También permite la mezcla de modelos de lenguaje y tiene diversos elementos personalizables.

- **HTK**

Se ha utilizado el software HTK (Hidden Markov Model ToolKit) para la obtención de los modelos acústicos, inspirándose en el proceso básico de entrenamiento de su manual de referencia, *HTKBook* [2], pero con diversas modificaciones para adaptarlo a la base de datos y a los objetivos específicos del proyecto.



Figura 4.1: Logotipo de HTK3.

Originalmente fue desarrollado en el "Machine Intelligence Laboratory" del Departamento de Ingeniería de la Universidad de Cambridge. Se trata de una herramienta de uso libre pero no comercializable cuya última actualización tuvo lugar en febrero de 2009.

A pesar de ello es una de las herramientas más potentes de libre uso que existen en el mercado, utilizándose con éxito en tareas de reconocimiento de voz, síntesis de voz, reconocimiento de texto manuscrito y otras muchas tareas relacionadas con el aprendizaje automático.

Internamente se compone de diversas herramientas que permiten desde la generación de gramáticas cerradas, modelos de lenguaje sencillos con bigramas, grabación de archivos de audio y editores de etiquetas de archivos hasta las herramientas más potentes como *HERest* (utilidad que aplica el algoritmo de Baum-Welch) que es la parte central y más costosa de todo el proceso de entrenamiento.

También consta de un reconocedor propio que utiliza el algoritmo de Viterbi,

HVite, que también sirve para realizar alineamientos forzados que mejoran la calidad del entrenamiento al alinear los fonemas de los ficheros y asignarles marcas de tiempo.

■ Julius

Para el reconocimiento se decidió prescindir, tras un periodo de prueba, de las herramientas básicas incluidas en el HTK (*HVite* y *HDecode*) y utilizar en su lugar otra herramienta más moderna, potente, rápida, libre, actualizada y con más funcionalidades que es Julius.



Figura 4.2: Logotipo del software Julius.

Se trata de un reconocedor de habla continua de gran vocabulario (*Large Vocabulary Continuous Speech Recognition - LVCSR*) de código abierto y de alto rendimiento cuya particularidad consiste en efectuar dos pasadas de reconocimiento, utilizando en la primera bigramas (pares de palabras con una probabilidad de ocurrencia asociada) y en la segunda trigramas (tríos de palabras) de derecha a izquierda o entidades superiores de N-gramas hasta deca-gramas.

Con diccionarios de hasta 60.000 palabras se aproxima al tiempo real en la mayoría de las situaciones. A pesar de que el tiempo real no es objetivo de este proyecto, dicho programa ha ayudado a reducir el tiempo de finalización del mismo.

Utiliza modelos acústicos en formato ASCII de HTK y modelos de lenguaje en formato ARPA y soporta distintos tipos de HMMs, como los de estados compartidos y los de gaussianas compartidas.

Fue desarrollado inicialmente en la universidad de Kyoto y ahora es mantenido por el Instituto Nagoya de Tecnología.

4.2. Proceso de entrenamiento

En esta sección se detallan los pasos seguidos para generar los modelos acústicos y de lenguaje que se utilizarán en los experimentos.

4.2.1. Discriminación de los archivos de audio

Para evitar que el proceso de entrenamiento modele en exceso fonemas con demasiado ruido de fondo o ruidos explosivos como clics y golpes, se ha realizado un purgado previo de la base de datos en la que se eliminan todos los archivos con etiquetas que corresponden a archivos excesivamente mal grabados, aquellos con una SNR menor de 10 dB y los que tenían archivos de audio vacíos o incompletos.

4.2.2. Generación de archivos “.lab”

La base de datos de Telefónica tiene dos tipos de ficheros: los .wav que contienen los datos de la forma de onda y los .info que contienen los metadatos asociados a los ficheros de forma de onda. Por cada fichero “.wav” hay un fichero “.info”.

Lo que se debe hacer es extraer el campo de transcripción de los archivos “.info” que acompañan a los “.wav”. Hubo que prestar especial atención a la codificación interna de los ficheros. Nuestro sistema utiliza por defecto el sistema de codificación UTF-8 y los ficheros venían en ISO-8859-1. Es recomendable convertirlos a la codificación utilizada en la propia terminal de Linux para facilitar el funcionamiento de HTK.

Después se coloca cada palabra de la transcripción en una línea diferente y se termina el archivo con un punto, dado que éste es el formato que HTK utiliza.

bocadillo

de

atún

.

4.2.3. Generación del diccionario

Para generar el diccionario de entrenamiento se separan todas las palabras de las transcripciones en líneas diferentes y se ordenan por orden alfabético eliminando duplicados.

Posteriormente, se fonetizan siguiendo las normas del castellano. En este proyecto de fin de carrera se ha utiliza un conjunto de 28 fonemas diferentes para el español. Como el sistema de entrenamiento utiliza el modelo de pausa corta “sp”, se añade este al final de cada palabra.

El diccionario toma la siguiente forma:

```

abadejo a bb a dd e x o sp
abandono a bb a n d o n o sp
...
zurrar z u rr a r sp
zurrón z u rr o n sp
silence sil
!ENTER sil
!EXIT sil

```

Las palabras "silence" "!ENTER" "!EXIT" son símbolos especiales. *HVite* usa "silence" para el alineamiento forzado y tanto *HVite* como Julius y otros reconocedores usan "!ENTER" Y "!EXIT" para indicar el comienzo y final de frases.

4.2.4. Generación de los archivos de transcripciones y listas

HTK precisa que todas las transcripciones estén juntas en un archivo de extensión ".mlf", comenzado por la etiqueta "#!MLF!#" seguido del nombre de archivo entre comillas y la transcripción como en los archivos ".lab":

```

#!MLF!#
"/001.lab"
bocadillo
de
atún
.
"/002.lab"
visitó
la
ciudad
de
cuenca
.

```

Después, aplicando los diccionarios, la herramienta *HLEd* (HTK Label Editor) separa cada palabra en sus fonemas:


```
#!MLF!#  
"/001.lab"  
sil  
b  
o  
k  
a  
dd  
i  
ll  
o  
sp  
dd  
e  
sp  
a  
t  
u  
n  
sp  
sil
```

Por otro lado, se generan todas las listas de archivos que precisa HTK para el entrenamiento: Las de archivos “.wav”, archivos “.mfc” y archivos “.lab”.

4.2.5. Generación del modelo de lenguaje

Para generar el modelo de lenguaje, primero se recopilan todas las frases de la base de datos de entrenamiento (Tabla 3.1) en un único archivo y se edita para que comiencen y terminen con las etiquetas “!ENTER” y “!EXIT”.

Inicialmente se utilizaba la herramienta *HBUILD* de HTK para construir modelos de lenguaje, pero sólo permite generar modelos de lenguaje de bi-gramas. Por ello a continuación se procedió a utilizar la herramienta CMU-LM y se generó un modelo de lenguaje acotado (en el que todas las palabras a reconocer están en el modelo y no hay ninguna palabra del modelo fuera de las frases a reconocer) basado en trigramas para el reconocimiento en Auto-test y de Test.

4.2.6. Extracción de características

Tras realizar diversas pruebas iniciales, se decidió efectuar una extracción de características mediante Mel Frequency Cepstral Coefficients, de ahora en adelante MFCCs.

La principal ventaja de utilizar coeficientes cepstrales es que normalmente están decorrelados, lo que permite que se puedan utilizar covarianzas diagonales en los HMMs.

Para el cálculo de los MFCCs se empleó una ventana de Hamming con pre-énfasis y normalización de energía. El tamaño de la ventana de análisis es de 25 ms con un desplazamiento de 10 ms.

Los MFCCs son generados mediante la herramienta de HTK, *HCopy* que efectúa dicha extracción de características generando un archivo con el mismo nombre que el “.wav” pero con extensión “.mfc”. Estos archivos son necesarios para el entrenamiento mediante *HERest* y sirven para acelerar el reconocimiento al no tener que calcularlos de nuevo cada vez.

Los ficheros de audio están en formato WAV a 8 kHz y los vectores así generados tienen 39 componentes.

4.2.7. Generación de los HMM

A continuación se detallan cada uno de los pasos que se han seguido para conseguir los HMM finales que se utilizarán en la fase de pruebas.

4.2.7.1. Generación de mono-fonemas

El primer paso consiste en crear un archivo prototipo que contenga un ejemplo del vector de características a utilizar, una representación de cuántos estados va a tener el modelo (tres en nuestro caso más uno inicial y uno final) así como la forma de la matriz de transiciones.

A continuación, dado que no se dispone de un alineamiento inicial de los ficheros de audio con sus marcas de tiempo, se llama a la función *HCompV* de HTK que se encarga de transformar este modelo prototipo inicial en otro en el que se añaden todas las medias y las varianzas globales que se han extraído de toda la base de datos de entrenamiento y establece todas las gaussianas a este valor. De disponer de un alineamiento inicial se podría utilizar la función *HInit*.

Una vez obtenidos estos datos iniciales, se genera el MMF copiando el modelo del *proto* en cada uno de los fonemas que utilizamos.

Se genera también a partir del archivo *proto* y del archivo *'vFloors'* calculado por *HCompV* el archivo *'macros'* que se utiliza para, en cada nueva reestimación, indicar

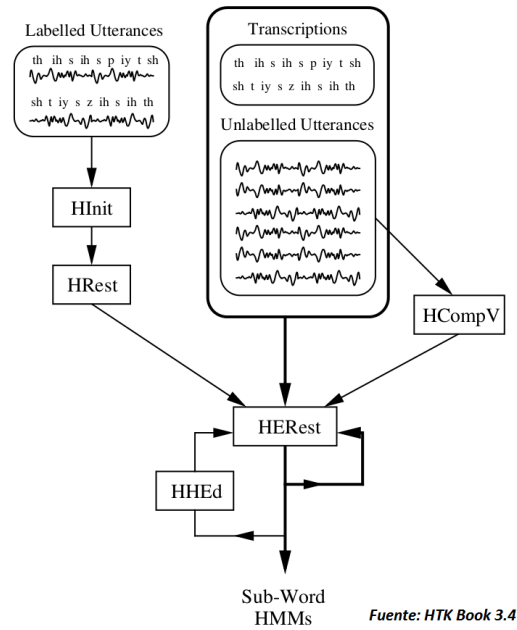


Figura 4.3: Entrenamiento de HMMs de fonemas.

la forma del archivo 'hmmdefs' a generar.

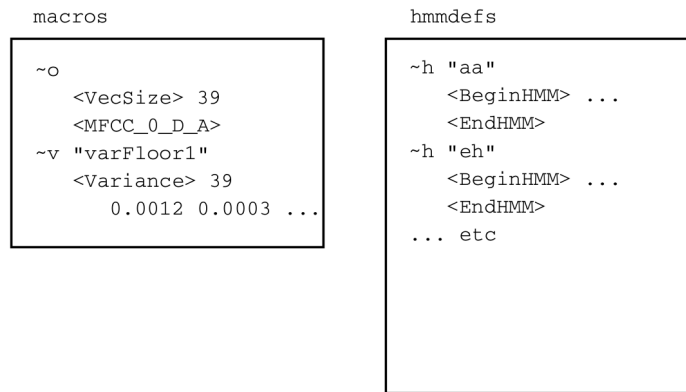


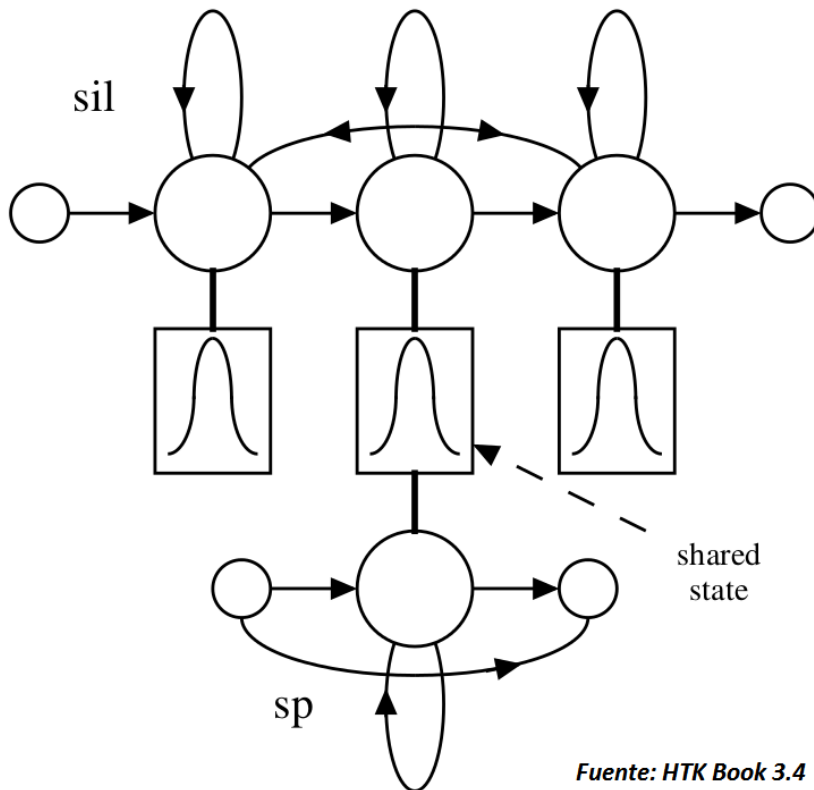
Figura 4.4: Ejemplo de archivo de 'macros' y de 'hmmdefs'.

A continuación ya se puede comenzar la reestimación de los modelos mediante la herramienta *HERest*, que aplicando el algoritmo de Baum-Welch [2] se encarga de calcular los modelos para cada uno de los fonemas que se ha definido a partir de los datos de entrenamiento.

4.2.7.2. Optimización de los modelos de silencio

Para hacer el sistema más robusto ante posibles ruidos impulsivos, se ligan los estados 2 a 4 y 4 a 2. Esto permite que sean estados individuales los que absorban los ruidos impulsivos. También, ligar el estado 4 al 2 permite que esto ocurra sin salir del modelo.

Este ligado se realiza mediante un *script* que se envía la herramienta *HHEd*.



Fuente: HTK Book 3.4

Figura 4.5: Mejora del modelo de silencio.

A la par se añade un modelo de pausa corta de un estado llamado 'sp' que comparte su estado central con el fonema 'sil' pero que se trata de un "tee-model" que es aquel que tiene una transición directa de la entrada a la salida.

Después se vuelven a reestimar los modelos utilizando la herramienta *HERest*.

4.2.7.3. Alineamiento forzado y entrenamiento inicial

Una vez se dispone de los modelos de silencio mejorados, se pasa a realizar un alineamiento forzado. Esto sirve, aprovechando los modelos que ya se han logrado

generar, para alinear a nivel de fonema la ocurrencia de cada uno de ellos dentro de cada archivo de audio y extraer marcas de tiempo para mejorar así el trabajo que realiza *HERest*.

Dicho alineamiento forzado se efectúa mediante la herramienta de HTK *HVite* que permite hacer un reconocimiento inicial de los archivos de la base de datos para extraer las marcas de tiempo de cada fonema. El archivo alineado de salida se llama "aligned.mlf".

Una vez efectuado dicho proceso, se vuelven a reestimar los modelos.

4.2.7.4. Generación de tri-fonemas

El siguiente objetivo es generar un set de tri-fonemas dependientes del contexto.

En primer lugar se utiliza la herramienta *HLEd* para expandir el archivo de etiquetas recién alineado en sus posibles tri-fonemas. Además genera una lista de todos los tri-fonemas vistos en las frases de entrenamiento.

La representación final de los fonemas se define como "*word-internal*" y toma la siguiente forma:

sil h+o h-o+l o-l+a l-a sp m+u m-u+n u-n+d n-d+o d-o sp

Aunque para algunos reconocedores se podrían listar sólo los tri-fonemas encontrados en las frases de entrenamiento, al utilizar en las pruebas iniciales *HDecode*, se optó por directamente generar todos los tri-fonemas posibles de las combinaciones de los mono-fonemas sin 'sp' y sin 'sil'. Esta lista que contiene todos los posibles tri-fonemas se llama "*fulllist*".

Mediante un script que se aplica a la herramienta *HHEd* se procede a generar todos los tri-fonemas mediante la clonación de su fonema central, volviendo a reestimar con *HERest* a continuación. En este punto ya se ha conseguido una lista de HMMs continuos de tri-fonemas dependientes del contexto.

4.2.7.5. Ligado de estados de tri-fonemas

Con el fin de reducir el coste computacional y de caracterizar mejor los tri-fonemas de los que no se dispone de datos de entrenamiento suficientes, se procede a efectuar un ligado de estados.

Aplicando un árbol de decisión, se unen los estados similares de los tri-fonemas que no tienen realizaciones suficientes a aquellos que tienen más realizaciones. Ajustando el umbral 'RO' en el script de *HHEd*, se puede ajustar el nivel de clusterización. Un valor muy alto puede causar que todos los tri-fonemas se acaben asociando a sus

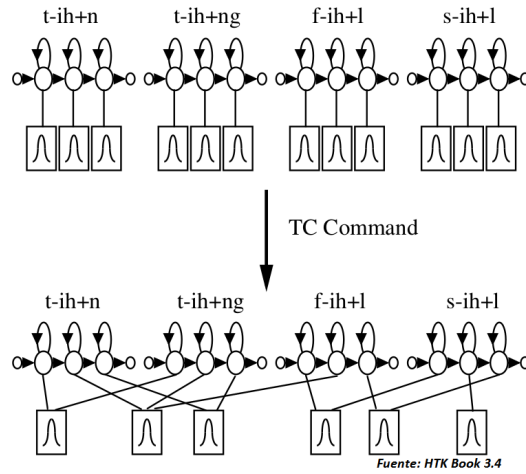


Figura 4.6: Ejemplo de modelo de estados ligados.

mono-fonemas iniciales, lo que evidentemente provocaría reconocimientos más veloces pero en principio de peor calidad.

Dejando este umbral a 0 se eliminan sólo los estados que no han sido entrenados, aproximándolos a sus estados más parecidos. En función de cada base de datos este factor puede variar por lo que es recomendable realizar un barrido en un rango de valores de esta variable para encontrar su óptimo.

mktri.hed

```
CL list
TI T_a {(*-a+*,a+*,*-a).transP}
TI T_b {(*-b+*,b+*,*-b).transP}
...
```

tree.hed

```
R0 150 stats

QS "L_NonBoundary" { *-* }
QS "R_NonBoundary" { **+ }
QS "L_Silence" { sil-* }
QS "R_Silence" { **sil }
QS "L_Vocales" { a-*, e-*, i-*, o-*, u-* }
QS "R_Vocales" { **a, **e, **i, **o, **u }
...
```

4.2.7.6. Incremento paulatino del número de gaussianas

Como paso final, para mejorar la calidad del sistema, lo que se realiza es un incremento del número de gaussianas.

Una gaussiana puede no ser suficiente para modelar todas las posibles realizaciones de ese fonema. Por ello, HTK permite incrementar el número de gaussianas para mejorar el modelo, incrementando así la tasa de reconocimiento.

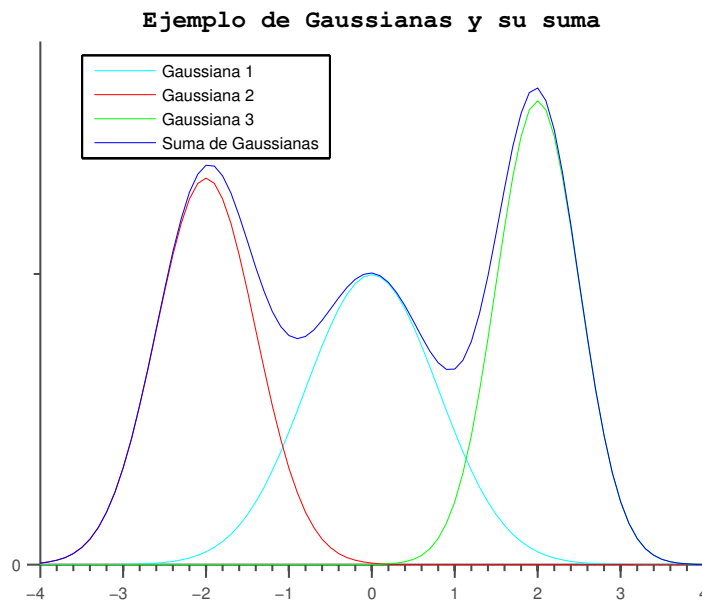


Figura 4.7: Ejemplo de suma de múltiples gaussianas.

