Universidad del País Vasco
Euskal Herriko Unibertsitatea

Departamento de Electricidad y Electrónica

Zientzi Fakultatea * Facultad de Ciencias

Modelos k-explorables en sentido estricto integrados en un sistema de reconocimiento automático del habla

Memoria presentada para optar al grado de doctora por:

Amparo Varona Fernández

Dirigida por:

Dra. Inés Torres Barañano.

Leioa, febrero 2000

Índice

AGRADECIMIENTOS XI
PRÓLOGO XIII

Capítulo 1 Introducción y objetivos 1

1.1.- El sistema de reconocimiento 3
1.2.- Presentación de objetivos 5
1.3.- Entorno de trabajo 6

Capítulo 2 El modelo de lenguaje 7

2.1.- Métodos para la modelización del lenguaje 7
2.1.1.- Modelos estadísticos 8
2.1.2.- Modelos gramaticales 14
2.1.2.1.- Concepto de gramática, lenguaje y autómata 15
2.1.2.2.- Gramáticas incontextuales en RAH 18
2.1.2.3.- Gramáticas regulares en RAH 18
2.1.3.- Enfoque estadístico y/o gramatical 20
2.2.- Problemática de la estimación de las distribuciones de probabilidad 21
2.2.1.- Métodos convencionales para el suavizado 23
2.2.1.1.- Interpolación 24
2.2.1.2.- Back-off 26
2.3.- La perplejidad como medida de evaluación de los modelos de lenguaje 26

Capítulo 3 Modelo de lenguaje k-Explorable en Sentido Estricto (k-EE) 31

3.1.- Definición e inferencia de los lenguajes k-EE 32
3.1.1.- El Autómata de Estados Finitos Estocástico k-EE (AEFE k-EE) 34
3.2.- Lenguajes *k*-EE suavizados 36
3.2.1.- Aplicación del back-off sintáctico a los modelos k-EE 38
3.2.2.- Modelo integrado y suavizado 39
3.2.3.- Modelos de lenguaje k-EE y N-gramas 43
3.3.- Representación eficiente del autómata de estados finitos determinista K-EE suavizado ... 44
3.3.1.- Representación de los modelos K-EE en una estructura tipo array 44
3.3.2.- La función de transición δ a través de la representación interna del modelo K-EE 47
3.4.- Evaluación de los modelos *k*-EE 48
3.4.1.- Memoria 49
3.4.2.- Perplejidad 50

Capítulo 4 Sistema de RAH: Prototipo TORNASOL.....51
4.1.- Decodificación Acústico-Fonética.....52
4.1.1.- Técnicas de modelado.....53
4.1.2.- Selección de unidades.....54
4.1.2.1.- Unidades independientes de contexto.....54
4.1.2.2.- Unidades dependientes de contexto.....55
4.1.3.- Condiciones experimentales.....56
4.1.3.1.- La secuencia acústica: adquisición y preproceso.....57
4.1.3.2.- MOM: aprendizaje y reconocimiento.....58
4.1.3.3.- Unidades acústicas: fonemas y unidades contextuales obtenidas a partir de árboles de decisión.....62
4.1.3.4.- Criterios de evaluación.....63
4.1.4.- Resultados de reconocimiento.....64
4.1.4.1.- Unidades fonéticas incontextuales: fonemas.....65
4.1.4.2.- Unidades contextuales.....67
4.2.- Decodificación de palabras aisladas.....68
4.2.1.- Criterios de evaluación.....68
4.2.2.- Resultados de reconocimiento.....69
4.2.2.1.- Unidades incontextuales: fonemas.....69
4.2.2.2.- Unidades contextuales: árboles de decisión.....72
4.3.- Decodificación integrando conocimiento acústico y sintáctico.....74
4.3.1.- Integración de los modelos acústicos y de lenguaje.....74
4.3.1.1.- Implementación de la integración entre DAF y ML.....74
4.3.1.2.- Organización del léxico.....76
4.3.1.3.- Algoritmos de búsqueda.....76
4.3.2.- Condiciones experimentales y criterios de evaluación.....81
4.3.3.- Resultados experimentales con ajuste del parámetro que regula el ancho de haz.....83
4.4.- Evaluación de las ponderaciones de las probabilidades acústicas y/o de ML.....86
4.4.1.- Efecto de las ponderaciones sobre las probabilidades del ML.....88
4.4.1.1.- Escalado exponencial: $P(\Omega)^{\alpha}$.....89
4.4.1.2.- Escalado lineal: $\beta P(\Omega)$.....95
4.4.1.3.- Escalado lineal y exponencial: $\beta P(\Omega)^{\alpha}$.....95
4.4.1.4.- Aplicación de una función escalón.....96
4.4.2.- Resultados ponderando las probabilidades acústicas.....99
4.4.2.1.- Función exponencial: $P(A/\Omega)^{\gamma}$.....100
4.4.2.2.- Utilización de distintos conjuntos de parámetros acústicos.....104
4.4.3.- Utilización de modelos acústicos más precisos.....106
4.4.3.1.- Resultados con MOM semicontinuos.....106
4.4.3.2.- Resultados con unidades contextuales.....106
4.4.4.- Conclusiones.....107
4.5.- Arquitectura.....110