El procedimiento utilizado para incrementar las gaussianas ha sido la herramienta *HHEd* con el comando 'MU' (Mixture Up). Se incrementa en uno el número de gaussianas y luego se realizan una o más reestimaciones con *HERest*. Después se repite el proceso hasta llegar al número de gaussianas requeridas.

Tras varias pruebas que se detallan en el punto siguiente 4.3 se ha llegado a la conclusión de que para nuestra base de datos, el número óptimo de gaussianas es de 16.

4.3. Reconocimiento de verificación

Una vez terminado el proceso de entrenamiento, se procede a la evaluación de la calidad del sistema generado.

Como se ha mencionado anteriormente, los primeros pasos en la evaluación y

mejora de este reconocedor se dieron utilizando las herramientas básicas incluidas *HVite* y *HDecode*. Se abandonaron por su pobre rendimiento tanto en velocidad como en tasas obtenidas, cambiándolos por otro reconocedor más versátil, moderno y en constante evolución como Julius.

Julius permite utilizar los modelos acústicos generados mediante HTK por lo que se puede hacer una evaluación directa de la calidad de los mismos. Dispone de infinidad de parámetros a configurar, siendo muy importante ajustarlos a la forma en que se haya entrenado en HTK.

Tiene la posibilidad de extraer directamente el vector de características de los archivos “.wav” lo que permite tratar cualquier archivo sin tener que efectuar la extracción de este vector mediante la herramienta *HCopy* cada vez que se varíe el tamaño de ventana u otros parámetros.

Se ha realizado un script para convertir la salida de Julius a formato HTK con el fin de poder utilizar el programa *HResults* de HTK para evaluar la salida.

Dicho programa da una salida en formato estándar métrico US NIST FOM.

```
----- Overall Results -----
SENT: %Correct=72.40 [H=724, S=276, N=1000]
WORD: %Corr=95.83, Acc=92.12 [H=6619, D=55, S=233, I=256, N=6907]
=====
```

Siendo “SENT - Correct” la tasa de frases correctas totales:

$$\%Correct = \frac{H}{N} \times 100 \% \quad (4.1)$$

”WORD - Corr” la de palabras correctas:

$$\%Corr = \frac{H}{N} \times 100 \% \quad (4.2)$$

”WORD - Acc” es la precisión, teniendo en cuenta también el número de inserciones negativamente:

$$Acc = \frac{H - I}{N} \times 100 \% \quad (4.3)$$

Se ha utilizado un diccionario de unas 12.000 palabras, y un modelo de lenguaje de trigramas acotado y adaptado a las frases, donde no hay palabras desconocidas y todas las palabras que se deben reconocer están en el diccionario.

Se han generado 6 grupos distintos de control para la evaluación del reconocedor. Los 4 primeros son de Auto-test, es decir, se prueba a reconocer archivos que se han

empleado para entrenar los modelos acústicos. Los otros dos grupos son de archivos que no forman parte de la base de datos de entrenamiento.

La segmentación se ha hecho entorno a la SNR de los archivos para ver de paso cómo afecta ésta a la calidad del reconocimiento. Además, todas las listas de test están balanceadas, es decir, que contienen un número proporcional de los archivos de cada base de datos de origen. Contienen 1.000 archivos cada una, que se corresponde con aproximadamente 70 minutos de audio:

Tabla 4.1: Grupos de archivos de prueba.

Grupo	Lista de Archivos	Nº Archivos	Duración (min)	Rango SNR (dB)
Auto-test	Autotest_10_20	1000	70	10 - 20
	Autotest_20_30	1000	70	20 - 30
	Autotest_30_40	1000	70	30 - 40
	Autotest_40_99	1000	70	40 - 99
Test	Test_30_40	1000	70	30 - 40
	Test_ruido	1000	70	< 10

Una vez definidos los grupos de control, se efectuó un experimento para definir el número óptimo de gaussianas a utilizar:

- **Elección del número óptimo de gaussianas:**

Este experimento se basó en entrenar los HMM incrementando de una en una hasta las 64 gaussianas. Este entrenamiento llevó cerca de 72 horas.

Una vez generados los modelos, se utilizaron para el reconocimiento todos ellos entre 1 y 64 gaussianas. Se reconocieron los seis conjuntos de archivos y se dibujaron gráficas para ver la tendencia.

Aunque en los reconocimientos de Auto-test se observa que incrementa la tasa de acierto siempre que se incrementa el número de gaussianas, al contrastarlo con las bases de datos de Test se observa que a partir de las 16 gaussianas la tasa de aciertos comienza a degradarse.

Esto se debe a que al incrementar el número de gaussianas así como las reestimaciones, se están sobre-adaptando los HMM a la base de datos de entrenamiento, lo que provoca que al llegar nuevos locutores o frases no entrenadas se reconocerán peor, y esto dista del objetivo del sistema.

Contrastando todas las gráficas y tablas se decidió tomar 16 gaussianas, dado que era el número óptimo en la mayoría de los casos hasta el que siempre subía la tasa de aciertos y luego volvía a bajar.

Aunque ocasionalmente 18 o 21 gaussianas dieran tasas similares, se decidió utilizar 16 porque a misma tasa de aciertos el procesado será más rápido.

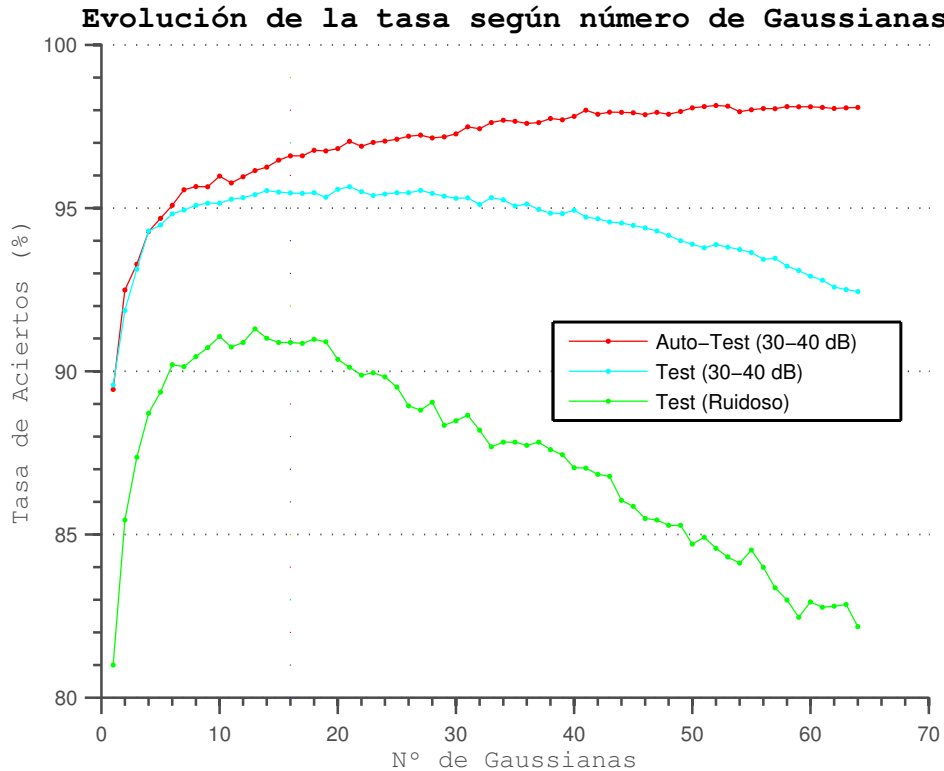


Figura 4.8: Gráfica selección de gaussianas.

Para más información consultar el apéndice A.

■ **Elección del número óptimo de archivos de entrenamiento:**

Este experimento se realizó para tratar de determinar cuál es el número óptimo de ficheros de entrenamiento.

Se entrenaron 26 modelos con distinto número de archivos hasta las 16 gaussianas y se hicieron reconocimientos de prueba para validar la calidad del sistema. Se extrajeron las siguientes gráficas de los mismos.

En ellas se puede observar una tendencia logarítmica entre el número de archivos y el incremento de la tasa de aciertos. Aunque para los archivos de mayor calidad el incremento en la tasa de aciertos tiende a estabilizarse a partir de los 120.000 ficheros de entrenamiento con incrementos cada vez menores en los aciertos. Para los archivos de peor calidad comprendidos en el grupo “Test Ruidosos” y en el grupo “Auto-test 10-20 dB” todavía se observan grandes incrementos cercanos al 1% al pasar de 220.000 a 240.000 ficheros de entrenamiento.

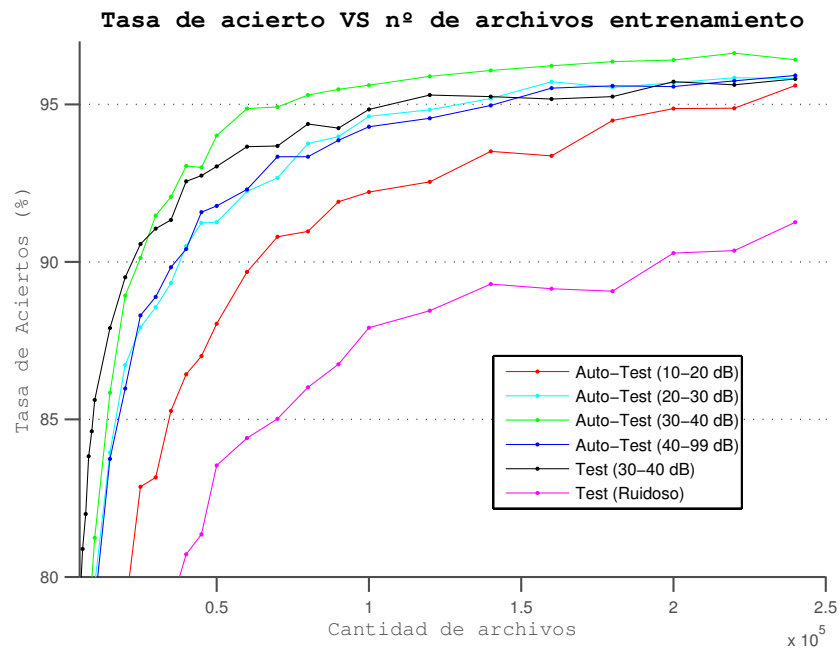


Figura 4.9: Gráfica para la selección del número de archivos de entrenamiento en unidades naturales.

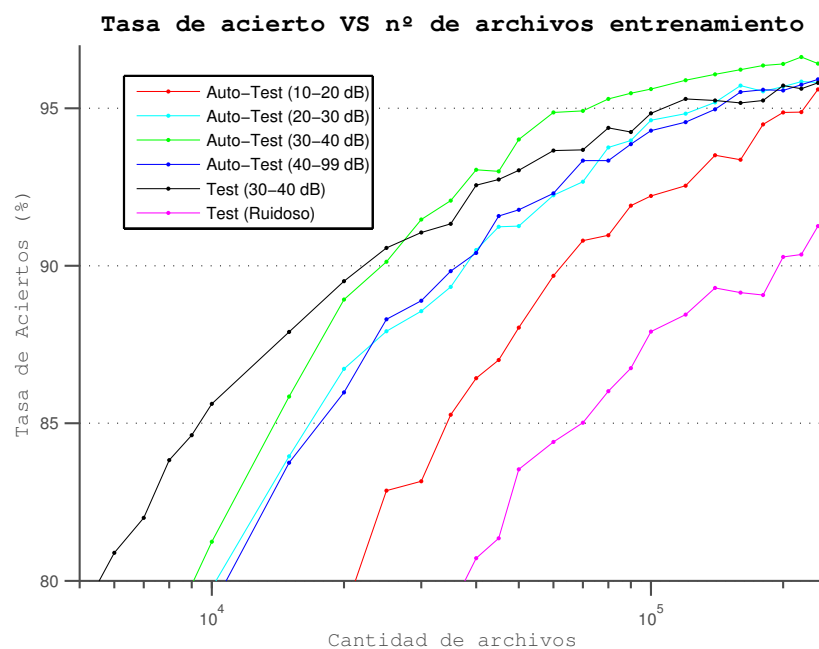


Figura 4.10: Gráfica para la selección del número de archivos de entrenamiento en unidades logarítmicas.

Esto sugiere que hay margen para añadir más ficheros de entrenamiento con el fin de mejorar especialmente el reconocimiento en situaciones de mala relación señal a ruido así como en los casos que además de una pobre SNR se dan también eventos fonéticos, como golpes, chasquidos...

Desafortunadamente no se dispone de más archivos de entrenamiento con los que prolongar esta prueba. Por tanto se deben utilizar todos los archivos disponibles para el entrenamiento.

Para los resultados completos consultar el apéndice B.

Tras estos experimentos iniciales, reconociendo las listas de archivos explicadas anteriormente se han podido generar diversas modificaciones en el sistema hasta obtener tasas entre el 95 % y el 98 % de acierto en Auto-test y entre el 95 % y 97 % en Test. Una vez llegados a este punto, se consideró que el sistema estaba listo para la fase de experimentación.

Los intervalos de confianza se han calculado para el 95 %.

Tabla 4.2: Resultados pruebas test para 16 gaussianas.

Grupo de archivos	% WORD	% ACC	% SENT
SNR 10-20 dB (Auto-Test)	95,74 ± 1.04	90,28	86,26
SNR 20-30 dB (Auto-Test)	96,18 ± 0.66	94,04	77,78
SNR 30-40 dB (Auto-Test)	96,60 ± 0.71	93,92	77,19
SNR 40-99 dB (Auto-Test)	95,94 ± 0.72	94,38	82,23
SNR 30-40 dB (Test)	95,46 ± 0.89	92,27	73,47
SNR Ruidosos (Test)	90,88 ± 1.44	83,47	76,35

Capítulo 5

Optimización del Sistema y Resultados

En este capítulo se llevará a cabo la descripción detallada de la optimización del sistema de referencia generado (Capítulo 4). Siguiendo la motivación de este proyecto, la optimización mencionada tendrá como base la adaptación de los modelos de lenguaje a los tópicos o temáticas de los contenidos a reconocer. Para ello, se tratará la problemática existente, se explicará el escenario en el que se desarrollará esta tarea; y por último, se llevarán a cabo una serie de experimentos, cuyos resultados serán analizados y presentados gráficamente.

5.1. Descripción del problema

Como se explicó en la subsección 2.2.2.4, el objetivo de los modelos estadísticos de lenguaje es asignar probabilidades a secuencias de palabras y su aplicación más destacada es en reconocimiento de habla, donde estos proporcionan probabilidades que tienen mayor prioridad, ayudando así a evitar confusiones con pronunciaciones ambiguas, similares acústicamente.

En ocasiones, y de acuerdo a la tarea que se vaya a desarrollar, puede convenir llevar a cabo una adaptación de estos modelos [15]. Las razones por las cuales es interesante la adaptación, son las siguientes:

- El lenguaje permite expresar la misma idea de formas distintas, afectando en la descripción de las cosas. El vocabulario evoluciona con el tiempo, apareciendo nuevas palabras, nuevos usos de ellas, etc.
- Una misma palabra o cadena de palabras puede dar lugar a distintos significados,

dependiendo del tópico. Por ejemplo, “ratio de interés”, es diferente en una aplicación bancaria que en una campaña publicitaria de un producto.

- Las personas tienden a ajustar su lenguaje en función de la tarea del momento. La sintaxis empleada en una conversación técnica es diferente a la empleada en una conversación informal con un amigo.
- El estilo del discurso puede cambiar debido a la variedad de los factores como el estatus socio-económico, estado emocional, etc.

Además, se ha comprobado en investigaciones anteriores [16], que las discordancias en la precisión del reconocimiento de un tópico concreto, se debe en mayor medida al modelado del lenguaje, mientras que el modelado acústico tiene poca influencia en la tasa de error de palabra (*Word Error Rate* - *WER*).

Por otro lado, la adaptación de los modelos de lenguaje supone una reducción considerable de la perplejidad (en torno al 15-20 %, en algunos trabajos de las Universidades de Carnegie Mellon y de Southern California), para un tópico concreto; y esto se traduce en una mejora en el reconocimiento, disminuyendo la tasa de error de palabra.

Existen multitud de trabajos de investigación relacionados con esto, por ejemplo, la detección instantánea de tópicos en discursos e incluso, la detección del idioma del locutor o locutores. Siguiendo la línea de algunos trabajos, que emplean vídeos de YouTube [17] para estudiar este reconocimiento basado en el tópico, también se hará uso de contenidos audiovisuales de Internet.

En las siguientes subsecciones se explicará cómo se adaptarán los modelos estadísticos del lenguaje al tópico específico en cada reconocimiento, con el fin de optimizar el sistema de referencia generado.

5.2. Escenario y generación de modelos

En primer lugar, antes de llevar a cabo los experimentos que cuantifiquen la optimización del sistema de referencia, es preciso definir un escenario en el que se va a trabajar. En él se detallarán los tópicos o temáticas de los contenidos a reconocer y los elementos necesarios para la generación de los modelos del lenguaje que se van a emplear.

5.2.1. Tópicos

Se ha pretendido, desde el primer momento, realizar experimentos en temas cotidianos conocidos por todos y que sus contenidos fuesen concretos. De esta forma, los

resultados obtenidos servirán para conocer los pros y contras que tienen estos tópicos a la hora de ser reconocidos y generar su transcripción, así como su posible integración en posibles aplicaciones prácticas.

Los tópicos elegidos han sido los siguientes:

- Economía: Han sido escogidos contenidos audiovisuales de conferencias de temas de actualidad como la situación de crisis en España, la gestión del Gobierno y la actuación de los bancos, entre otros.
- Deporte: Este material será extraído del espacio de programación del telediario, correspondiente a dos cadenas españolas de televisión. Tratarán noticias de la actualidad deportiva de los deportes más mediáticos.
- Moda: Dichos contenidos se tomarán de vídeos de bloggeras aficionadas a la moda y los productos de cosmética. En ellos se contarán las tendencias del momento, combinaciones estilísticas y lugares donde encontrar productos y complementos.

5.2.2. Corpus de entrenamiento de los modelos

Una vez conocidos los tópicos de los contenidos, es preciso disponer de volúmenes de texto que constituyan los diferentes corpus necesarios para el entrenamiento de los modelos del lenguaje. Se necesitarán corpus para cada tópico y así obtener modelos del lenguaje específicos, los cuales se adaptarán a la tarea que se desee.

Los corpus empleados para llevar a cabo este estudio serán los siguientes:

- **Corpus genérico**: Su nombre se debe a que dispondrá de un vocabulario cotidiano y general. Consistirá en una agrupación de textos del periódico *El País* principalmente, del corpus de *CORLEC* del *Laboratorio de Lingüística Informática* de la UAM, y de transcripciones de diálogos de películas y entrevistas. Cuenta con un tamaño máximo de 100 millones de palabras aproximadamente.
- **Corpus de economía**: Contiene un amplio volumen de datos correspondiente a la sección de Economía del periódico *El País* fundamentalmente, unido a textos obtenidos de otras fuentes como *El Economista*, *Diario Expansión*, *Cinco Días* y algunos blogs de economía. Cuenta con un tamaño máximo de 20 millones de palabras.
- **Corpus de deporte**: Se obtiene un amplio volumen de texto extraído de noticias de la sección deportiva del periódico *El País*, junto a los de otros diarios

deportivos como *As*, *Marca*, *Mundo Deportivo* y *Sport*. A esto se le unen contenido de otras webs deportivas.

Cuenta con un tamaño máximo de 20 millones de palabras.

- **Corpus de moda:** Este corpus es generado mayoritariamente, como todos los anteriores, por textos de la sección Moda y Tendencias del periódico *El País*. A esto, se ha de añadir contenidos de otras webs de moda como *Tendencias*, *Vogue*, e incluso de blogs dedicados al tema.

Este volumen de datos tiene un tamaño máximo de 8 millones de palabras¹, al igual que los correspondientes al resto de tópicos.

Estos corpus serán manejados según se requiera, siendo su tamaño un factor variable; es decir, se dispondrá de mayor o menor volumen en función de las pruebas a realizar.

5.2.3. Dicionarios específicos

Como ha sido explicado en la subsección 2.2.2.3, a partir de un determinado vocabulario puede obtenerse uno de los componentes fundamentales de la arquitectura del reconocedor: el diccionario.

De acuerdo a los experimentos que se desarrollarán posteriormente, se hará uso de diferentes diccionarios como son los que siguen:

- **Diccionario genérico:** Está constituido por el vocabulario del corpus genérico, explicado anteriormente, además de:
 - ✧ 600 nombres y apellidos más frecuentes de España, extraídos de la base de datos del *Instituto Nacional de Estadística (INE)*.
 - ✧ 400 ciudades más pobladas de España y sus comunidades autónomas y provincias, obtenidos de *Wikipedia*.
 - ✧ 100 ciudades más pobladas del mundo y las 300 ciudades más ricas del mundo, con todos los países y continentes obtenidos de *Wikipedia*.
 - ✧ Diccionario base de Español de España de 72.000 palabras de *LibreOffice*.
 - ✧ Diccionario de las 76.000 palabras más comunes del Español según la base de datos *CREA* de la *Real Academia Española (RAE)*.

¹Es preciso mencionar de antemano, que al disponer de un corpus de menor tamaño (menos de la mitad que en los otros dos tópicos), puede verse repercutida la tasa de acierto de palabra en los ficheros relativos a este tópico.

✧ Diccionario de la base de datos de entrenamiento propia.

Con el fin de crear este tipo de diccionario, se ha llevado a cabo un proceso de eliminación de palabras repetidas, y gracias a la ordenación por frecuencia, se han generado dos diccionarios de referencia: un diccionario genérico de 60.000² palabras más frecuentes del Español de España y otro de 130.000 palabras, que permita ampliar el abanico de reconocimiento del reconocedor. Estos dos diccionarios de vocabulario genérico serán los que se junten con las palabras específicas de cada tópico. Además de estos dos diccionarios mencionados, se realizarán otros de diversos tamaños, ya que serán necesarios en el Experimento 4 (subsección 5.4.6).

- **Diccionarios específicos economía, deporte y moda:** Estos diccionarios son generados con el diccionario genérico de 60.000 palabras, mencionado anteriormente, más las N palabras específicas de cada tópico, que no aparecen en él.

En la siguiente tabla se muestran todas las que se han recopilado de cada tópico:

Tabla 5.1: Número de palabras específicas de cada tópico.

Tópico	N
Economía	2.000
Deporte	4.000
Moda	4.000

- **Diccionarios acotados de los contenidos de test de economía, deporte y moda:** Estos diccionarios están formados por el vocabulario exacto que se emplea en cada una de las temáticas de los contenidos audiovisuales a reconocer, sin contener palabras extra.

Todos estos diccionarios han sido sometidos a un proceso de eliminación de palabras repetidas y de alguna corrección ortográfica debida a erratas de las fuentes de dónde se han obtenido.

Cabe destacar que, debido al fonetizador de reglas del que se dispone, propio del Español, ha sido necesaria una adaptación de los diccionarios específicos de cada tópico. Dichos tópicos cuentan con numerosos términos extranjeros (unos 4.500 vocablos),

²Se ha escogido este número de palabras debido a que el reconocedor ahorra memoria RAM si se usan menos de 65.000 y según las especificaciones de Julius, hasta esta cifra podría garantizarse el reconocimiento a tiempo real.

por lo que se han fonetizado minuciosamente con el set de fonemas disponibles; evitando la fonetización automática, para que afecte en la menor medida posible a la decodificación.

También se han tenido en cuenta los términos correspondientes a siglas (en torno a 400), que se han tenido que fonetizar manualmente tras un deletreo previo.

5.3. Base de Datos de Test

Tras la elección de los tópicos con los que se va a trabajar, se han escogido los contenidos audiovisuales de Internet acordes a ellos.

Una vez descargados estos contenidos, los cuales son vídeos en formato MP4, se ha llevado a cabo el proceso correspondiente para la extracción del audio en formato WAV, que es el soportado por el sistema de reconocimiento que se ha desarrollado, y a una frecuencia de 8 kHz, 16 bits y 1 canal, coincidiendo así con la de los ficheros de la Base de Datos de entrenamiento (Capítulo 3) .

De todo el volumen de audio, se han realizado conjuntos de voz de 70 minutos de cada tópico (subsección 5.2.1) análogamente a los conjuntos de test empleados en la evaluación del sistema de referencia, y de esta forma poder comparar algunas de las tasas de acierto de palabra que se obtengan. Cabe destacar que, en todo momento, se ha intentado mantener la variabilidad de locutor en la medida de lo posible.

Además, una breve caracterización de estos ha servido para conocer mejor los vídeos con los que se va a trabajar y ayudado a la comprensión de los resultados que se obtendrán más adelante.

Tópico	Locutores	Voz masculina	Voz femenina	Proximidad micro	Música de fondo
Economía	Menos de 10	Sí	No	Sí	No
Deporte	Más de 10	Sí	Sí	Sí	Sí
Moda	Menos de 10	No	Sí	No	Sí

Tabla 5.2: Características de los contenidos audiovisuales del entorno de test.

Una vez descritas las condiciones iniciales y los agentes influyentes, se va a definir la serie de experimentos llevados a cabo, con su correspondiente análisis de resultados.

5.4. Experimentos

Tras definir el entorno de trabajo, se procede al análisis exhaustivo de la batería experimental que se ha desarrollado. En primer lugar, se hará una introducción que

refleje las condiciones iniciales de las que parte el sistema de reconocimiento de referencia. Sucesivamente, irán tomando parte en la optimización las distintas pruebas realizadas, con sus consiguientes resultados.

5.4.1. Condiciones iniciales: Características del sistema de referencia

Antes de disponerse a realizar los experimentos, será interesante conocer aquellas características del sistema de referencia (Capítulo 4), que puedan suponer limitaciones en la eficacia a la hora del reconocimiento y permitan conocer dónde se encuentra el punto máximo posible, en el cual se consiga una tasa de acierto de palabra mayor.

1. Voz leída frente a voz espontánea: Como se ha explicado en el Capítulo 3, la base de datos de entrenamiento estaba formada por habla leída; mientras que en estos experimentos se va a proceder al reconocimiento de contenidos audiovisuales de Internet, los cuales son mayoritariamente de voz espontánea. Esto supondrá una serie de efectos en la pronunciación como: coarticulación, velocidad del habla, que da lugar a efectos fonéticos como la influencia de fonemas consonánticos en vocálicos. Además, el habla espontánea se ve afectada por “falsos” comienzos en el discurso, repetición de palabras, y palabras de relleno (ah, eh, mm, etc.) entre otros factores.
2. Distintos tipos de canal: Al igual que en el caso anterior, no hay coincidencia del canal empleado en todos los casos. El corpus de la base de datos de entrenamiento empleada contiene voz telefónica mayoritariamente, y en menor medida, voz limpia; mientras que el conjunto de test contiene voz grabada con distintos tipos de micrófonos tomada en medios de comunicación y en entornos académicos y caseros, con distintas condiciones de adquisición.

Al tratarse de distinto ancho de banda (3,8 kHz para telefonía fija), será precisa la unificación de los mismos, provocando así una posible pérdida de información fonética, que suponga una influencia importante en los resultados de decodificación.

3. Condiciones del entorno: A pesar de haber entrenado con una serie de fenómenos del entorno controlados, con el fin de modelar mejor el espacio acústico, en el corpus de test pueden aparecer algunos tipos y niveles de ruidos, multitud de voces e incluso sonidos musicales de fondo, los cuales pueden influir negativamente en el proceso de reconocimiento.

4. Vocablos extranjeros: La ausencia de extranjerismos en el corpus de entrenamiento, junto a la fonetización adaptada de ellos al set de fonemas del Español, supondrán un efecto extra (correspondencia fonética inexacta) en el proceso de decodificación, cuando estos aparezcan en los distintos tópicos.

5.4.2. Descripción general de los experimentos

En este apartado se mostrará una descripción resumida de los experimentos y el objetivo que pretende alcanzar cada uno de ellos.

En primer lugar, se quiere resaltar que los primeros experimentos se han realizado con una única pasada de reconocimiento (*Left-to-Right* en el diagrama de Trellis), debido al ahorro del coste computacional que supone el uso únicamente de bigramas en el modelo del lenguaje, obteniendo unos resultados preliminares para conocer la situación del sistema de reconocimiento implementado. A partir de ahí, y con una visión general de los parámetros óptimos de entrada (tamaños de corpus, de diccionarios...), los últimos experimentos se realizarán empleando una segunda pasada de reconocimiento (*Right-to-Left* en el diagrama de Trellis) haciendo uso de N-gramas en los modelos del lenguaje, con el fin de obtener los mejores resultados posibles, y cuantificar a su vez esta mejora.

Con el fin de conocer la tasa máxima alcanzable, se propone el siguiente experimento:

- **Experimento 1:** Consistirá en el reconocimiento de los tres conjuntos de test (economía, deportes y moda). Se emplean modelos de lenguaje y diccionarios acotados, los cuales han sido generados con los corpus exactos extraídos de las transcripciones de los mismos. El objetivo será ver la respuesta del sistema de reconocimiento, disponiendo de una entrada de voz espontánea no entrenada.

El propósito de los siguientes experimentos será encontrar el tamaño óptimo del corpus de entrenamiento de los modelos del lenguaje y de los diccionarios, tanto de vocabulario genérico como vocabulario de cada tópico:

- **Experimento 2:** Se realizará el reconocimiento en cada tópico con un modelo del lenguaje genérico y un diccionario genérico de 60.000 palabras. Aquí tendrá lugar una variación del tamaño del corpus empleado, desde 200.000 palabras hasta 100 millones de palabras, con el fin de obtener el modelo de lenguaje óptimo que consiga la tasa de acierto de palabra máxima.
- **Experimento 3:** Este experimento tendrá como objetivo, dar un paso más y complementar el Experimento 2. Consistirá en el reconocimiento de los contenidos de los tópicos con diccionarios específicos (formados por las 60.000 palabras

más frecuentes del Español más las palabras extra de cada tópico). A su vez, se realizarán modelos del lenguaje añadiendo corpus específico de cada tópico (fijo) al corpus genérico (variable) empleado anteriormente.

- **Experimento 4:** En esta ocasión se realiza una variación del tamaño del diccionario genérico, con el fin de detectar el tamaño óptimo que consigue la mayor tasa de acierto. Para ello se modificará el vocabulario desde 10.000 palabras hasta llegar a 130.000 palabras, aplicándolo en el reconocimiento de los conjuntos de Auto-test y Test (de las mismas características que la base de datos de entrenamiento).
- **Experimento 5:** Se realiza este experimento con el fin de obtener resultados concluyentes acerca de la penalización en la tasa de acierto de palabra, por la eliminación de una cantidad de palabras del vocabulario genérico, y la posterior ganancia que supone añadir las palabras específicas del tópico.

Como se ha indicado en la subsección 5.2.3, cada tópico tiene una cantidad de palabras específicas que no se encuentran en el vocabulario genérico. Como el tamaño de estos conjuntos de palabras no llega a ser excesivamente elevado en proporción al del vocabulario genérico, se eliminarán del diccionario genérico tantas palabras como el número total de esas palabras específicas, y se añadirán estas últimas; puesto que ir restando y añadiendo palabras gradualmente no aportaría datos relevantes a este estudio.

El objetivo de los experimentos siguientes es comprobar el efecto que tiene emplear mezclas ponderadas (también llamado interpolación en la bibliografía), obtener la ponderación óptima en cada tópico y a su vez las consecuencias de aplicar una segunda pasada de reconocimiento con los modelos de lenguaje y diccionarios que dan lugar a las mejores tasas de acierto de palabra:

- **Experimento 6:** En este experimento se estudiará la influencia que supone realizar los modelos del lenguaje como una mezcla ponderada entre el modelo genérico y el modelo propio de cada tópico, a diferencia de como se realizó en el Experimento 3, con la unión de diferentes corpus y el posterior entrenamiento de los modelos. De esta forma, se realizará una pasada de reconocimiento estableciendo un barrido variando la combinación de pesos entre el modelo genérico y el modelo del tópico (desde 0,2/0,8 hasta 0,8/0,2, en pasos de 0,1).

- **Experimento 7:** En este experimento se pretende llevar a cabo una segunda pasada adicional de reconocimiento³, empleando 3-gramas. Se ha elegido esta unidad de tres palabras, ya que en artículos científicos de la bibliografía empleada no recomendaban emplear más de 4-gramas por la carga computacional que supone y los buenos resultados que se pueden conseguir con ellos. El fin de este experimento es cuantificar la mejora que esta segunda pasada provoca, manteniendo la mezcla ponderada de modelos del lenguaje empleada en el Experimento 6.
- **Experimento 8:** Del mismo modo que se ha ejecutado el experimento anterior con una segunda pasada de reconocimiento, el cambio introducido, en este caso, será la integración de hasta 5-gramas en los modelos del lenguaje, con el fin de comparar resultados en las tasas de reconocimiento del Experimento 7.

El objetivo de este último experimento será comprobar qué tipo de modelo de lenguaje da lugar a las mejores tasas de acierto de palabra: modelos de mezcla ponderadas (genérico y tópico concreto) o un modelo global de mezclas ponderadas (genérico y todos los tópicos). Continuando sin duda con el empleo de la segunda pasada con N -gramas, donde N será la que proporciona mejores resultados ($N=5$, Experimento 8):

- **Experimento 9:** Consiste en la generación de un modelo del lenguaje global, empleando los modelos específicos de los tres tópicos que se han tratado hasta ahora, junto con el modelo genérico (corpus de 100 millones de palabras), variando las ponderaciones para ver qué pesos proporcionan los mejores resultados. Los pesos de cada modelo del lenguaje en la mezcla global serán partiendo de un peso alto para el modelo genérico, pesos iguales para deporte y moda y en menor medida tomará parte el modelo de economía. Así se irá aumentando los pesos de los modelos de deporte y moda, disminuyendo el peso del modelo específico y manteniendo casi fijo el peso del modelo de economía. Estos pesos han sido elegidos teniendo en cuenta los resultados de los Experimentos 6, 7 y 8 (subsecciones 5.4.8, 5.4.9 y 5.4.10). Por otro lado, se hará uso de un diccionario global formado por el diccionario genérico de 130.000 palabras y las 8.000 palabras específicas de los tres tópicos (2.000 de economía, 4.000 de deporte y 4.000 de moda, y eliminación posterior de las palabras repetidas). El objetivo será conocer si es más conveniente un modelo de este estilo, dónde se engloben todos

³Cabe destacar que Julius permite una doble pasada de reconocimiento: la primera *Left-to-Right* empleando bigramas, y una segunda *Right-to-Left* empleando N -gramas

los tópicos, o uno que únicamente se adapte al tópico concreto (Experimento 8).

Para finalizar, se realizará un último experimento en el cual se compararán los resultados obtenidos hasta aquí con los sistemas de los líderes tecnológicos de nuestros días, como Apple y Google:

- **Experimento 10:** Se tratará de realizar el reconocimiento de los contenidos de los tres tópicos empleados hasta ahora con los sistemas de reconocimiento siguientes: Dictado Automático de Apple⁴ y la Web API Speech Demonstration de Google⁵.

A continuación se mostrarán detalladamente los resultados obtenidos para experimento individualizado.

5.4.3. Experimento 1: Modelos del lenguaje acotados en cada tópico

Las condiciones para este experimento han sido:

Tabla 5.3: Condiciones establecidas en el Experimento 1.

Diccionario	Modelo Lenguaje	Nº Pasadas	Orden (N-gramas)
Acotado	Acotado	1	2

Los resultados obtenidos para cada tópico, reconocido con su respectivo diccionario y modelo de lenguaje, han sido los siguientes:

Tabla 5.4: Tasa de acierto de palabra con modelos del lenguaje acotados.

Tópico	%WORD
Economía	54,54
Deporte	58,36
Moda	33,98
Test Sistema Ref.	95,46

Como puede observarse en los resultados obtenidos, los tópicos de economía y deporte alcanzan una tasa de acierto de palabra mayor que el tópico de moda. Se puede concluir:

⁴<http://support.apple.com/es-es/HT5449>

⁵<http://www.google.com/intl/es/chrome/demos/speech.html>

- Partiendo de los resultados obtenidos en la evaluación del sistema de referencia, también con modelos del lenguaje y vocabulario acotados y ajustados a los propios conjuntos de test, en el mejor de los casos significa una bajada absoluta de la tasa de 37,1 puntos (tópico de deporte), mientras que en el peor de los casos supone una bajada de 61,5 puntos (tópico de moda). Esta caída de la tasa de reconocimiento es la penalización provocada por el reconocimiento de señal de habla espontánea, el canal y las condiciones del entorno (ruido, música, etc.), como se ha comentado en la subsección 5.4.1.
- La diferencia tan elevada en el tópico de moda es causada en primer lugar por las condiciones de adquisición del habla. Los contenidos audiovisuales han sido grabados por particulares, difiriendo de los medios de adquisición profesionales propios de las cadenas de televisión. Por último, se detecta la influencia del *pitch* de los locutores, ya que este se encuentra en torno a los 220 Hz, y el sistema tiene problemas para reconocer voz con estas características. La frecuencia fundamental media de un locutor puede adquirir cualquier valor en un rango del espectro [60 Hz, 300 Hz], por lo que para aquellas frecuencias que no están bien representadas en la base de datos de entrenamiento, aumentará la tasa de error [18].

A continuación se presenta una tabla relacionando el *pitch* de los contenidos que tratan de moda con la tasa de acierto de palabra en el reconocimiento, y se puede determinar que a medida que va aumentando el *pitch* promedio del locutor, va disminuyendo la tasa de acierto de palabra (Figura 5.1).

Tabla 5.5: Relación directa entre la tasa de acierto de palabra y el *pitch* promedio, en el tópico de Moda.