Capítulo 5 Evaluación del suavizado del ML en el sistema de RAH.....113
5.1.- Estudio de las distribuciones de probabilidad del ML.....114
5.1.1.- Atenuación de las probabilidades de las transiciones al estado de back-off.....115
5.1.2 .- Importancia del suavizado.....118
5.2.- Factores de descuento aplicables bajo el back-off sintáctico.....119
5.2.1.- Descuento sobre todos los eventos vistos.....122
5.2.1.1.- Descuento de Witten-Bell.....122

5.2.1.2.- Descuento Sencillo 123
5.2.1.3.- Descuento Absoluto 124
5.2.1.4.- Descuento Lineal 124
5.2.2.- Descuento sobre los eventos peor estimados 125
5.2.2.1.- Descuento de Katz 125
5.2.2.2.- Descuento Acotado 126
5.3.- Resultados experimentales 127
5.3.1.- Resultados sin reescalado de las probabilidades del ML 127
5.3.1.1.- Descuento Sencillo. 128
5.3.1.2.- Descuento Absoluto 129
5.3.1.3.- Descuento Lineal 131
5.3.1.4.- Descuento Acotado 132
5.3.1.5.- Resumen de resultados (α=1) 134
5.3.2.- Resultados reescalando las probabilidades del ML 136
5.3.2.1.- Resumen de resultados (α>1) 139
5.3.3.- Experimentación con una base de datos menor (tarea MLA) 141
5.3.3.1.- Resultados sin reescalado de las probabilidades del ML 142
5.3.3.2.- Resultados reescalando las probabilidades del ML 143

Capítulo 6 Evaluación de los modelos de lenguaje k-EE podados 145
6.1.- Evaluación experimental utilizando el descuento de Witten-Bell 146
6.1.1.- Escalando las probabilidades del ML: $P(ML)^{\alpha}$ 149
6.2.- Evaluación de los distintos descuentos 152
6.2.1.- Resultados sin reescalado de las probabilidades del ML 152
6.2.2.1.- Descuento Sencillo 152
6.2.2.2.- Descuento Absoluto 153
6.2.2.3.- Descuento Lineal 155
6.2.2.4.- Descuento Acotado 156
6.2.2.5.- Resumen de resultados (α=1) 158
6.2.2.- Resultados reescalando las probabilidades del ML 161
6.2.2.1.- Resumen de resultados (α>1) 164

Capítulo 7 Conclusiones y desarrollos futuros 167
7.1.- Conclusiones y aportaciones 168
7.2.- Líneas de trabajo futuras 170

Referencias bibliográficas 171

Apéndice A Bases de datos 183
A.1.- Corpus asociado al modelado acústico. 184
A.1.1.- Corpus de entrenamiento de los modelos acústicos 184
A.1.2.- Corpus de evaluación de los modelos acústicos 186
A.2.- Corpus asociados al modelado de lenguaje 186
A.2.1.- Tarea MLA 187
A.2.2.- Tarea BDGEO 191
A.3.- Distribución de fonemas de los corpus de voz 199

Apéndice B Representación de la estructura del modelo integrado 201

B.1.- La estructura del modelo K-EE suavizado como una red finita.....201

B.2.- Construcción de la estructura mediante un array.....204

Apéndice C Resultados de decodificación dependientes del conjunto de frases de evaluación.....207

C.1.- Evaluación de los distintos métodos de descuento sobre los ML 210

Índice de tablas

Tabla 3.1.- Número de estados y memoria necesaria para representar el autómata de la aplicación MLA para diferentes valores de K. 49

Tabla 3.2.- Número de estados y memoria necesaria para representar el autómata de la aplicación BDGEO para diferentes valores de K. 49

Tabla 4.1.- Conjunto de unidades para el Castellano (notación AFI), sus grafías y la notación utilizada por el transcriptor. 63

Tabla 4.2.- Resultados de reconocimiento $\%PRA_{DAF}$ en DAF utilizando modelos discretos. 66

Tabla 4.3.- Resultados de reconocimiento $\%PRA_{DAF}$ en DAF utilizando modelos semicontinuos. 66

Tabla 4.4.- Resultados de DAF con diferentes conjuntos de unidades subléxicas, para el test acústico 67

Tabla 4.5.- Resultados de reconocimiento de palabras aisladas aplicando distintos factores de peso entre palabras (MLA) 70

Tabla 4.6.- Resultados de reconocimiento de palabras aisladas aplicando distintos factores de peso entre palabras (BDGEO). 70

Tabla 4.7.- Representa el número de palabras y de fonemas que aparecen en las frases de referencia (Transcripción ortográfica) y en las reconocidas aplicando distinto factor de peso entre palabras) 71

Tabla 4.8.- Representa el número de palabras de longitud X en fonemas que aparecen tanto en las frases reconocidas como en las de referencia. 72

Tabla 4.9.- Resultados de %PRE utilizando unidades contextuales 73

Tabla 4.10.- Resultados obtenidos para la tarea MLA, número medio de nodos activos, tiempo medio por frame y %PRE para distintos valores de K. 84

Tabla 4.11.- Resultados obtenidos para la tarea BDGEO: número medio de nodos activos, tiempo medio por frame y %PRE para distintos valores de K. 84

Tabla 4.12.- Resultados del número medio de nodos activos, del tiempo medio por frame y del %PRE con distintos factores de ancho de haz (*ah*) para la tarea MLA 90

Tabla 4.13.- Resultados del número medio de nodos activos, del tiempo medio por frame y del %PRE con distintos factores de ancho de haz (*ah*) para la tarea BDGEO. 91

Tabla 6.1.- Evaluación de los modelos integrados de orden K podados (K=2...6) con diferentes factores de poda (fp). Se representa el número de estados (QK), la memoria, la perplejidad (PP), el Numero Medio de Nodos Activos por frame (NMNA), el Tiempo Medio de decodificación por frame (TMF) y el porcentaje real de error (%PRE). 147

Tabla 6.2.- Resumen de los resultados de %PRE obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5, 6). Se añaden los valores del intervalo de confianza $[PRE^{-}, PRE^{+}]_{95\%}$ para medir la significancia estadística de los resultados (5655 palabras) 147

Tabla 6.3.- Evaluación de los modelos integrados de orden K (K=3...5) podados para α=5 (mejor respuesta del sistema) 151

Tabla A.1.- Distribución en fonemas de los conjuntos de voz presentados en el apéndice A. 199