Fichero	% WORD	Pitch Promedio (Hz)
moda05.wav	38,20	209
moda04.wav	39,63	215
moda03.wav	37,08	220
moda01.wav	32,92	223
moda02.wav	22,94	240

5.4.4. Experimento 2: Modelo del lenguaje genérico para todos los tópicos

Las condiciones para este experimento han sido:

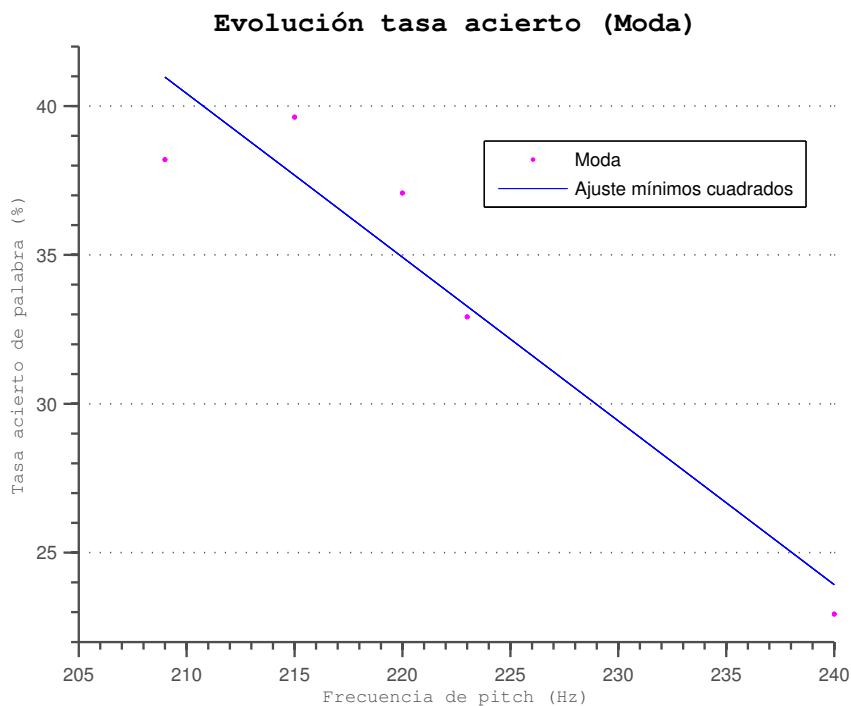


Figura 5.1: Tendencia de la tasa de acierto de palabra en función del *pitch* promedio, en el tópicos de Moda.

Tabla 5.6: Condiciones establecidas en el Experimento 2.

Diccionario	Modelo Lenguaje	Nº Pasadas	Orden (N-gramas)
Genérico (60.000 palabras)	Genérico (Corpus variable)	1	2

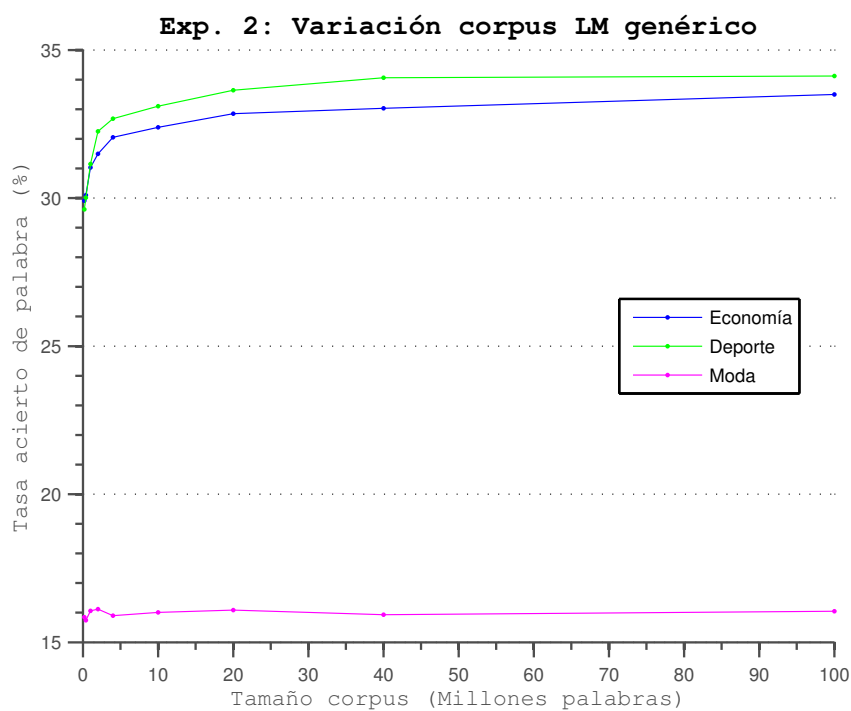


Figura 5.2: Variación del corpus de entrenamiento del LM genérico para cada tópicos.

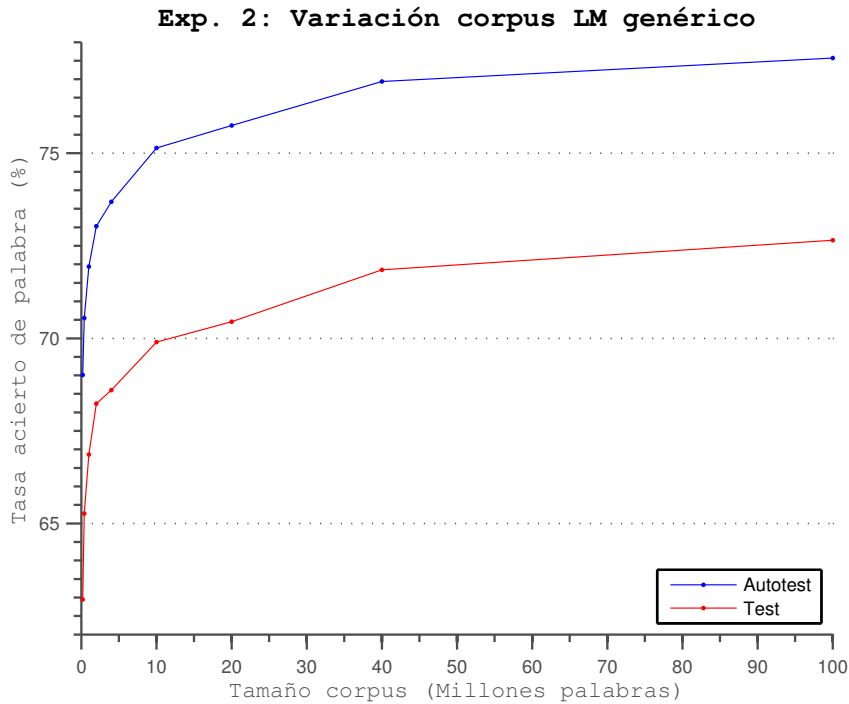


Figura 5.3: Variación del corpus de entrenamiento LM genérico para Auto-test y Test.

En los resultados obtenidos en la Figura 5.2 se puede observar cómo, tanto en los tópicos de economía y deporte cuyas tasas del 33,5 % y 34,12 % respectivamente, el reconocimiento funciona mejor conforme se aumenta el tamaño del corpus, siendo óptimo el máximo tamaño para el que se han obtenido resultados (100 millones de palabras). Mientras que en la temática de moda se tienen tasas del 16,12 %, para un corpus menor que el anterior (2 millones de palabras). Es posible concluir que:

- Los resultados obtenidos en el tópico de moda, se ven reducidos casi a la mitad comparados con los del resto de tópicos; esto es debido a los factores que han sido detallados en el Experimento 1 (subsección 5.4.3).
- El efecto de emplear un modelo del lenguaje abierto y no ajustado a los contenidos que se quieren reconocer, supone en la tasa de acierto una reducción relativa del 40 % aproximadamente. Por otro lado, al aplicar los mismos modelos del lenguaje al conjunto de test empleado en la evaluación del sistema de referencia (con las mismas características que la base de datos de entrenamiento), resulta una reducción relativa del 23,9 % en la tasa de acierto de palabra, siendo esta del 72,65 % (Figura 5.3). Comparando ambos resultados, se produce una diferencia del 16,1 % dependiendo de la semejanza o no de los datos de test

a los datos de entrenamiento (conjunto de frases de test frente a conjunto de contenidos audiovisuales).

5.4.5. Experimento 3: Modelos del lenguaje mezclando corpus

Las condiciones para este experimento han sido:

Tabla 5.7: Condiciones establecidas en el Experimento 3.

Diccionario	Modelo Lenguaje	Nº Pasadas	Orden (N-gramas)
Específico	Específico (Corpus genérico variable)	1	2

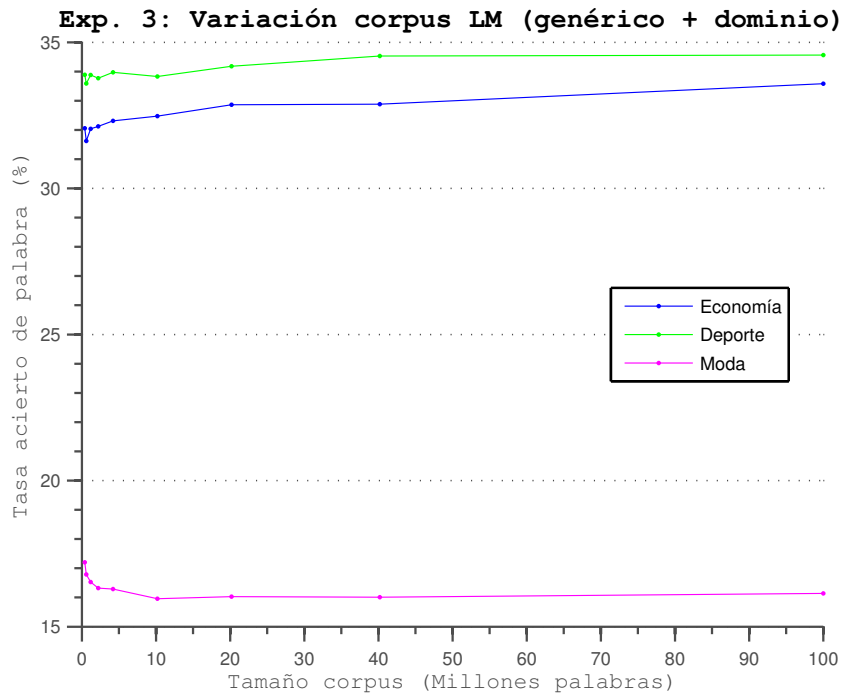


Figura 5.4: Variación del corpus de entrenamiento LM genérico y específico, con diccionario fijo (60.000 palabras).

Si se realiza una observación de la Figura 5.4, se puede determinar de forma generalizada que a mayor volumen de corpus genérico, se obtiene una mayor tasa de acierto de palabra. A su vez, al aumentar el tamaño del corpus específico de cada temática se obtiene una mejora adicional. Se concluye que:

- Los mejores resultados se corresponden a la generación de modelos del lenguaje basada en la mezcla de: en primer lugar, el corpus genérico mayor disponible (100

millones de palabras) y en segundo lugar, un corpus específico. En referencia a este último, el paso de un volumen de 200.000 palabras a 20 millones de palabras, supone un incremento de 1 punto aproximadamente en la tasa de acierto en algunos casos. Este efecto es más notable en el tópico de deporte, ya que al disponer de vocablos muy distintivos (deportivos y extranjerismos), al aumentar la proporción del corpus de deporte en el entrenamiento del modelo del lenguaje, se amplía el abanico de posibilidades adaptándose mejor a la tarea a reconocer.

- Esta ligera introducción de corpus específico no supone un incremento importante respecto al experimento anterior, donde el modelo del lenguaje empleado era únicamente formado por un corpus genérico.

5.4.6. Experimento 4: Diccionario genérico

Las condiciones para este experimento han sido:

Tabla 5.8: Condiciones establecidas en el Experimento 4.

Diccionario	Modelo Lenguaje	Nº Pasadas	Orden (N-gramas)
Genérico (Variable)	Acotado	1	2

Se obtiene que:

- Manteniendo el modelo de lenguaje genérico empleado en los experimentos anteriores (corpus de 100 millones de palabras), que es el óptimo del Experimento 2 (Figura 5.3), se obtiene que conforme aumenta el tamaño de diccionario, se alcanzan tasas mayores. La velocidad de crecimiento de la tasa de acierto de palabra, está inversamente relacionada con el tamaño del diccionario: a mayor volumen del vocabulario, el crecimiento de la tasa sigue aumentando pero muy lentamente.
- Las tasas de acierto obtenidas para un diccionario de 130.000 palabras (Figura 5.5), tanto para Auto-test (80,54 %) como para Test (75,01 %), suponen mejoras relativas del 3,8 % y 3,2 % respecto de utilizar un diccionario de 60.000 palabras como en el Experimento 2.

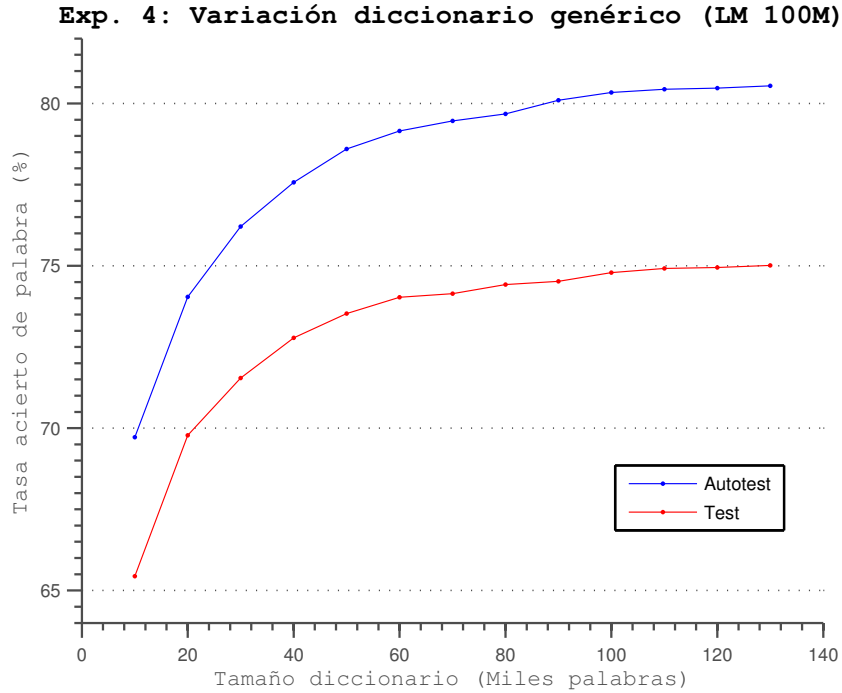


Figura 5.5: Variación del diccionario genérico con LM genérico fijo (100 millones de palabras).

5.4.7. Experimento 5: Diccionario genérico con vocabulario específico

Las condiciones para este experimento han sido:

Tabla 5.9: Condiciones establecidas en el Experimento 5.

Diccionario	Modelo Lenguaje	Nº Pasadas	Orden (N-gramas)
Específico (Modificación Genérico)	Específico (Unión corpus)	1	2

Los resultados de este experimento serán recogidos en la siguiente tabla:

Tabla 5.10: Penalización y ganancia en la tasa de acierto de palabra (%WORD) al eliminar palabras genéricas y añadir específicas.

Tópico	60.000 palabras	(60.000-N) palabras	(60.000-N) + N_{esp} palabras
Economía	32,86	32,80	32,86
Deporte	33,64	33,69	34,18
Moda	16,09	16,01	16,03

Tras analizar los resultados de la Tabla 5.10, se llega a las siguientes conclusiones:

- La penalización por eliminar del diccionario genérico de 60.000 palabras, N^6 palabras tantas como las específicas de los tópicos de economía y moda es de 0,06 puntos y 0.08 puntos respectivamente. Por otro lado en el tópico de deporte, este proceso no su pone una penalización, si no una ganancia de 0,05 puntos.
- La ganancia por añadir las palabras propias de cada tópico (N_{esp}) es de: 0,06 puntos en economía, 0,02 en moda y 0,49 en deporte. Esto último es debido a que en el reconocimiento de esta temática es beneficioso disponer de menor número de palabras genéricas y más vocabulario referente a nombres de deportistas, a los numerosos equipos, y a tecnicismos de cada disciplina, entre otros. Mientras que para los otros dos tópicos, carece de excesiva relevancia, tanto por la calidad de los audios como por la influencia en el vocabulario general.

5.4.8. Experimento 6: Modelos de lenguaje con mezclas ponderadas

Las condiciones para este experimento han sido:

Tabla 5.11: Condiciones establecidas en el Experimento 6.

Diccionario	Modelo Lenguaje	Nº Pasadas	Orden (N-gramas)
Específico	Específico (Mezcla ponderada LMs)	1	2

Observando la Figura 5.6 se puede concluir que:

- En el tópico de economía la mezcla óptima que da lugar a la mayor tasa de reconocimiento es: de entre el 50-70 % del modelo de vocabulario genérico y 30-50 % del modelo del lenguaje específico de este tópico (se elegirá la proporción 60-40 % al encontrarse en el punto medio). La razón de esta distribución de los pesos se debe a que los contenidos reconocidos, al tratarse de temas de la actualidad económica explicados de una manera llana, contienen vocablos propios de lo que se ha categorizado en este proyecto como lenguaje genérico.
- En los tópicos de deporte y moda, en cambio, resulta sumamente importante darle prioridad a los propios modelos del lenguaje específicos. En ambos casos, la mezcla ponderada óptima será del 20-30 % del modelo del lenguaje genérico y 70-80 % del modelo específico del tópico concreto. La razón de esto es que el contenido reconocido en este caso que contiene estructuras y numerosos

⁶Este número ha sido especificado en la Tabla 5.1

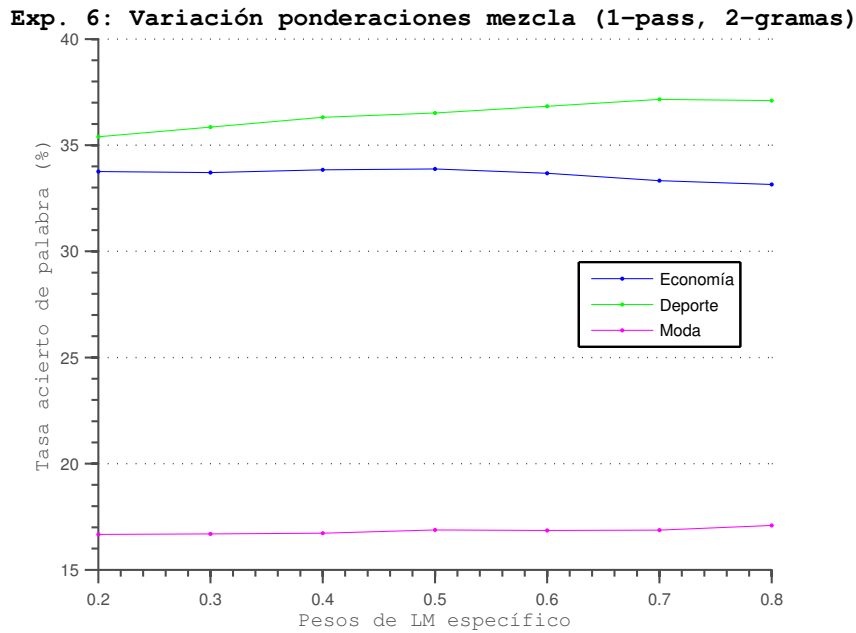


Figura 5.6: Variación de las ponderaciones de la mezcla de LMs (genérico y específico).

nombres propios tanto de deportistas, modelos, productos o incluso vocablos o expresiones hechas en otros idiomas.

- Cabe destacar que en el tópico de moda se ha empleado un corpus de 8 millones de palabras frente a los 20 millones del corpus de deporte, siendo un factor a tener en cuenta también en la tasa de acierto de los ficheros de este tópico.
- Este experimento aporta conocimiento sobre la mejora que supone la adaptación de los modelos empleando una mezcla ponderada, respecto de la generación de un único modelo a partir de la mezcla de corpus: en economía se produce una mejora de 0,1 puntos, nada relevante comparado con la obtenida en deporte o moda, 1,7 y 1 punto respectivamente.

5.4.9. Experimento 7: Modelos de lenguaje con mezclas ponderadas (2^a pasada, 3-gramas)

Las condiciones para este experimento han sido:

Tabla 5.12: Condiciones establecidas en el Experimento 7.

Diccionario	Modelo Lenguaje	Nº Pasadas	Orden (N-gramas)
Específico	Específico (Mezcla ponderada LMs)	2	3

Tabla 5.13: Comparativa de tasas de acierto de palabra (%WORD) en el reconocimiento con 1^a pasada única y 2^a pasada (3-gramas).