Índice de figuras

Figura 1.1.- Esquema general de un sistema de Reconocimiento Automático del Habla. 4
Figura 2.1.- Relación entre gramática, lenguaje y autómata. 17
Figura 2.2.- Representa las transiciones a partir de un nodo con una determinada historia h. 22
Figura 3.1.- $L_{1\text{-}EE}$, $L_{2\text{-}EE}$, $L_{3\text{-}EE}$ inferidos a partir de un conjunto de muestras positivas 33
Figura 3.2.- a) Transiciones a partir del estado inicial y entre estados representando secuencias de palabras de longitud inferior a *k-1*. b) Transiciones entre estados representando secuencias de longitud *k-1*. 35
Figura 3.3.- Corpus de entrenamiento seleccionado para realizar la construcción de los autómatas *k*-EE 36
Figura 3.4.- Los autómatas 1-EE, 2-EE, 3-EE y 4-EE obtenidos de R^+ y Σ de la figura 3.3. 37
Figura 3.5.- Las transiciones entre eventos vistos en un estado q ($w_j \in \Sigma_q$) conectan el estado con estados en el mismo submodelo K-EE. Las transiciones para los elementos no vistos (etiquetadas por U) conectan con su estado de back-off en el modelo (K-1)-EE. 42
Figura 3.6.- La red completa representando la estructura del modelo integrado y suavizado (K=4). 43
Figura 3.7.- Se representan los parámetros del modelo integrado y suavizado de la figura 3.6 46
Figura 3.8.- Función de búsqueda a través de la estructura presentada en la figura 3.7. 47
Figura 3.9.- Función de búsqueda extendida (δ_Viterbi) para dar como salida las probabilidades y estados destino para todas las palabras $|\Sigma|$ como posibles alternativas en decodificación. 48
Figura 3.10.- Representa la perplejidad asociada a los corpus de evaluación del ML (Prueba_Texto) y de voz (Prueba_Voz) que se van a utilizar en el sistema de RAH para las tareas MLA y BDGEO. 50
Figura 4.1.- (a) Espacio acústico cuantificado vectorialmente. (b). Espacio acústico convertido en conjuntos de mezclas de fdp. (c) Espacio acústico con densidad de probabilidad gausianas independientes. 60
Figura 4.2.- Topología de los MOM utilizados en la experimentación. 60
Figura 4.3.- Resultados de reconocimiento para cuatro frases elegidas arbitrariamente de la tarea BDGEO. 71
Figura 4.3.- a) Esquema de un sistema de RAH en el que los conocimientos acústicos y léxicos se integran en un primer módulo y el análisis sintáctico en un segundo módulo. b) Esquema de reconocimiento que integra en un único módulo todas las fuentes de conocimiento (acústica, léxica y sintáctica). 75
Figura 4.4.- En las transiciones entre los estados del ML se integran los modelos de las unidades léxicas correspondientes. 82
Figura 4.5.- El algoritmo de Viterbi obtiene el camino de máxima probabilidad a través de las unidades léxicas representadas en las transiciones entre estados del ML 82
Figura 4.6.- Resultados obtenidos evaluando el porcentaje real de errores (%PRE) para las tareas (MLA y BDGEO), y diferentes modelos integrados de orden K (K=2..6) y factores de ancho de haz (*ah*). 83
Figura 4.7.- Resultados de decodificación obtenidos para cuatro frases elegidas arbitrariamente de la tarea BDGEO. 85
Figura 4.8.- a) Reducción de las probabilidades del ML debido a los factores α y β. b) Incremento de las probabilidades acústicas debido al parámetro γ. 88
Figura 4.9.- Evolución del error con distintos valores de *ah* y de α para los modelos integrados de orden K con K=2, 4 y 6 (óptimo). Tarea de aplicación MLA. 89
Figura 4.10.- Evolución del error con distintos valores de *ah* y de α para los modelos integrados de orden K con K=2, 3 y 4 (óptimo). Tarea de aplicación BDGEO. 90
Figura 4.11.- Representa la evolución con a (1...7) del número medio de nodos activos (base logarítmica) para cada uno de los modelos integrados de orden K (K=2.. 6) y el porcentaje de error observado (%PRE).. 91
Figura 4.12.- Resultados de reconocimiento para cuatro frases elegidas arbitrariamente de la tarea BDGEO con un modelo K-EE con K=4 y *ah*=0.55. 93
Figura 4.13.- Representa las probabilidades acumuladas en el desarrollo del trellis de Viterbi. 94
Figura 4.14.- Representa la evolución del error aplicando un barrido de los valores de β para los modelos integrados de orden K con K=2, 3 y 4 para la tarea BDGEO. 95
Figura 4.15.- Representa la evolución del error para los modelos integrados de orden K con K=2, 3 y 4 cuando se aplican ambos factores (α>1 y β<1) sobre la tarea de BDGEO. 96
Figura 4.16.- Detalle de la evolución del error para los modelos integrados de orden K con K=2, 3 y 4 cuando se aplican ambos factores (α=5, 6 y 7 (alrededor del óptimo) y β>1) para la tarea de BDGEO. 96
Figura 4.17.- Representa los resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 y distintos valores del parámetro e. 97
Figura 4.18.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 cuando se modifican los valores del parámetro α en la parte superior del escalón (BDGEO). 98
Figura 4.19.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 para una función escalón suavizada (2α) en la parte inferior (BDGEO). 98