Tópico	1-pass (2-gramas)	2-pass (3-gramas)
Economía	33,84	36,76
Deporte	37,10	39,73
Moda	17,09	19,20

Analizando los resultados de la Tabla 5.13, se obtiene lo siguiente:

- En cuanto al tópico de economía, la mejora supone 2,82 puntos sobre realizar una única pasada de reconocimiento. En deportes, como se puede observar, se consigue aumentar la tasa de reconocimiento en 2,63 puntos y en el tópico de moda, 2,11 puntos. Generalizando, se puede concluir que generando modelos del lenguaje con 3-gramas se consigue un aumento considerable de la tasa de acierto de palabra entre 2 y 3 puntos absolutos.
- Esta mejora es debida a que, en estos tópicos y en cualquier otro, existen palabras que son más probables que otras que aparezcan juntas en el discurso. Tomando como ejemplo los temas de economía y deportes, que siguen un discurso más formal y estructurado, debido a que muchos de estos contenidos han sido extraídos de telediarios, se mantienen expresiones como “el conjunto blanco” (trigrama que hace referencia al Real Madrid) o “durante la crisis” (un trigrama muy empleado en el tópico de economía). Con ellos, se permite una mayor precisión en el reconocimiento de patrones, modelando estos conjuntos de palabras con una mayor probabilidad.

5.4.10. Experimento 8: Modelos de lenguaje con mezclas ponderadas (2^a pasada, 5-gramas)

Las condiciones para este experimento han sido:

Tabla 5.14: Condiciones establecidas en el Experimento 8.

Diccionario	Modelo Lenguaje	Nº Pasadas	Orden (N-gramas)
Específico	Específico (Mezcla ponderada LMs)	2	5

Una vez corroborado el incremento de la tasa de acierto de palabra en el Experimento 7, los resultados obtenidos con la integración de hasta 5-gramas en los modelos del lenguaje se pueden observar en las gráficas (Figuras 5.7, 5.8, 5.9). La evolución de la tasa de acierto de palabra, a medida que se va aumentando la N del modelo del lenguaje de N -gramas, para cada tópico, es creciente.

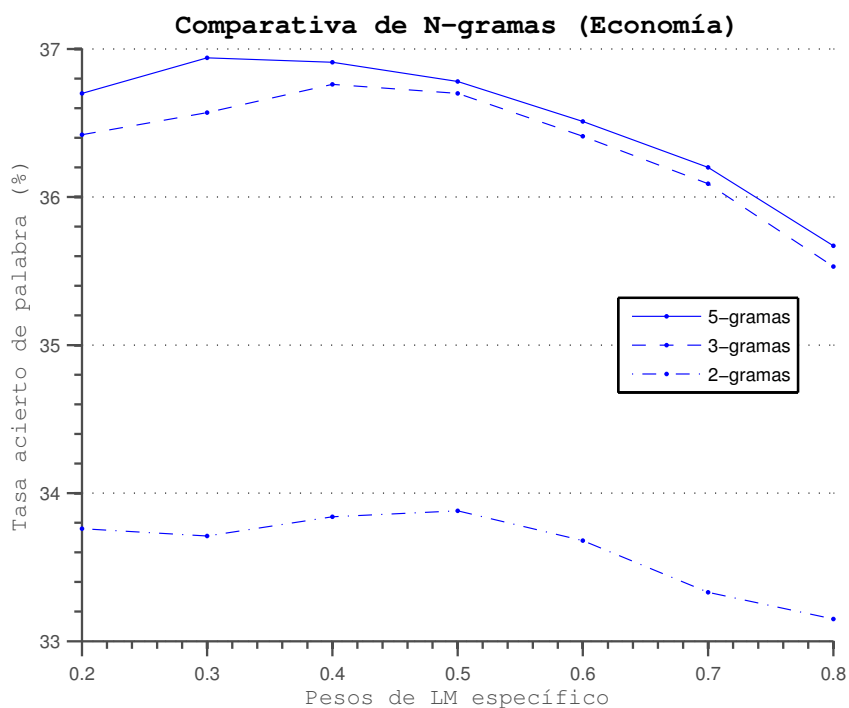


Figura 5.7: Comparativa del reconocimiento con N-gramas en el tópico de Economía.

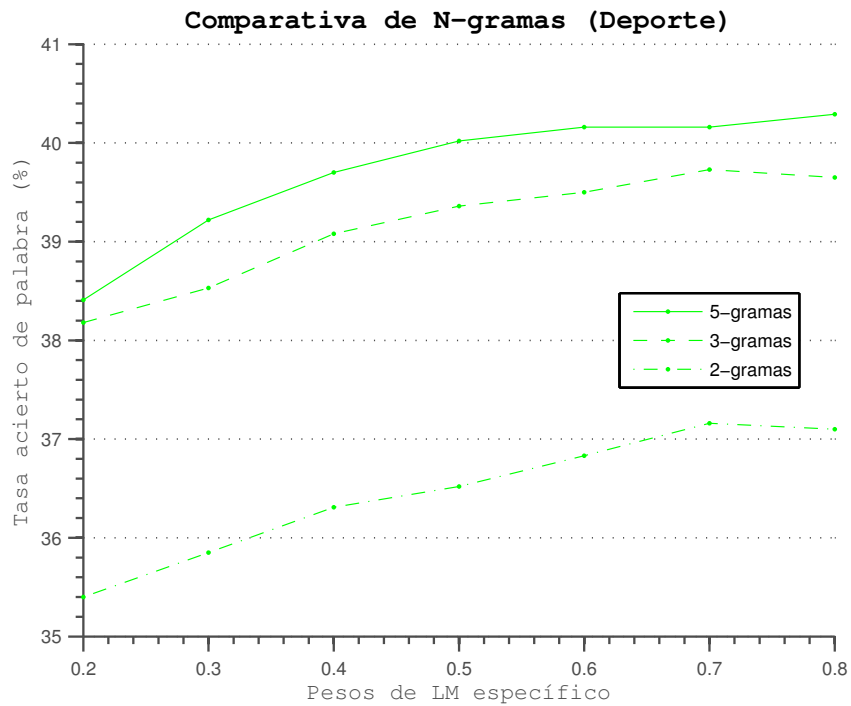


Figura 5.8: Comparativa del reconocimiento con N-gramas en el t3pico de Deporte.

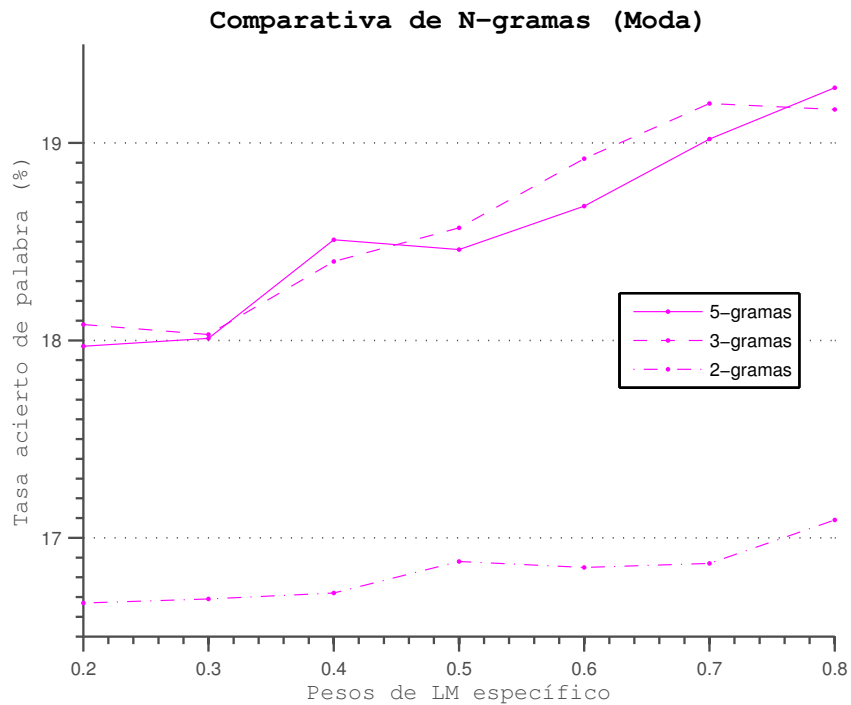


Figura 5.9: Comparativa del reconocimiento con N-gramas en el t3pico de Moda.

- En primer lugar, la cuantificación de esta mejora se realizará manteniendo los mismos pesos de cada modelo que en el experimento anterior, en las mezclas ponderadas de adaptación. De esta forma, se consiguen las siguientes tasas de acierto de palabra: en economía 3 % respecto a emplear una única pasada de reconocimiento con bigramas (casi 2 décimas mejor que 3-gramas); en deporte 3,2 % respecto a la primera pasada (más de 6 décimas mejor que 3-gramas); y en moda 2,2 % respecto a la primera pasada (tan sólo 1 décima mejor que 3-gramas).
- Se adoptará en la metodología de reconocimiento el entrenamiento de los modelos de lenguaje de hasta 5-gramas, ya que a efectos prácticos no supone ningún esfuerzo adicional ni gasto de recursos, y además proporcionan los mejores resultados en el reconocedor, hasta el momento, para esta serie de tópicos.

5.4.11. Experimento 9: Modelo del lenguaje global con mezclas ponderadas

Las condiciones para este experimento han sido:

Tabla 5.15: Condiciones establecidas en el Experimento 9.

Diccionario	Modelo Lenguaje	Nº Pasadas	Orden (N-gramas)
Global	Global (Mezcla ponderada LMs)	2	5

Se pretende conocer si es más conveniente un modelo de este estilo, donde se engloben todos los tópicos, o uno que únicamente se adapte al tópico concreto. En ambos casos, mezclándose con un modelo de vocabulario genérico.

Tabla 5.16: Tasas de acierto de palabra (%WORD) según las ponderaciones de la mezcla global de todos los LMs. El formato de la cabecera indica los pesos de cada modelo en %, siendo el orden el siguiente: Genérico-Deporte-Moda-Economía.

Tópico	50-20-20-10	40-25-25-10	30-30-30-10	25-35-35-5	15-40-40-5
Economía	37,03	37,17	36,57	36,39	35,74
Deporte	38,53	38,82	38,82	38,79	38,71
Moda	18,02	18,33	18,46	18,67	18,67

Tras analizar los resultados obtenidos se tiene que:

- La mejor tasa de acierto de palabra para los tres tópicos se consigue con una mezcla ponderada de la siguiente manera: la parte principal con un 40 % estará

ocupada por el modelo del lenguaje de vocabulario genérico, un 25 % tanto para los tópicos de deporte y moda respectivamente, y un 10 % del modelo del lenguaje de economía.

- Con esta mezcla ponderada, se obtienen mejores resultados que emplear únicamente un modelo del lenguaje genérico, como se ha podido comprobar también en experimentos anteriores. En economía se obtiene una tasa 8 décimas mayor, en deporte 2,3 puntos mayor y en moda se alcanzan 9 décimas más.
- En cambio, este modelo del lenguaje completo generado con el vocabulario global de 138.000 palabras, no alcanza el comportamiento de los modelos del lenguaje confeccionados a partir de la mezcla de uno genérico con el propio del tópico. Se puede observar que con el modelo completo, para economía, deporte y moda se alcanzan tasas de 37,17 %, 38,82 % y 18,33 % respectivamente; mientras que para mezclas ponderadas para cada modelo por separado dichas tasas de acierto son de 36,91 % (prácticamente sin diferencia), 40,29 % y 19,28 % respectivamente.

5.4.12. Experimento 10: Comparativa de resultados con los líderes del mercado

Como se ha explicado en el Capítulo 2, esta tecnología del reconocimiento de habla natural tiene una gran importancia en la actualidad. La inteligencia artificial despierta un gran interés, y su uso de manera intuitiva ha supuesto que nos sirvamos de ella sin darnos cuenta. Tanto Siri y el dictado automático de Apple, como los servicios de Google para búsquedas o “*Voice Actions*”, son los principales sistemas que conviven con la mayoría de nosotros tanto en terminales móviles como en ordenadores.

Por este motivo se realizará una comparativa entre los mejores resultados obtenidos por el sistema descrito y optimizado en este proyecto, y los sistemas disponibles por Google y Apple.

- **Preparación del entorno:** Para la realización de estas pruebas, se hará uso de la Web Speech API Demonstration⁷ y el servicio de Dictado Automático de Apple⁸ (Mac OS X Yosemite), y se configura el Español de España como idioma. Ambos sistemas tendrán como entrada los contenidos íntegros de audio de la Base de Datos de Test (sección 5.3) de los tópicos mencionados a lo largo de este documento, y la salida será cada una en función del sistema correspondiente.

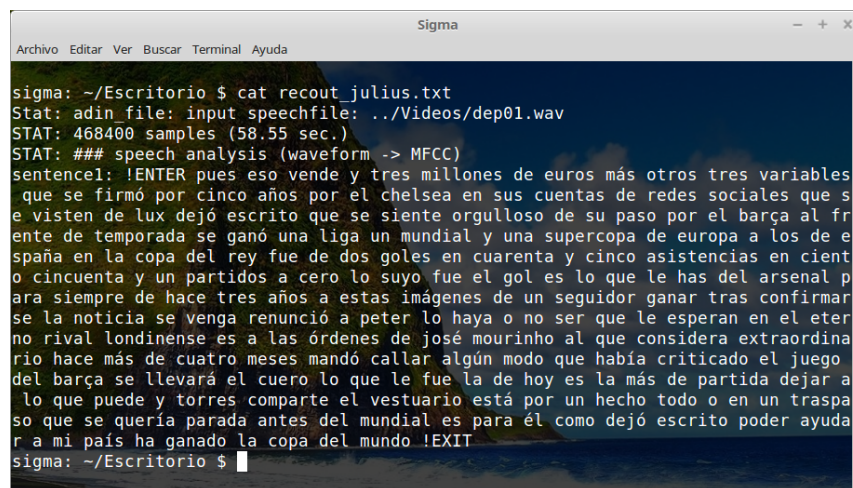
⁷<http://www.google.com/intl/es/chrome/demos/speech.html>

⁸<http://support.apple.com/es-es/HT5449>

Cabe destacar que se han llevado a cabo diversos ajustes del volumen de los ficheros para evitar fenómenos de saturación que pudiesen afectar a la calidad del reconocimiento.

En este experimento se hace uso de una herramienta de cortado de ficheros de Sigma Technologies, mientras que en el resto de experimentos, que han sido explicados hasta ahora, usaban el cortador automático de Julius. Esto puede afectar a los resultados que se obtienen a continuación.

Seguidamente se presenta el reconocimiento de un fragmento de los 70 minutos utilizados del tópico de deporte, obtenido a la salida de los diferentes sistemas a comparar.



```

sigma: ~/Escritorio $ cat recout_julius.txt
Stat: adin file: input speechfile: ../Videos/dep01.wav
STAT: 468400 samples (58.55 sec.)
STAT: ### speech analysis (waveform -> MFCC)
sentence1: !ENTER pues eso vende y tres millones de euros más otros tres variables
que se firmó por cinco años por el chelsea en sus cuentas de redes sociales que s
e visten de lux dejó escrito que se siente orgulloso de su paso por el barça al fr
ente de temporada se ganó una liga un mundial y una supercopa de europa a los de e
spaña en la copa del rey fue de dos goles en cuarenta y cinco asistencias en cient
o cincuenta y un partidos a cero lo suyo fue el gol es lo que le has del arsenal p
ara siempre de hace tres años a estas imágenes de un seguidor ganar tras confirmar
se la noticia se venga renunció a peter lo haya o no ser que le esperan en el eter
no rival londinense es a las órdenes de josé mourinho al que considera extraordina
rio hace más de cuatro meses mandó callar algún modo que había criticado el juego
del barça se llevará el cuero lo que le fue la de hoy es la más de partida dejar a
lo que puede y torres comparte el vestuario está por un hecho todo o en un traspas
o que se quería parada antes del mundial es para él como dejó escrito poder ayuda
r a mi país ha ganado la copa del mundo !EXIT
sigma: ~/Escritorio $

```

Figura 5.10: Salida resumida del sistema de reconocimiento desarrollado en este proyecto.

Web Speech API Demonstration

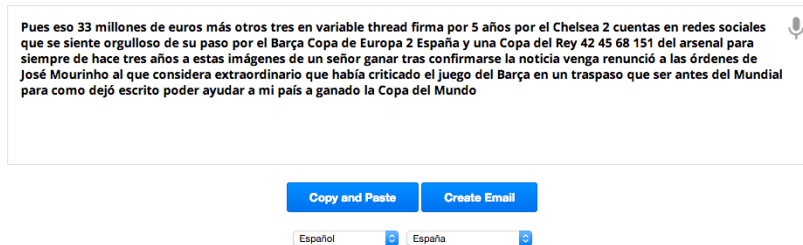


Figura 5.11: Salida del reconocimiento de la Web Speech API Demonstration de Google.

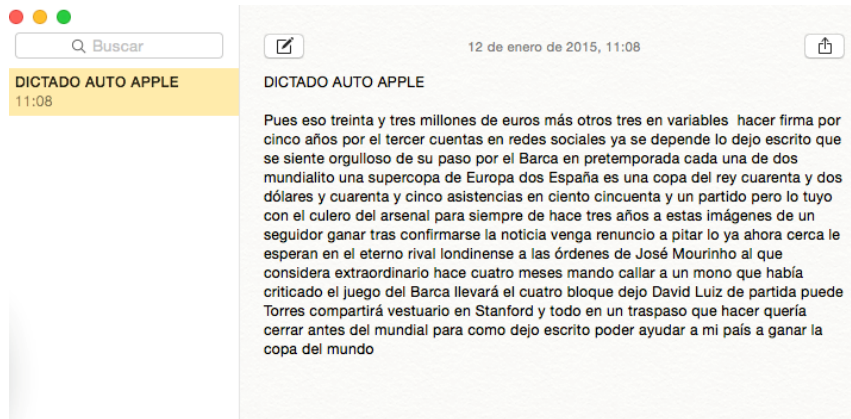


Figura 5.12: Salida del reconocimiento del Dictado Automático de Apple.

- Resultados:** Tras conocer las salidas de los reconocedores mencionados, será preciso un post-procesado para adaptarlas al formato empleado por la herramienta de evaluación utilizada. Como puede observarse en la Tabla 5.17, la tasa de acierto de palabra de nuestro sistema coincide prácticamente con la obtenida en la API de Google para los tópicos de economía y deporte, mientras que nuestro sistema obtiene mejores resultados en el tópico de moda.

Tabla 5.17: Comparativa de tasas de acierto de palabra (%WORD) en el reconocimiento de los tres tópicos, con los sistemas de Google y Apple.

Tópico	Σ	Google	Apple
Economía	39,23	39,67	33,94
Deporte	45,29	45,25	32,80
Moda	20,24	11,10	18,19

Como se ha explicado en la preparación del entorno, la herramienta de cortado de ficheros de Sigma Technologies, permite mejorar los resultados respecto a los del Experimento 8 (subsección 5.4.10), que eran hasta ahora los mejores. Por otro lado, se tienen en cuenta dos observaciones: la primera, que tanto los sistemas de Google como de Apple, a la salida obtienen un menor número de inserciones de palabras, lo que contribuye a una mejor legibilidad del resultado obtenido. Y la segunda, parece ser que el sistema desarrollado en este proyecto es un poco más robusto a condiciones de entorno no ideales (ruido, música de fondo...), ya que en fragmentos de audio con estas condiciones, no muestran la salida.

Capítulo 6

Aplicación del Sistema

Como se comentó en los objetivos del proyecto (sección 1.2), una vez implementado, ajustado, evaluado y optimizado el sistema de reconocimiento de habla natural, foco de este proyecto, se ha introducido en una solución comercial de la empresa.

La solución se encarga de indexar contenidos audiovisuales de Internet con el fin de realizar búsquedas en ellos. El sistema de reconocimiento transcribe en *background* el audio de los contenidos, generando las marcas de tiempo de cada palabra. Esto otorgará una mayor robustez y fiabilidad a la indexación, obteniendo mejores resultados.

A continuación, se muestra una demostración del funcionamiento de la aplicación y el papel que juega la transcripción de los contenidos audiovisuales en ella.

6.1. Demostración

En la demostración se utilizan vídeos correspondientes a diversas noticias extraídas del telediario. Una vez reconocidos e indexados dichos vídeos, se permite la búsqueda en diversas partes de la noticia: en el título y descripción del vídeo, y en el propio audio con su marca de tiempo asociada.

La interfaz muestra un diseño práctico e intuitivo para el usuario (Figuras 6.1 y 6.2). En primer lugar, en la parte superior se dispone de un campo de texto para introducir la cadena completa a buscar, una búsqueda aditiva o disyuntiva, etc. En la parte central, se presenta el visualizador de vídeos y una barra adicional que marca en color azul cada una de las apariciones de la búsqueda en el instante de tiempo correspondiente. En la parte inferior del vídeo, se mostrará información relevante de la noticia (título, fecha, temática y breve descripción). Por último, a la derecha se mostrará a modo de *scroll*, los resultados de la búsqueda ordenados por orden de

fiabilidad de aparición de esa búsqueda en el contenido del vídeo.

Ejemplo

Supongamos como ejemplo que introducimos en la barra de búsqueda la cadena “Presidente del Gobierno” (Figura 6.1), en ese preciso instante aparecerá en primera plana el vídeo que dispone de mayor probabilidad de que aparezca aquello que se ha solicitado.



Figura 6.1: Imagen de la aplicación. Búsqueda concreta.

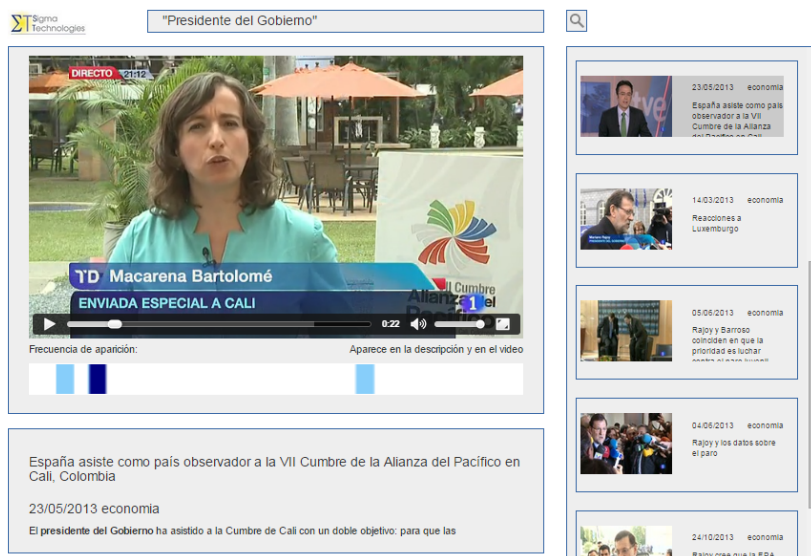


Figura 6.2: Imagen de la aplicación. Diferentes apariciones en el vídeo.

Puede comprobarse como a la derecha aparecerá marcado el vídeo que se está reproduciendo, y en la parte inferior derecha del visualizador se informa de que la búsqueda aparece tanto en el audio como en la descripción de la noticia, apareciendo en ella, la búsqueda marcada en negrita.

Además, como se muestra en la Figura **6.2**, la segunda marca de tiempo pasa a un color azul más oscuro, respecto a la Figura **6.1**, cambiando el locutor que menciona la cadena “Presidente del Gobierno”.

Capítulo 7

Conclusiones y Trabajo Futuro

7.1. Conclusiones

En este proyecto se ha conseguido el objetivo fundamental de desarrollar un sistema de reconocimiento de habla natural capaz de transcribir contenidos de audio en Internet. Para ello, se ha estudiado cómo afecta la adaptación del modelo del lenguaje al tópico.

Para este fin, en primer lugar se realizó un estudio detallado del estado del arte, analizando la arquitectura de los sistemas de reconocimiento de voz existentes en la literatura de referencia, y la aplicación de los Modelos Ocultos de Markov en ellos.

Tras este análisis, se estableció una metodología en el proceso de entrenamiento con su consiguiente proceso de evaluación directa. De este modo, se desarrolla un sistema de referencia, el cual ha sido sometido a una serie de ajustes paramétricos con el fin de obtener los mejores resultados sobre el conjunto escogido de la base de datos para tal fin.

Teniendo presente el objetivo del proyecto, junto con las motivaciones existentes para llevarlo a cabo, se ha analizado cómo afecta el empleo de modelos del lenguaje adaptados al tópico del contenido que se quiera reconocer. De la evaluación del sistema de referencia y la optimización llevada a cabo es posible extraer una serie de conclusiones acerca del sistema de reconocimiento implementado aquí:

- La principal conclusión que se obtiene de la implementación de este sistema y la evaluación del mismo tras la optimización, es que cumpliendo con el objetivo fundamental, una adaptación del modelo del lenguaje al tópico o temática que se pretenda reconocer, supone una mejora relativa considerable (entre un 1,13% y un 8,73%, según el tópico) respecto al reconocimiento con un modelo del lenguaje elaborado con vocabulario genérico.

Además se ha calculado la perplejidad de los modelos adaptados al tópico y se produce una reducción de casi el 6% en el mejor de los casos para un tamaño de texto de prueba de unas 13.000 palabras, corroborándose así la reducción de la perplejidad presente en los trabajos de investigación de universidades prestigiosas como se hizo mención en la sección 5.1.

- Con los modelos del lenguaje empleados en los experimentos, y según los resultados de estos, se puede concluir que conforme se aumenta el volumen de los corpus de entrenamiento de los modelos y el vocabulario preciso para crear el diccionario, aumenta la tasa de acierto de palabra en el reconocimiento. Por ello, los modelos elegidos serán generados con unos 100 millones de palabras (tamaño total del corpus genérico disponible) y un vocabulario de 138.000 palabras (total de palabras recopiladas).
- Realizando una comparación entre el sistema de referencia (Capítulo 4) y el sistema optimizado (Capítulo 5), se observa una diferencia media en la tasa de acierto de palabra de unos 50 puntos absolutos. Como se explicó en los resultados, este fenómeno es debido a las discordancias entre una base de datos de entrenamiento, compuesta por voz leída telefónica, frente a una base de datos de test, empleada para testear la optimización y formada por voz espontánea. Los principales agentes ocasionantes de esto son: la velocidad del habla variable, los sonidos pobremente articulados, la aparición de correcciones, los falsos comienzos de frase y la falta de exactitud al seguir las reglas del lenguaje.

Además, los modelos del lenguaje se han generado con textos de habla formal de cada tópico, mientras que los contenidos reconocidos son del mismo tópico pero de habla espontánea.

- Se ha podido comprobar que en ficheros de audio en los que el locutor posee un *pitch* a partir de los 220 Hz., se produce una reducción de la tasa de acierto de palabra en el reconocimiento. Esto es lo ocurrido con los contenidos del tópico de moda, cuyas locutoras poseen un *pitch* promedio superior a lo mencionado, que afecta proporcionalmente en la tasa de acierto, reduciéndola (Figura 5.1).
- El último experimento realizado trataba de comparar el sistema optimizado con los sistemas de reconocimiento de voz de las compañías líderes: Web API Demonstration de Google (disponible online) y el Dictado Automático de Apple (integrado en Mac OS X). En los resultados obtenidos se ha podido comprobar que con la adaptación de modelos, se consiguen alcanzar tasas muy similares a las de la API de Google en los principales tópicos. Por otro lado, se ha logrado

superar, en el mejor de los casos, en 12 puntos absolutos al Dictado de Apple (véase Tabla 5.17).

- Por último, esta tecnología tiene numerosas aplicaciones, y se ha querido demostrar una de ellas, incluyendo el sistema desarrollado en una solución comercial que permite mejorar la búsqueda en contenidos audiovisuales (Capítulo 6).

7.2. Trabajo futuro

Al concluir este proyecto, habida cuenta del potencial del sistema implementado, aparecen distintas líneas de trabajo por las que se puede continuar investigando:

- Atendiendo al volumen de horas de voz del que se dispone en la base de datos actual, junto con las gráficas del Apéndice A (referencia), se ha podido demostrar que se consiguen mejores resultados incrementando el número de horas de voz de entrenamiento. De este modo, se puede deducir que el sistema implementado mejoraría si se ampliase la base de datos de entrenamiento, especialmente si se trata de habla espontánea y no habla leída como en la actualidad.
- Teniendo presente las entradas de voz con un *pitch* promedio de más de 220 Hz., sería conveniente aplicar modelos acústicos adaptados. En esta línea, ya se ha comenzado a trabajar, como puede verse en [19].
- Dado que los contenidos que se van a reconocer proceden directamente de Internet y no se puede asegurar su calidad, sería recomendable implementar nuevos métodos de supresión de ruido para poder así mejorar la fiabilidad del reconocimiento.
- Por último, según la mejora que suponen las Redes Neuronales Profundas (*Deep Neural Networks - DNN*) frente a los HMMs en el reconocimiento de habla natural, en torno al 30 % según la literatura. Por ello, sería adecuado investigar en esta línea y aplicar esta tecnología en la generación de los modelos acústicos. A pesar de que el software de HTK no permite el empleo de DNNs, existe otro software como Kaldi, que las considera como uno de los principales puntos de desarrollo y proporciona dicha funcionalidad.

Bibliografía

- [1] A. Sethy, P. G. Georgiou, and S. S. Narayanan, “Building topic specific language models from webdata using competitive models,” in *Proceedings of InterSpeech*, (Lisbon, Portugal), pp. 1293–1296, Oct. 2005.
- [2] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, *et al.*, *The HTK book*, vol. 2. Entropic Cambridge Research Laboratory Cambridge, 1997.
- [3] A. Lee, *The Julius book*. 2010.
- [4] P. Lamere, P. Kwok, W. Walker, E. B. Gouvêa, R. Singh, B. Raj, and P. Wolf, “Design of the cmu sphinx-4 decoder,” in *INTERSPEECH*, Citeseer, 2003.
- [5] G. Saon and J.-T. Chien, “Large-vocabulary continuous speech recognition systems: A look at some recent advances,” *Signal Processing Magazine, IEEE*, vol. 29, pp. 18–33, Nov 2012.
- [6] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, vol. 14. PTR Prentice Hall Englewood Cliffs, 1993.
- [7] G. Hinton, L. Deng, D. Yu, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, November 2012.
- [8] M. Adda-Decker and L. Lamel, “The use of lexica in automatic speech recognition,”
- [9] G. A. Fink, “n-gram models,” in *Markov Models for Pattern Recognition*, pp. 107–127, Springer, 2014.
- [10] J. González-Rodríguez, D. T. Toledano, and J. Ortega-García, “Voice biometrics,” in *Handbook of Biometrics*, pp. 151–170, Springer, 2008.
- [11] X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov models for speech recognition*, vol. 2004. Edinburgh university press Edinburgh, 1990.
- [12] S. Austin, R. Schwartz, and P. Placeway, “The forward-backward search algorithm,” in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pp. 697–700, IEEE, 1991.

- [13] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on*, vol. 13, no. 2, pp. 260–269, 1967.
- [14] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The annals of mathematical statistics*, pp. 164–171, 1970.
- [15] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech communication*, vol. 42, no. 1, pp. 93–108, 2004.
- [16] G. J. L. L. Lefevre, F., "Towards task-independent speech recognition," in *In: Proc. 2001 Internat. Conf. Acoust. Speech Signal Process., Salt Lake City, UT*, 2001.
- [17] K. Thadani, F. Biadsy, and D. Bikel, "On-the-fly topic adaptation for youtube video transcription," 2012.
- [18] C. García and D. Tapias, "La frecuencia fundamental de la voz y sus efectos en reconocimiento de habla continua," *División de Tecnología del Habla Telefónica Investigación y Desarrollo, Revista de Procesamiento de Lenguaje Natural*, vol. 26, pp. 163–168, 2000.
- [19] J. Antón Martín, "Desarrollo de un sistema de reconocimiento de habla natural independiente del locutor," Master's thesis, EPS-UAM, 2015.

Apéndice A

Experimentos elección de número de gaussianas

En este apéndice se muestran las tablas completas que han llevado a la elección del número óptimo de gaussianas:

Tabla A.1: Tasas de acierto de palabras (%WORD) en función del número de gaussianas empleadas durante la fase de entrenamiento del modelo acústico. Archivos de la lista Auto-Test con SNR 10-20 dB.

Nº gaussianas	% WORD	% ACC	% SENT	Nº gaussianas	% WORD	% ACC	% SENT
1	85,21	78,96	78,08	33	96,56	91,46	86,49
2	88,94	83,40	80,64	34	96,59	91,10	86,47
3	91,04	85,49	82,21	35	96,77	91,73	86,77
4	92,25	87,88	84,14	36	96,74	91,88	86,87
5	92,96	89,19	86,04	37	96,74	91,70	86,57
6	93,85	90,40	87,36	38	96,67	91,56	86,47
7	94,44	90,84	87,54	39	96,85	91,60	86,37
8	94,13	89,89	86,95	40	96,81	91,17	85,57
9	94,34	89,79	87,35	41	96,49	90,53	84,75
10	94,09	89,43	85,73	42	96,45	89,93	83,45
11	94,30	89,14	85,92	43	96,74	90,49	83,47
12	94,39	88,83	86,07	44	96,67	90,00	82,87
13	94,94	89,41	86,19	45	96,81	89,68	82,67
14	95,40	89,73	86,30	46	96,80	89,68	82,29
15	95,32	89,99	86,36	47	96,79	89,74	82,07
16	95,74	90,28	86,26	48	96,79	89,59	81,97
17	95,62	90,21	85,73	49	96,86	89,30	80,95
18	95,74	91,02	87,06	50	96,86	89,34	80,75
19	96,02	91,62	87,56	51	96,93	89,68	81,23
20	95,84	91,29	86,95	52	96,71	89,71	81,41
21	95,84	91,29	87,35	53	96,72	89,40	81,33
22	96,12	90,95	86,72	54	96,79	89,65	80,93
23	96,12	90,92	86,53	55	96,90	89,76	81,43
24	95,91	90,41	86,56	56	96,93	89,79	81,31
25	95,95	90,61	86,85	57	97,18	90,15	81,52
26	96,30	90,61	86,35	58	97,21	90,28	81,70
27	96,44	91,15	85,94	59	97,18	90,00	81,29
28	96,23	90,85	86,02	60	97,18	89,86	81,41
29	96,27	90,91	85,74	61	97,36	89,79	81,11
30	96,28	90,75	85,69	62	97,32	89,87	81,13
31	96,46	91,25	87,09	63	97,40	90,08	80,93
32	96,39	91,14	86,49	64	97,40	90,08	81,43

Tabla A.2: Tasas de acierto de palabra (%WORD) en función del número de gaussianas empleadas durante la fase de entrenamiento del modelo acústico. Archivos de la lista Auto-Test con SNR 20-30 dB.

Nº gaussianas	% WORD	% ACC	% SENT	Nº gaussianas	% WORD	% ACC	% SENT
1	87,05	84,45	60,12	33	96,77	94,77	78,70
2	90,97	88,13	66,47	34	96,75	94,72	78,90
3	92,06	89,11	66,53	35	96,89	95,00	79,40
4	93,30	90,36	68,94	36	96,92	95,03	79,50
5	93,55	90,79	70,42	37	96,98	94,94	79,10
6	94,21	91,36	72,21	38	97,00	95,02	79,50
7	94,77	91,94	73,54	39	96,97	95,02	79,80
8	94,91	92,64	74,55	40	96,97	95,05	80,00
9	94,94	92,57	74,60	41	96,95	94,99	80,28
10	95,00	92,80	75,23	42	97,14	95,27	80,28
11	95,34	92,78	75,43	43	97,03	95,17	79,60
12	95,51	92,99	75,68	44	97,11	95,22	79,60
13	95,73	93,38	76,18	45	97,28	95,39	80,00
14	95,87	93,50	76,68	46	97,31	95,32	79,98
15	95,99	93,91	77,48	47	97,42	95,34	79,68
16	96,18	94,04	77,78	48	97,48	95,57	80,08
17	95,82	93,68	77,48	49	97,50	95,56	79,98
18	96,06	94,17	77,80	50	97,45	95,52	79,98
19	96,03	94,23	78,40	51	97,43	95,51	80,38
20	96,12	94,20	78,20	52	97,32	95,38	80,28
21	96,32	94,28	77,88	53	97,42	95,43	80,48
22	96,55	94,50	78,00	54	97,40	95,43	80,28
23	96,46	94,52	78,58	55	97,34	95,40	80,18
24	96,47	94,41	78,50	56	97,37	95,37	80,58
25	96,50	94,47	78,30	57	97,29	95,13	79,88
26	96,61	94,57	78,48	58	97,28	95,18	79,68
27	96,51	94,50	78,58	59	97,23	95,21	79,68
28	96,58	94,59	78,70	60	97,20	95,09	79,28
29	96,77	94,72	79,40	61	97,17	95,15	79,38
30	96,81	94,84	79,10	62	97,25	95,11	79,50
31	96,72	94,70	78,60	63	97,39	95,45	80,08
32	96,84	94,75	78,90	64	97,43	95,52	80,08

Tabla A.3: Tasas de acierto de palabra (%WORD) en función del número de gaussianas empleadas durante la fase de entrenamiento del modelo acústico. Archivos de la lista Auto-Test con SNR 30-40 dB.

Nº gaussianas	% WORD	% ACC	% SENT	Nº gaussianas	% WORD	% ACC	% SENT
1	89,44	87,44	66,33	33	97,62	95,06	79,60
2	92,49	90,52	72,02	34	97,69	95,16	80,10
3	93,28	90,41	70,30	35	97,66	95,19	79,60
4	94,28	91,16	71,66	36	97,59	95,15	79,62
5	94,68	91,31	73,11	37	97,62	95,16	80,12
6	95,08	91,95	73,43	38	97,74	95,40	80,60
7	95,56	92,52	74,22	39	97,70	95,28	80,42
8	95,66	92,54	74,52	40	97,81	95,42	80,62
9	95,65	92,72	75,73	41	98,00	95,52	81,02
10	95,98	92,89	74,97	42	97,87	95,42	80,52
11	95,77	92,36	74,02	43	97,94	95,69	81,21
12	95,96	92,68	74,57	44	97,93	95,63	80,62
13	96,15	93,06	75,53	45	97,92	95,57	80,64
14	96,25	93,28	75,73	46	97,86	95,51	80,74
15	96,47	93,72	76,63	47	97,93	95,63	80,92
16	96,60	93,92	77,19	48	97,87	95,61	80,52
17	96,60	94,00	77,51	49	97,96	95,74	81,14
18	96,77	94,06	77,99	50	98,07	95,81	81,75
19	96,75	94,26	77,71	51	98,11	95,82	81,75
20	96,82	94,28	78,47	52	98,14	95,81	81,54
21	97,04	94,54	78,59	53	98,12	95,88	81,75
22	96,89	94,41	78,49	54	97,95	95,72	81,24
23	97,01	94,43	78,41	55	98,01	95,90	81,75
24	97,05	94,60	78,69	56	98,05	95,94	81,85
25	97,11	94,74	79,10	57	98,04	95,86	81,86
26	97,20	94,79	79,02	58	98,11	95,98	82,26
27	97,23	94,68	79,32	59	98,10	95,88	81,96
28	97,15	94,58	78,69	60	98,10	95,95	82,36
29	97,18	94,63	78,69	61	98,08	95,88	82,06
30	97,27	94,67	78,69	62	98,05	95,89	82,16
31	97,49	94,85	79,50	63	98,07	95,91	82,06
32	97,43	94,90	79,30	64	98,08	95,89	82,26

Tabla A.4: Tasas de acierto de palabra (%WORD) en función del número de gaussianas empleadas durante la fase de entrenamiento del modelo acústico. Archivos de la lista Auto-Test con SNR 40-99 dB.