Figura 4.20.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 para una función escalón suavizada (1+α) en la parte inferior (BDGEO)..... 99
Figura 4.21.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ (*ah*=0.5).....101
Figura 4.22.- Resultados de error para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ y del ML α (*ah*=0.5). 102
Figura 4.23.- Detalle de los resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ y del ML α=4, 5, 6 y 7 (*ah*=0.5).....102
Figura 4.24.- Resultados de error para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ y del ML α (*ah*=0.6). 103
Figura 4.25.- Resultados de error para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ y del ML α (*ah*=0.8). 103
Figura 4.26.- Resultados de error para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ y del ML α (*ah*=0.9). 103
Figura 4.27.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 cuando se utilizan diferentes conjuntos de componentes acústicas (*ah*=0.5). 104
Figura 4.28.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 cuando se utilizan diferentes conjuntos de componentes acústicas y un barrido experimental sobre α (*ah*=0.5).....105
Figura 4.29.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 cuando se utilizan diferentes conjuntos de componentes acústicas (*ah*=0.6). 105
Figura 4.30.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 para α=4, 5, 6 y 7 utiliando modelos semicontinuos(BDGEO).....106
Figura 4.31.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores de α, cuando se utilizan las unidades contextuales para la tarea BDGEO. 107
Figura 4.32.- Resultados de error obtenidos para los modelos integrados de orden K con K=2 , 3 y 4 a partir de las unidades contextuales (detalle para α=4, 5, 6, 7) BDGEO.....107
Figura 4.33.- Arquitectura del prototipo TORNASOL. Se muestra la comunicación entre el cliente y el servidor a través de una red local.....111
Figura 5.1.- Resultados de perplejidad con los modelos de orden K, K-EE (K=2...6) y distintos factores de atenuación de las probabilidades de transición al estado de back-off (descuento de Witten-Bell).....116
Figura 5.2.- Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) cuando se utilizan distintos factores(ab). 116
Figura 5.3.- Resultados de error obtenidos para los modelos integrados de orden K (K=2...6) con el parámetro de atenuación de las probabilidades de transición al estado de back-off *ab*=1 y *ab*=0.01.....118
Figura 5.4.- Resultados de error obtenidos para la tarea BDGEO cuando se utilizan modelos integrados de orden K (K=2...6) suavizados con el descuento de Witten-Bell (S) y sin suavizar (NS).....119
Figura 5.5.- Resultados de perplejidad obtenidos con los modelos integrados de orden K (K=2...6) suavizados utilizando el descuento de Witten-Bell y descuento Sencillo sobre Prueba_Texto y Prueba_Voz.....128
Figura 5.6.- Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) utilizando el descuento Sencillo y descuento de Witten-Bell.....129
Figura 5.7.- Resultados de perplejidad (PP) y %PRE obtenidos para los modelos integrados de orden K (K=2...4) con los descuentos de Witten-Bell y descuento Sencillo.....129
Figura 5.8.- Resultados de perplejidad (PP) obtenidos para los modelos integrados de orden K (K=2...6) suavizados con el descuento de Witten-Bell y descuento Absoluto. b) PP en detalle para aquellos descuentos que dan resultados más parecidos.....130
Figura 5.9.- Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) utilizando el descuento Absoluto y descuento de Witten-Bell.....130
Figura 5.10.- Resultados de perplejidad (PP) obtenidos para los modelos integrados de orden K (K=2...6) suavizados con el descuento de Witten-Bell y descuento Lineal.....131
Figura 5.11.- Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) utilizando el descuento Lineal y descuento de Witten-Bell.....132
Figura 5.12.- Resultados de perplejidad (PP) obtenidos con los modelos integrados de orden K (K=2...6) suavizados utilizando el descuento de Witten-Bell y descuento Acotado.....133
Figura 5.13.- Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) con el descuento Acotado y el descuento de Witten-Bell. 133
Figura 5.14.- Resumen de las perplejidades más significativas obtenidas con los modelos integrados de orden K (K=2...6) utilizando los descuentos de Witten-Bell, Sencillo, Absoluto, Lineal y Acotado. 134
Figura 5.15.- Detalle de las perplejidades más significativas obtenidas con los modelos integrados de orden K (K=3...6) utilizando los descuentos de Witten-Bell, Sencillo, Absoluto, Lineal, y Acotado.....134
Figura 5.16.- Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) utilizando los descuentos de Witten-Bell, Sencillo, Absoluto (b=0.4 y b=0.01), Acotado (d=0.7 y d=0.99) y Lineal (l=0.1 y l=0.01).. 135