Nº gaussianas	% WORD	% ACC	% SENT	Nº gaussianas	% WORD	% ACC	% SENT
1	88,78	85,69	78,08	33	96,72	95,69	86,49
2	91,59	89,44	80,64	34	96,67	95,47	86,47
3	92,31	90,39	82,21	35	96,62	95,55	86,77
4	93,47	91,76	84,14	36	96,82	95,69	86,87
5	93,98	92,36	86,04	37	96,99	95,87	86,57
6	94,52	93,15	87,36	38	96,96	95,85	86,47
7	94,96	93,64	87,54	39	96,86	95,72	86,37
8	95,11	93,66	86,95	40	96,84	95,77	85,57
9	95,27	93,82	87,35	41	96,95	95,83	84,75
10	95,26	93,79	85,73	42	97,05	95,92	83,45
11	95,52	94,02	85,92	43	96,91	95,62	83,47
12	95,40	93,83	86,07	44	97,08	95,81	82,87
13	95,71	94,07	86,19	45	97,1	95,89	82,67
14	95,87	94,29	86,30	46	97,05	95,87	82,29
15	95,88	94,29	86,36	47	97,06	95,94	82,07
16	95,94	94,38	86,26	48	97,1	96,08	81,97
17	96,14	94,77	85,73	49	97,2	96,07	80,95
18	95,99	94,58	87,06	50	97,15	96,05	80,75
19	96,01	94,49	87,56	51	97,22	96,05	81,23
20	96,10	94,6	86,95	52	97,15	96,03	81,41
21	96,16	95,09	87,35	53	97,2	96,03	81,33
22	96,22	95,11	86,72	54	97,22	96,05	80,93
23	96,22	95,15	86,53	55	97,24	96,03	81,43
24	96,32	95,28	86,56	56	97,24	96,10	81,31
25	96,30	95,23	86,85	57	97,18	95,98	81,52
26	96,52	95,54	86,35	58	97,25	96,09	81,70
27	96,54	95,45	85,94	59	97,20	96,11	81,29
28	96,44	95,29	86,02	60	97,15	96,00	81,41
29	96,49	95,31	85,74	61	97,20	96,08	81,11
30	96,41	95,24	85,69	62	97,18	96,01	81,13
31	96,46	95,42	87,09	63	97,26	96,16	80,93
32	96,67	95,64	86,49	64	97,23	96,18	81,43

Tabla A.5: Tasas de acierto de palabra (%WORD) en función del número de gaussianas empleadas durante la fase de entrenamiento del modelo acústico. Archivos de la lista Test con SNR 30-40 dB.

Nº gaussianas	% WORD	% ACC	% SENT	Nº gaussianas	% WORD	% ACC	% SENT
1	89,58	87,61	68,97	33	95,32	91,57	71,41
2	91,86	89,48	70,27	34	95,25	91,44	71,08
3	93,13	90,02	70,47	35	95,06	91,25	70,51
4	94,29	90,78	70,80	36	95,12	91,21	70,15
5	94,48	90,95	71,14	37	94,96	90,81	69,92
6	94,82	91,50	72,57	38	94,84	90,50	68,98
7	94,94	91,74	72,14	39	94,83	90,45	68,71
8	95,08	92,07	73,10	40	94,93	90,67	68,88
9	95,15	92,03	72,80	41	94,72	90,30	69,15
10	95,15	91,81	72,47	42	94,67	90,17	69,01
11	95,27	91,70	71,23	43	94,57	90,01	68,81
12	95,32	91,61	71,36	44	94,54	89,87	68,37
13	95,41	91,96	71,51	45	94,47	89,82	68,20
14	95,53	91,98	71,84	46	94,39	89,66	67,74
15	95,49	92,23	73,07	47	94,30	89,55	68,00
16	95,46	92,27	73,47	48	94,16	89,37	67,37
17	95,45	92,27	72,75	49	94,00	89,10	67,14
18	95,47	92,22	73,37	50	93,89	88,90	66,97
19	95,33	92,03	72,87	51	93,78	88,65	67,04
20	95,57	92,30	72,72	52	93,88	88,81	66,83
21	95,65	92,54	73,19	53	93,80	88,56	66,73
22	95,50	92,27	73,12	54	93,73	88,14	66,16
23	95,38	92,22	73,35	55	93,64	88,12	65,60
24	95,43	92,17	73,17	56	93,43	87,50	65,13
25	95,47	92,24	73,32	57	93,46	87,63	64,93
26	95,47	92,36	73,32	58	93,22	87,20	64,23
27	95,54	92,33	73,29	59	93,08	86,86	64,29
28	95,45	92,13	72,89	60	92,91	86,75	63,59
29	95,36	91,81	72,14	61	92,79	86,63	63,49
30	95,30	91,82	72,19	62	92,58	86,35	63,05
31	95,31	91,75	72,09	63	92,50	86,16	63,25
32	95,11	91,33	71,26	64	92,44	86,03	62,95

Tabla A.6: Tasas de acierto de palabra (%WORD) en función del número de gaussianas empleadas durante la fase de entrenamiento del modelo acústico. Archivos de la lista Test con archivos ruidosos.

Nº gaussianas	% WORD	% ACC	% SENT	Nº gaussianas	% WORD	% ACC	% SENT
1	81,00	72,14	61,43	33	87,68	77,76	69,88
2	85,44	79,37	70,34	34	87,83	77,55	69,18
3	87,36	80,98	71,94	35	87,83	77,41	68,98
4	88,71	82,82	75,25	36	87,73	77,31	69,38
5	89,36	83,93	75,98	37	87,83	76,83	69,28
6	90,20	85,15	77,68	38	87,6	76,38	69,18
7	90,14	84,63	76,95	39	87,44	76,51	68,47
8	90,45	84,44	76,95	40	87,04	75,80	67,77
9	90,72	84,41	76,75	41	87,03	74,90	67,30
10	91,06	85,15	76,75	42	86,84	74,93	67,50
11	90,74	84,39	76,15	43	86,78	75,07	67,23
12	90,88	84,78	76,15	44	86,04	73,83	66,23
13	91,30	84,97	77,15	45	85,86	73,96	66,03
14	91,01	84,07	76,05	46	85,49	73,27	65,63
15	90,88	83,89	76,75	47	85,44	73,03	65,13
16	90,88	83,47	76,35	48	85,28	72,48	64,96
17	90,85	83,47	75,75	49	85,28	71,91	63,90
18	90,98	83,25	75,65	50	84,70	71,27	63,70
19	90,90	83,65	75,25	51	84,91	70,83	62,90
20	90,36	83,12	74,82	52	84,57	69,93	62,60
21	90,12	82,51	74,22	53	84,31	69,59	61,90
22	89,88	81,93	73,52	54	84,12	69,27	62,40
23	89,95	81,32	73,15	55	84,52	69,64	62,40
24	89,83	81,29	72,72	56	83,99	68,93	61,30
25	89,51	80,37	72,02	57	83,36	68,19	61,00
26	88,94	79,20	71,41	58	82,99	67,06	59,40
27	88,81	79,11	71,39	59	82,46	66,38	58,70
28	89,05	79,95	70,88	60	82,93	67,08	59,26
29	88,34	78,68	70,68	61	82,77	65,81	58,76
30	88,48	78,84	70,21	62	82,80	66,63	58,76
31	88,65	79,29	70,18	63	82,85	66,74	59,06
32	88,19	78,71	70,11	64	82,17	65,32	57,80

Apéndice B

Experimentos elección de número de ficheros de entrenamiento

En este apéndice se muestran las gráficas y tablas completas de la elección del número de archivos de entrenamiento:

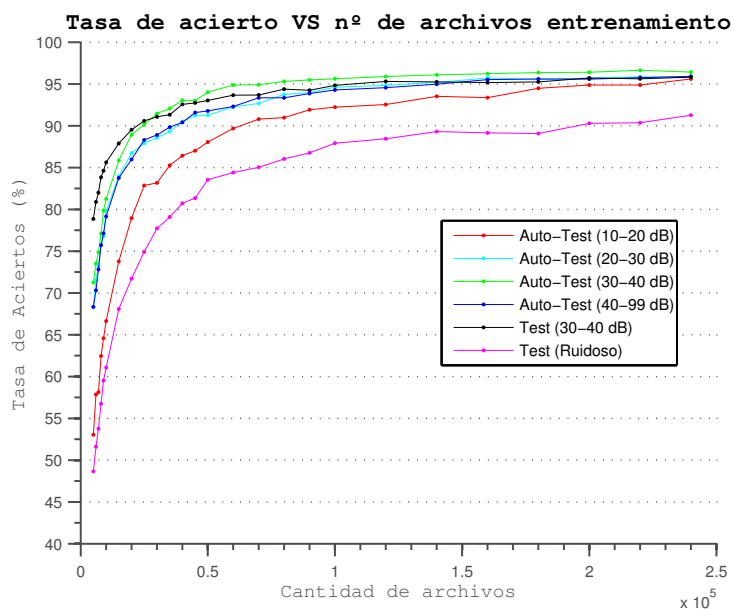


Figura B.1: Gráfica en la que se puede observar la evolución de la tasa de acierto (%WORD) en función del número de archivos con el que se ha entrenado el modelo acústico.

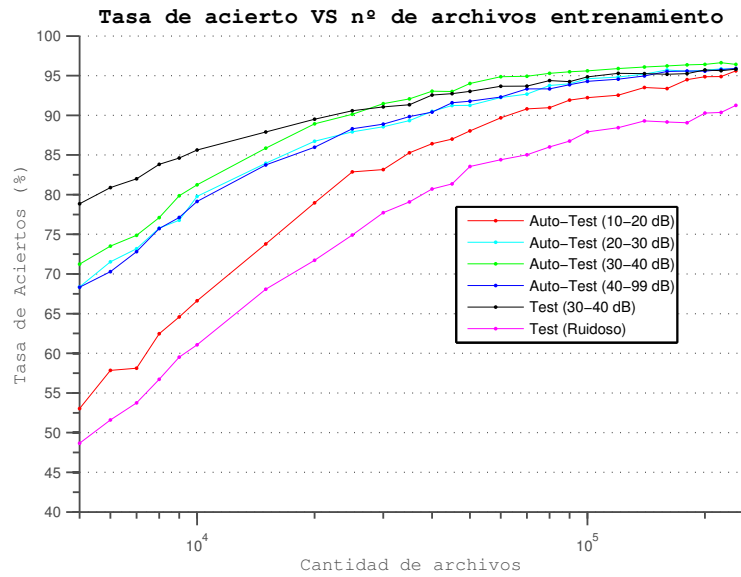


Figura B.2: Gráfica en la que se puede observar la evolución de la tasa de acierto (%WORD) en función del número de archivos con el que se ha entrenado el modelo acústico. Se muestra en escala semilogarítmica para poder apreciar mejor la tendencia de la gráfica.

Tabla B.1: Tasa de acierto de palabras (%WORD) frente a número de archivos de entrenamiento por grupos de archivos de prueba.

Nº Archivos	Auto 10-20	Auto 20-30	Auto 30-40	Auto 40-99	Test 30-40	Test Ruido
5.000	53,03 ± 2.66	68,36 ± 2.21	71,24 ± 2.18	68,33 ± 2.29	78,85 ± 1.94	48,66 ± 2.34
6.000	57,85 ± 2.71	71,53 ± 2.18	73,52 ± 2.19	70,30 ± 2.31	80,89 ± 1.84	51,59 ± 2.38
7.000	58,13 ± 2.71	73,19 ± 2.18	74,87 ± 2.16	72,82 ± 2.26	82,00 ± 1.87	53,75 ± 2.37
8.000	62,47 ± 2.65	75,78 ± 2.12	77,11 ± 2.13	75,72 ± 2.27	83,83 ± 1.80	56,73 ± 2.39
9.000	64,58 ± 2.67	76,77 ± 2.12	79,87 ± 2.04	77,13 ± 2.24	84,62 ± 1.73	59,51 ± 2.39
10.000	66,63 ± 2.62	79,78 ± 2.05	81,24 ± 2.02	79,15 ± 2.21	85,62 ± 1.72	61,07 ± 2.41
15.000	73,78 ± 2.50	83,95 ± 1.91	85,85 ± 1.77	83,75 ± 2.06	87,90 ± 1.59	68,07 ± 2.34
20.000	78,96 ± 2.37	86,73 ± 1.77	88,93 ± 1.62	85,98 ± 1.95	89,51 ± 1.55	71,72 ± 2.34
25.000	82,86 ± 2.24	87,92 ± 1.72	90,13 ± 1.46	88,30 ± 1.79	90,57 ± 1.47	74,91 ± 2.24
30.000	83,16 ± 2.20	88,56 ± 1.66	91,47 ± 1.37	88,89 ± 1.71	91,06 ± 1.39	77,72 ± 2.17
35.000	85,27 ± 2.09	89,33 ± 1.60	92,07 ± 1.35	89,83 ± 1.66	91,33 ± 1.31	79,09 ± 2.08
40.000	86,43 ± 1.93	90,50 ± 1.54	93,05 ± 1.24	90,41 ± 1.63	92,56 ± 1.28	80,72 ± 2.06
45.000	87,01 ± 1.97	91,24 ± 1.35	93,00 ± 1.26	91,58 ± 1.52	92,74 ± 1.22	81,35 ± 2.03
50.000	88,04 ± 1.88	91,26 ± 1.38	94,01 ± 1.15	91,78 ± 1.49	93,03 ± 1.24	83,54 ± 1.90
60.000	89,68 ± 1.81	92,24 ± 1.24	94,87 ± 1.02	92,30 ± 1.36	93,66 ± 1.23	84,41 ± 1.86
70.000	90,80 ± 1.68	92,67 ± 1.22	94,92 ± 0.98	93,34 ± 1.22	93,68 ± 1.10	85,02 ± 1.85
80.000	90,97 ± 1.66	93,76 ± 1.14	95,30 ± 1.02	93,34 ± 1.24	94,38 ± 1.12	86,02 ± 1.75
90.000	91,91 ± 1.56	93,98 ± 1.08	95,48 ± 0.89	93,86 ± 1.17	94,25 ± 1.10	86,75 ± 1.72
100.000	92,22 ± 1.51	94,62 ± 1.01	95,61 ± 0.85	94,29 ± 1.14	94,84 ± 0.97	87,91 ± 1.67
120.000	92,54 ± 1.50	94,83 ± 0.95	95,89 ± 0.79	94,56 ± 1.04	95,30 ± 0.97	88,45 ± 1.60
140.000	93,51 ± 1.36	95,19 ± 0.92	96,08 ± 0.80	94,97 ± 0.88	95,25 ± 0.96	89,30 ± 1.55
160.000	93,37 ± 1.29	95,72 ± 0.83	96,23 ± 0.82	95,52 ± 0.83	95,17 ± 0.97	89,15 ± 1.57
180.000	94,49 ± 1.24	95,54 ± 0.79	96,36 ± 0.72	95,59 ± 0.83	95,25 ± 0.99	89,07 ± 1.59
200.000	94,87 ± 1.21	95,69 ± 0.72	96,41 ± 0.72	95,57 ± 0.77	95,72 ± 0.94	90,28 ± 1.48
220.000	94,88 ± 1.19	95,85 ± 0.70	96,63 ± 0.73	95,75 ± 0.71	95,62 ± 0.86	90,36 ± 1.48
240.000	95,60 ± 1.08	95,82 ± 0.65	96,42 ± 0.71	95,92 ± 0.63	95,81 ± 0.83	91,26 ± 1.43

Apéndice C

Resultados exhaustivos de la optimización

Este apéndice contiene los resultados tabulados de todos los experimentos realizados en el Capítulo 5.

Tabla C.1: Resultados Experimento 1. Tasas de acierto de palabra (%WORD) empleando modelos del lenguaje acotados (entrenados con corpus formado con vocabulario únicamente empleado en los ficheros de test) para cada tópico.

Tópico	Fichero	%WORD
Economía	eco01.wav	44,09
	eco02.wav	64,66
	eco03.wav	55,31
	Promedio	54,54
Deporte	dep01.wav	52,55
	dep02.wav	57,34
	dep03.wav	57,25
	dep04.wav	47,47
	dep05.wav	51,51
	dep06.wav	64,85
	dep07.wav	67,93
	dep08.wav	61,66
	Promedio	58,36
Moda	moda01.wav	32,92
	moda02.wav	22,94
	moda03.wav	37,08
	moda04.wav	39,63
	moda05.wav	38,20
	Promedio	33,98

Tabla C.2: Resultados Experimento 2. Tasas de acierto de palabra (%WORD) variando el corpus de entrenamiento del modelo del lenguaje genérico, para cada tópico. La variación del corpus está comprendida entre 200.000 palabras (Gen 200K) y 100 millones de palabras (Gen 100M).

Tópico	Fichero	Gen 200K	Gen 400K	Gen 1M	Gen 2M	Gen 4M	Gen 10M	Gen 20M	Gen 40M	Gen 100M	
Economía	eco01.wav	23,28	23,15	24,22	24,13	25,14	25,14	25,61	25,68	26,10	
	eco02.wav	34,79	35,39	36,54	37,51	37,51	38,10	38,79	39,14	39,67	
	eco03.wav	32,26	32,37	32,85	33,38	34,05	34,53	34,69	34,80	35,25	
	Promedio	29,91	30,10	31,03	31,49	32,05	32,39	32,85	33,03	33,50	
	Deporte	dep01.wav	28,47	28,71	29,56	29,44	29,93	30,54	31,27	32,26	32,73
		dep02.wav	26,59	26,59	27,98	28,47	28,87	28,37	28,27	28,47	28,17
		dep03.wav	23,14	23,33	26,18	26,96	27,55	28,04	29,31	29,31	30,49
		dep04.wav	23,19	24,22	25,05	25,37	26,65	26,65	27,29	26,84	27,55
dep05.wav		24,61	24,91	26,16	27,92	27,77	28,43	29,39	29,76	29,17	
dep06.wav		32,40	32,78	33,55	35,86	36,19	36,63	37,73	37,62	38,22	
dep07.wav		39,14	39,36	40,21	41,43	40,86	42,00	41,93	43,64	43,07	
dep08.wav		33,33	33,78	34,77	35,84	36,62	37,03	37,11	37,81	37,36	
Moda	Promedio	29,62	30,01	31,15	32,25	32,68	33,10	33,64	34,06	34,12	
	moda01.wav	16,35	14,10	14,42	13,94	13,83	13,89	13,99	14,21	14,32	
	moda02.wav	12,50	12,76	13,05	12,90	12,02	12,17	12,57	12,35	12,43	
	moda03.wav	19,33	19,08	19,42	19,38	19,80	19,76	19,50	19,20	19,72	
	moda04.wav	17,33	16,99	17,09	17,53	17,09	17,64	17,53	17,30	17,36	
	moda05.wav	16,53	15,53	16,21	16,63	16,79	16,47	16,74	16,58	16,37	
	Promedio	15,85	15,74	16,06	16,12	15,90	16,01	16,09	15,93	16,05	

Tabla C.3: Resultados Experimento 2. Tasas de acierto de palabra (%WORD) variando el corpus de entrenamiento del modelo del lenguaje genérico, para 'Auto-test' y 'Test'. 'Auto-test' hace referencia al conjunto de 1.000 ficheros de la base de datos de entrenamiento. 'Test' hace referencia al conjunto de 1.000 ficheros, extraídos antes del entrenamiento. La variación del corpus está comprendida entre 200.000 palabras (Gen.200K) y 100 millones de palabras (Gen.100M).

	Gen 200K	Gen 400K	Gen 1M	Gen 2M	Gen 4M	Gen 10M	Gen 20M	Gen 40M	Gen 100M
Auto-test	69,01	70,55	71,94	73,03	73,69	75,14	75,75	76,94	77,57
Test	62,95	65,27	66,86	68,24	68,60	69,90	70,45	71,85	72,65

Tabla C.4: Resultados Experimento 3. Tasas de acierto de palabra (%WORD) variando el corpus de entrenamiento del modelo del lenguaje genérico y mezclándolo con un corpus de cada tópico, de 200.000 palabras (Text). La variación del corpus genérico está comprendida entre 200.000 palabras (Gen 200K) y 100 millones de palabras (Gen 100M).