Figura 5.17.- Resultados de perplejidad (PP) y %PRE obtenidos para los modelos integrados de orden K con los distintos descuentos evaluados: Lineal (dL l=0.1), de Witten-Bell (dWB), Acotado (dAc d=0.7), Absoluto (b=0.4), Sencillo (dS), Lineal (dL l=0.01), Acotado (dAc d=0.99) y Absoluto (dAb b=0.4). 136
Figura 5.18.- Resultados de %PRE obtenidos en función del NMNA para los modelos integrados de orden K (K=2, 3 y 4) con el descuento Sencillo y el descuento de Witten-Bell. 136
Figura 5.19.- Resultados de perplejidad (PP) y %PRE obtenidos para los modelos integrados de orden K (K=2...4) utilizando los descuentos de Witten-Bell (α=6) y descuento Sencillo (α=5)..... 137
Figura 5.20.- Resultados de %PRE obtenidos en función del NMNA para los modelos integrados de orden K utilizando el descuento Absoluto con distintos valores del parámetro b y el descuento de Witten-Bell..... 138
Figura 5.21.- Resultados de %PRE obtenidos en función del NMNA para los modelos integrados de orden K (K=2, 3 y 4) utilizando el descuento Acotado con distintos valores del parámetro d y el descuento de Witten-Bell..... 138
Figura 5.22.- Resultados de %PRE obtenidos en función del NMNA para los modelos integrados de orden K (K=2, 3 y 4) utilizando el descuento Lineal con distintos valores del parámetro l y el descuento de Witten-Bell..... 139
Figura 5.23.- Detalle de los resultados del %PRE más significativos obtenidos para los modelos integrados de orden K (K=2..6) utilizando los descuento de Witten-Bell, Sencillo, Absoluto, Lineal y Acotado..... 139
Figura 5.24.- Resultados de perplejidad (PP) y %PRE obtenidos para los modelos integrados de orden K (K=2...4) con los distintos descuentos evaluados. 140
Figura 5.25.- Resumen de las perplejidades más significativas obtenidas con los modelos integrados de orden K (K=2...6) utilizando el descuento de Witten-Bell, Sencillo, Absoluto, Lineal y Acotado..... 142
Figura 5.26.- Detalle de las perplejidades más significativas obtenidas con los modelos integrados de orden K (K=2...6) utilizando el descuento de Witten-Bell, Sencillo, Absoluto, Lineal y Acotado..... 142
Figura 5.27.- Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) utilizando el descuento de Witten-Bell, Sencillo, Absoluto, Acotadoy Lineal 143
Figura 5.28.- Resumen de los resultados del %PRE más significativos obtenidos para los modelos integrados de orden K (K=2. 6) utilizando los descuento de Witten-Bell, Sencillo, Absoluto, Lineal y Acotado..... 144
Figura 6.1.- Perplejidad obtenida con los modelos integrados de orden K (K=2...6) no podados (*fp*=1) y podados (*fp*>1)..... 148
Figura 6.2.- a) %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, b) Numero medio de nodos activos en función de K, c) %PRE en función del NMNA. 148
Figura 6.3.- Evolución de la perplejidad y %PRE para los modelos integrados de orden K (K=3, 4 y 5) con distintos valores del parámetro *fp* (descuento de Witten-Bell). 149
Figura 6.4.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) para distintos valores del factor de poda (*fp*) y distintos valores del parámetro α. 150
Figura 6.5.- Evolución del porcentaje de %PRE y PP para los modelos integrados de orden K (K=3, 4 y 5) en función del valor del factor de poda (*fp*) (α=5)..... 151
Figura 6.6.- Resultados de perplejidad obtenidos con los modelos integrados de orden K (K=2, 3, 4, 5 y 6) suavizados con el descuento Sencillo. 152
Figura 6.7.- %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Sencillo..... 153
Figura 6.8.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5, 6) suavizados con el a) descuento Absoluto b=0.4, b) descuento Absoluto b=0.01. 154
Figura 6.9.- %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Absoluto b=0.4. 154
Figura 6.10.-%PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Absoluto b=0.01..... 154
Figura 6.11.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5, 6) suavizados con el a) descuento Lineal l=0.1 b) descuento Lineal l=0.01..... 155
Figura 6.12.- %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Lineal l=0.1..... 156
Figura 6.13.- %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Lineal l=0.01..... 156
Figura 6.14.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5, 6) suavizados con el a) descuento Acotado d=0.70, b) descuento Acotado b=0.99..... 157
Figura 6.15.- %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Absoluto d=0. 157
Figura 6.16.- %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Lineal l=0.01..... 157
Figura 6.17.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) no podados, *fp*=1..... 158
Figura 6.18.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) con distintos factores de poda *fp*>1..... 158
Figura 6.19.- Detalle de los resultados de perplejidad presentados en la figura 6.18..... 158