Tópico	Fichero	200K+Text	400K+Text	1M+Text	2M+Text	4M+Text	10M+Text	20M+Text	40M+Text	100M+Text
Economía	eco01.wav	23,28	23,15	24,22	24,13	25,54	25,21	25,34	25,50	26,35
	eco02.wav	34,79	35,39	36,54	37,51	37,99	38,45	39,14	39,12	39,58
	eco03.wav	32,26	32,37	32,85	33,38	33,88	34,30	34,64	34,55	35,34
Deporte	Promedio	29,91	30,10	31,03	31,49	32,31	32,47	32,86	32,88	33,58
	ddep01.wav	28,47	28,71	29,56	29,44	30,78	30,66	31,51	32,85	32,60
	ddep02.wav	26,59	26,59	27,98	28,47	29,46	28,97	28,77	28,27	27,98
	ddep03.wav	23,14	23,33	26,18	26,96	29,51	29,71	30,10	30,69	30,49
	ddep04.wav	23,19	24,22	25,05	25,37	28,12	26,46	27,23	27,29	27,29
	ddep05.wav	24,61	24,91	26,16	27,92	28,43	29,02	29,83	29,17	29,46
	ddep06.wav	32,40	32,78	33,55	35,86	37,34	38,00	38,17	38,39	39,48
	ddep07.wav	39,14	39,36	40,21	41,43	42,43	42,93	43,14	44,07	43,71
Moda	ddep08.wav	33,33	33,78	34,77	35,84	38,22	37,68	37,81	38,55	38,22
	Promedio	29,62	30,01	31,15	32,25	33,97	33,83	34,18	34,53	34,56
	moda01.wav	13,35	14,10	14,42	13,94	14,16	13,51	14,26	14,37	14,42
	moda02.wav	12,50	12,76	13,05	12,90	12,13	12,13	12,39	12,43	12,39
	moda03.wav	19,33	19,08	19,42	19,38	20,36	19,67	19,29	19,50	19,67
	moda04.wav	17,33	16,99	17,09	17,53	17,77	17,77	17,80	17,50	17,57
	moda05.wav	16,53	15,53	16,21	16,63	17,00	16,42	16,21	16,16	16,63
	Promedio	15,85	15,74	16,06	16,12	16,29	15,96	16,03	16,01	16,14

Tabla C.5: Resultados Experimento 3. Tasas de acierto de palabra (%WORD) empleando el corpus de entrenamiento del modelo del lenguaje genérico y mezclándolo con un corpus de cada tópico, de 200K palabras (100M+Text), o mezclando con un corpus de cada tópico, de 20M palabras (100M+BigText). El corpus del modelo del lenguaje genérico ha sido elegido de 100M palabras al ser el que proporcionaba mayor tasa en los experimentos anteriores.

Tópico	Fichero	100M+Text	100M+BigText
Economía	eco01.wav	26,35	26,19
	eco02.wav	39,58	40,22
	eco03.wav	35,34	35,39
	Promedio	33,58	33,76
Deporte	dep01.wav	32,60	33,45
	dep02.wav	27,98	28,97
	dep03.wav	30,49	30,69
	dep04.wav	27,29	28,51
	dep05.wav	29,46	29,24
	dep06.wav	39,48	40,20
	dep07.wav	43,71	44,43
	dep08.wav	38,22	39,82
	Promedio	34,56	35,41
Moda	moda01.wav	14,42	14,42
	moda02.wav	12,39	12,35
	moda03.wav	19,67	19,72
	moda04.wav	17,57	17,53
	moda05.wav	16,63	16,63
	Promedio	16,14	16,13

Tabla C.6: Resultados Experimento 4. Tasas de acierto de palabra (%WORD) variando el tamaño del diccionario genérico, para 'Auto-test' y 'Test'. 'Auto-test' hace referencia al conjunto de 1.000 ficheros de la base de datos de entrenamiento. 'Test' hace referencia al conjunto de 1.000 ficheros, con las mismas características que los de la base de datos de entrenamiento, pero extraídos antes del entrenamiento. La variación del diccionario está comprendida entre 10.000 palabras (Dic 10K) y 130.000 de palabras (Dic 130K).

Ficheros	Dic 10K	Dic 20K	Dic 30K	Dic 40K	Dic 50K	Dic 60K	Dic 70K
Auto-test	69,72	74,04	76,21	77,57	78,60	79,15	79,46
Test	65,44	69,78	71,54	72,78	73,53	74,03	74,16
Ficheros	Dic 80K	Dic 90K	Dic 100K	Dic 110K	Dic 120K	Dic 130K	
Auto-test	79,68	80,10	80,34	80,44	80,47	80,54	
Test	74,42	74,52	74,79	74,92	74,95	75,01	

Tabla C.7: Resultados Experimento 5. Tasas de acierto de palabra (%WORD) variando el tamaño del diccionario genérico (60.000 palabras en este caso) y agregando gradualmente palabras de cada tópic, con el fin de cuantificar la penalización al quitar las palabras genéricas y la ganancia al introducir palabras del tópic en cuestión. La eliminación será en función de la frecuencia de las palabras en el idioma. *Por ejemplo:* Se eliminan 1.000 palabras del diccionario de 60.000 (Dic 59K) y se añaden 1.000 propias del tópic (Dic 59K+1K). En último lugar se eliminan N palabras, que corresponde a la cantidad total de palabras específicas de cada tópic (véase Tabla 5.1), y se añaden N_{esp} palabras, que hace referencia a las palabras concretas del tópic en cuestión.

Tópic	Fichero	Dic 59K	Dic 59K+1K	Dic 58K	Dic 58K+2K	Dic 57K	Dic 57K+3K	Dic 60K-N	Dic 60K-N+N _{esp}
Economía	eco01.wav	25,47	25,45	25,47	25,32	-	-	25,50	25,34
	eco02.wav	38,82	3861	38,82	39,10	-	-	38,82	39,14
	eco03.wav	34,58	34,53	34,58	34,63	-	-	34,61	34,64
Deporte	Promedio	32,78	32,68	32,78	38,85	-	-	32,80	32,86
	dep01.wav	31,63	31,27	31,63	31,27	31,27	30,78	31,51	31,51
	dep02.wav	28,17	29,17	28,17	29,27	28,47	29,37	28,17	28,77
	dep03.wav	29,12	29,22	29,41	29,22	29,12	29,02	29,51	30,10
	dep04.wav	27,16	27,55	27,29	27,23	27,16	26,27	27,03	27,23
	dep05.wav	28,73	29,17	28,88	29,39	29,02	29,24	29,24	29,83
	dep06.wav	37,95	38,11	37,62	37,62	37,45	38,33	37,89	38,17
	dep07.wav	42,14	42,50	42,14	43,00	42,00	42,64	42,14	43,14
Moda	dep08.wav	36,90	37,32	36,82	37,36	37,03	37,40	37,10	37,81
	Promedio	33,56	33,90	33,55	33,88	33,53	33,76	33,69	34,18
	moda01.wav	14,05	14,05	14,10	14,42	14,10	14,37	14,05	14,26
	moda02.wav	12,35	12,35	12,54	12,21	12,35	12,21	12,35	12,39
	moda03.wav	19,42	19,67	19,50	19,20	19,46	19,16	19,38	19,29
	moda04.wav	17,43	17,70	17,50	17,80	17,57	18,07	17,50	17,80
	moda05.wav	16,42	16,47	16,63	16,26	16,74	16,63	16,68	16,21
	Promedio	15,96	16,08	16,07	16,01	16,06	16,12	16,01	16,03

Tabla C.8: Resultados Experimento 6. Tasas de acierto de palabra (%WORD) empleando mezclas ponderadas de los modelos del lenguaje. Estos modelos serán generados de hasta 2-gramas. Se varía la ponderación del modelo del lenguaje genérico, para cada tópico. La variación de los pesos está comprendida entre 20% genérico/ 80% específico (Gen 0,2) y 80% genérico/ 20% específico (Gen 0,8).

Tópico	Fichero	Gen 0,2	Gen 0,3	Gen 0,4	Gen 0,5	Gen 0,6	Gen 0,7	Gen 0,8
Economía	eco01.wav	24,56	24,98	25,25	25,83	25,85	26,03	26,44
	eco02.wav	40,59	40,64	41,16	41,14	40,93	40,52	40,27
	eco03.wav	34,86	34,89	35,14	35,11	35,22	35,03	35,00
Deporte	Promedio	33,15	33,33	33,68	33,88	33,84	33,71	33,76
	dep01.wav	34,31	34,18	33,45	33,21	32,97	33,45	33,45
	dep02.wav	31,25	31,25	30,85	30,95	30,36	29,56	27,98
	dep03.wav	34,02	34,02	32,55	31,86	31,18	30,59	30,78
	dep04.wav	29,15	29,79	29,60	29,02	29,02	28,57	27,99
	dep05.wav	31,23	31,01	31,01	31,30	30,86	30,12	29,76
	dep06.wav	42,12	41,95	41,46	41,41	41,90	41,46	40,58
	dep07.wav	46,00	45,79	46,07	44,93	44,79	44,64	44,43
dep08.wav	41,26	41,54	41,34	41,13	40,70	40,07	39,90	
Moda	Promedio	37,10	37,16	36,83	36,52	36,31	35,85	35,40
	moda01.wav	15,28	15,07	14,91	14,26	13,94	13,46	14,21
	moda02.wav	12,87	12,43	12,46	13,01	12,98	12,90	13,12
	moda03.wav	20,74	20,70	20,83	20,92	20,87	20,74	20,49
	moda04.wav	18,78	18,48	18,58	18,48	18,38	18,48	18,04
	moda05.wav	17,79	17,79	17,40	17,52	17,10	17,52	17,31
Promedio	17,09	16,87	16,85	16,88	16,72	16,69	16,67	

Tabla C.9: Resultados Experimento 7. Tasas de acierto de palabra (%WORD) empleando mezclas ponderadas de los modelos del lenguaje. Estos modelos serán generados de hasta 3-gramas (haciendo uso de ellos en la 2ª pasada de reconocimiento). Se varía la ponderación del modelo del lenguaje genérico, para cada tópic. La variación de los pesos está comprendida entre 20% genérico/ 80% específico (Gen 0,2) y 80% genérico/ 20% específico (Gen 0,8).

Tópico	Fichero	Gen 0,2	Gen 0,3	Gen 0,4	Gen 0,5	Gen 0,6	Gen 0,7	Gen 0,8
Economía	eco01.wav	26,50	27,17	27,58	27,87	28,22	27,71	28,11
	eco02.wav	43,70	44,27	44,39	44,57	44,16	44,48	43,74
	eco03.wav	36,90	37,29	37,77	38,18	38,46	38,04	37,93
Deporte	Promedio	35,53	36,09	36,41	36,70	36,76	36,57	36,42
	dep01.wav	36,01	36,50	36,01	36,13	36,13	35,77	36,86
	dep02.wav	33,53	33,53	33,13	34,13	33,63	34,13	33,33
	dep03.wav	37,25	36,27	35,10	36,08	35,78	34,31	34,71
	dep04.wav	32,48	33,50	33,25	33,12	32,80	31,84	30,81
	dep05.wav	34,39	33,36	34,02	33,65	33,58	32,99	33,21
	dep06.wav	44,10	44,48	43,71	43,11	43,05	42,67	41,90
	dep07.wav	48,21	48,79	48,79	47,86	47,57	47,00	46,93
Moda	dep08.wav	43,72	43,64	43,76	43,47	42,98	42,49	41,79
	Promedio	39,65	39,73	39,50	39,36	39,08	38,53	38,18
	moda01.wav	17,53	18,02	17,16	16,14	16,30	15,23	15,28
	moda02.wav	14,60	14,38	14,34	14,26	14,08	13,90	13,79
	moda03.wav	22,54	22,58	22,84	22,54	22,28	22,28	21,56
	moda04.wav	20,88	21,11	20,24	20,30	19,70	19,39	20,07
	moda05.wav	20,51	20,09	20,30	19,52	19,83	19,36	19,62
	Promedio	19,17	19,20	18,92	18,57	18,40	18,03	18,08

Tabla C.10: Resultados Experimento 8. Tasas de acierto de palabra (%WORD) empleando mezclas ponderadas de los modelos del lenguaje. Estos modelos serán generados de hasta 5-gramas (haciendo uso de ellos en la 2ª pasada de reconocimiento). Se varía la ponderación del modelo del lenguaje genérico, para cada tópico. La variación de los pesos está comprendida entre 20% genérico / 80% específico (Gen 0,2) y 80% genérico / 20% específico (Gen 0,8).

Tópico	Fichero	Gen 0,2	Gen 0,3	Gen 0,4	Gen 0,5	Gen 0,6	Gen 0,7	Gen 0,8
Economía	eco01.wav	26,12	26,99	27,69	27,66	28,27	27,84	28,16
	eco02.wav	44,62	44,57	44,71	45,12	44,91	45,17	44,52
	eco03.wav	36,73	37,54	37,60	38,04	37,99	38,32	37,88
Deporte	Promedio	35,67	36,20	36,51	36,78	36,91	36,94	36,70
	dep01.wav	37,10	37,59	37,96	37,83	36,98	36,86	36,62
	dep02.wav	35,52	33,73	33,73	33,93	32,94	32,94	30,36
	dep03.wav	37,65	37,84	37,35	37,45	36,96	36,76	36,76
	dep04.wav	33,31	32,93	33,18	32,35	32,67	32,09	31,01
	dep05.wav	35,12	34,83	35,34	36,08	36,22	35,34	34,24
	dep06.wav	44,59	44,15	43,60	43,77	43,88	43,00	41,68
	dep07.wav	49,64	49,29	50,00	48,86	49,14	48,29	48,07
dep08.wav	43,23	44,05	43,68	43,60	42,45	42,36	42,12	
Moda	Promedio	40,29	40,16	40,16	40,02	39,70	39,22	38,41
	moda01.wav	17,43	16,57	16,57	16,51	16,57	15,34	15,39
	moda02.wav	14,82	14,45	14,49	14,15	14,08	13,60	13,71
	moda03.wav	22,20	22,16	22,24	22,07	22,20	21,73	21,81
	moda04.wav	21,55	20,84	20,34	20,10	20,32	19,81	19,36
	moda05.wav	20,36	21,25	19,78	19,52	19,41	19,46	19,73
Promedio	19,28	19,02	18,68	18,46	18,51	18,01	17,97	

Tabla C.11: Resultados Experimento 9. Tasas de acierto de palabra (%WORD) empleando mezclas ponderadas de los modelos del lenguaje, generando un modelo del lenguaje completo que englobe todos los tópicos. Se varía la ponderación de los distintos modelos del lenguaje. La variación de los pesos sigue un determinado formato en porcentaje en el siguiente orden: Genérico-Deporte-Moda-Economía. *Por ejemplo:* 50-20-20-10 hace referencia a un 50% genérico, 20% deporte y moda, y un 10% economía. En esta misma tabla se presenta el resultado del reconocimiento con un modelo genérico formado por el vocabulario total de 138.000 palabras (Gen 138K), que incluyen todas las específicas de los tres tópicos.

Tópico	Fichero	Gen 138K	50-20-20-10	40-25-25-10	30-30-30-10	25-35-35-5	15-40-40-5
Economía	eco01.wav	28,20	28,29	28,54	27,89	27,82	27,51
	eco02.wav	43,53	44,85	44,82	44,02	43,76	42,89
	eco03.wav	38,04	38,46	38,66	38,38	38,16	37,35
	Promedio	36,42	37,03	37,17	36,57	36,39	35,74
	dep01.wav	36,25	35,77	36,62	36,37	37,35	36,74
	dep02.wav	28,87	35,52	34,33	35,71	35,22	34,33
	dep03.wav	33,73	32,35	35,98	34,22	33,43	33,92
	dep04.wav	28,38	30,17	30,69	32,09	32,54	33,12
Deporte	dep05.wav	32,99	34,53	33,87	33,06	33,43	33,36
	dep06.wav	39,92	42,83	42,56	42,34	42,78	42,56
	dep07.wav	46,21	48,29	48,50	47,86	48,07	47,64
	dep08.wav	39,98	42,08	42,24	42,57	41,67	41,75
	Promedio	36,51	38,53	38,82	38,82	38,79	38,71
	moda01.wav	15,17	16,14	16,35	16,68	17,21	17,27
	moda02.wav	13,38	13,60	14,08	14,45	14,34	14,52
	moda03.wav	20,79	21,77	21,94	21,94	22,37	21,81
Moda	moda04.wav	18,82	19,73	20,00	20,30	20,27	20,27
	moda05.wav	19,46	18,89	19,31	18,78	19,25	19,10
	Promedio	17,48	18,02	18,33	18,46	18,67	18,59

Tabla C.12: Resultados Experimento 10. Comparativa de tasas de reconocimiento (%WORD) entre el sistema implementado optimizado y los sistemas de las grandes empresas del mercado, para cada t3pico.

T3pico	Fichero	Recog. Sigma	API Google	Dict. Auto Apple
Economía	eco01.wav	27,37	23,89	16,00
	eco02.wav	50,23	43,67	44,73
	eco03.wav	40,70	54,55	43,27
	Promedio	39,23	39,67	33,94
Deporte	dep01.wav	41,24	41,48	31,51
	dep02.wav	37,70	39,09	26,19
	dep03.wav	40,88	29,90	29,12
	dep04.wav	32,61	31,45	23,13
	dep05.wav	40,12	44,09	30,27
	dep06.wav	52,50	44,81	43,66
	dep07.wav	57,36	64,14	37,50
	dep08.wav	50,33	54,47	34,32
	Promedio	45,29	45,25	32,80
Moda	moda01.wav	19,30	8,95	18,55
	moda02.wav	16,32	6,62	18,90
	moda03.wav	22,71	8,73	12,15
	moda04.wav	21,99	12,06	18,18
	moda05.wav	20,99	20,99	24,24
	Promedio	20,24	11,10	18,19

Apéndice D

Presupuesto

1. Ejecución Material

- Compra de estación de trabajo 2.000€
- Material de oficina 200 €
- Total de ejecución material 2.200 €

2. Gastos generales

- 16 % sobre Ejecución Material 352 €

3. Beneficio Industrial

- 6 % sobre Ejecución Material 132 €

4. Honorarios Proyecto

- 1800 horas a 15 € / hora 27.000 €

5. Material fungible

- Gastos de impresión 80 €
- Encuadernación 30 €

6. Subtotal del presupuesto

- Subtotal Presupuesto 29.794 €

7. I.V.A. aplicable

- 21 % Subtotal Presupuesto.....6.256,80 €

8. Total presupuesto

- Total Presupuesto.....36.050,80 €

Madrid, Marzo de 2015

El Ingeniero Jefe de Proyecto

Fdo.: Juan Manuel Perero Codosero

Ingeniero de Telecomunicación

Apéndice E

Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, del DESARROLLO DE UN SISTEMA DE RECONOCIMIENTO DE HABLA NATURAL PARA TRANSCRIBIR CONTENIDOS DE AUDIO EN INTERNET. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho entorno. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.
2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.
3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.
4. La obra se realizará bajo la dirección técnica de un Ingeniero de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se

estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.
6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.
7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.
8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.
9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.
10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.
11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado

en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partidaalzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.
13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.
14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.
15. La garantía definitiva será del 4 % del presupuesto y la provisional del 2 %.
16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.
17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.
18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.
19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.
20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean

oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.
22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.
23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" y anteriormente llamado "Presupuesto de Ejecución Material" que hoy designa otro concepto.

Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.
2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.
4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.
5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.
6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.
7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.
8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.
9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.
10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.
11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.
12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.