Figura 6.20.- Resultados del %PRE y NMNA obtenidos con los modelos integrados de orden K (K=2, 3, 4, 5 y 6) no podados para α=1. 159
Figura 6.21.- Resultados del %PRE obtenidos con los modelos integrados de orden K (K=2, 3, 4, 5 y 6) para α=1 con distintos valores de *fp*....159
Figura 6.22.- NMNA obtenidos con los modelos integrados de orden K (K=2, 3, 4, 5 y 6) para α=1 con distintos valores de *fp*....160
Figura 6.23 .- Resultados de perplejidad (PP) y %PRE obtenidos con el modelo integrado de orden K=4 utilizando los descuentos: Lineal (dL l=0.1), de Witten-Bell (dWB), Acotado (dAc d=0.7), Absoluto (b=0.4), Sencillo (dS), Lineal (dL l=0.01), Acotado (dAc d=0.99) y Absoluto (dAb b=0.4).....160
Figura 6.24.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Sencillo, para distintos valores del factor de poda (*fp*) y distintos valores del parámetro α.....161
Figura 6.25.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Absoluto b=0.4, para distintos valores de *fp* y distintos valores del parámetro α.....162
Figura 6.26.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Acotado d=0.7, para distintos valores de *fp* y distintos valores del parámetro α. 162
Figura 6.27.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Lineal l=0.1, para distintos valores de *fp* y distintos valores del parámetro α.....162
Figura 6.28.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Absoluto b=0.01, para distintos valores de *fp* y distintos valores del parámetro α.....163
Figura 6.29.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Acotado d=0.99, para distintos valores de *fp* y distintos valores del parámetro α.....164
Figura 6.30.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Lineal l=0.01, para distintos valores de *fp* y distintos valores del parámetro α. 164
Figura 6.31.- Resultados de %PRE obtenidos cuando se integra en el sistema de RAH el modelo integrado de orden K=4 y *fp*=1 A la derecha se muestran los resultados más en detalle.....165
Figura 6.32.- Resultados de %PRE obtenidos cuando se integra en el sistema de RAH el modelo integrado de orden K=4 y distintos valores de *fp*>1.....165
Figura 6.33.- Detalle de los resultados de la figura 6.32.....165
Figura 6.34 .- Resultados de perplejidad (PP) y %PRE obtenidos con el modelo integrados de orden K=4 (K=2...4) utilizando distintos descuentos. 166
Figura A.1.- Resumen de las bases de datos de texto y voz utilizadas a lo largo de la memoria.....183
Figura A.2.- Datos correspondientes a la longitud en palabras de la base datos correspondiente a la tarea MLA.....188
Figura A.3.- Datos correspondientes a la longitud en palabras de la base datos correspondiente a la tarea BDGEO. 192
Figura B.1.- Corpus de entrenamiento seleccionado para realizar la construcción de los autómatas *k*-EE. 202
Figura B.2.- Bi-trie obtenido directamente del conjunto de muestras de entrenamiento R^+ de la figura B.1 después de realizar el primer paso de la transformación para un modelo K-EE con K=4.....202
Figura B.3.- La red representando la estructura del bi-trie no suavizado (K=4) que se obtiene a partir de la figura B.2. 203
Figura B.4.- La red completa representando la estructura del modelo integrado y suavizado (K=4) obtenido a partir del corpus de entrenamiento de la figura B.1.....204
Figura C.1.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2..6) tras la separación del conjunto de evaluación (bdgeo_600), en función de los resultados de perplejidad.....208
Figura C.2.- Resultados de error obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) y los diferentes conjuntos bdgeo_600, bdgeo_464 y bdgeo_136 (tarea BDGEO).....208
Figura C.3.- Resultados de error obtenido sobre los modelos integrados de orden K (K=2, 3, 4, 5 y 6) cuando se aplica el descuento de Witten-Bell sobre los tres conjuntos de frases.....209
Figura C.4.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2... 6) con los conjuntos de 600, 464 y 136 frases respectivamente. 210
Figura C.5.- Resultados de perplejidad en detalle obtenidos para los modelos integrados de orden K presentados en la figura C.4.....210
Figura C.6.- Resultados de error obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) con los conjuntos de 600, 464 y 136 frases respectivamente. 211
Figura C.7.- Resultados del NMNA obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) con los conjuntos de 600, 464 y 136 frases respectivamente. 211
Figura C.8.- Resultados de error obtenidos para el modelos integrados de orden K con K=4 con los conjuntos de 600, 464 y 136 frases respectivamente.....212
Figura C.9-. Resultados de error (en detalle) obtenidos para el modelo K-EE con K=4 con los conjuntos de 600, 464 y 136 frases respectivamente.....212

AGRADECIMIENTOS

El apoyo y dedicación de Manés ha sido fundamental para mi experiencia investigadora. Ella no sólo ha sido mi directora en este trabajo de Tesis sino la que ha estado guiando toda mi formación en el mundo del reconocimiento de voz trabajando como miembro del Grupo de Reconocimiento Automático del Habla- Mintzo Ezagutza Automatikoren Taldea del que ella es fundadora y responsable.

He de destacar el apoyo de Germán Bordel en los inicios de mi trabajo de tesis, al dejarme participar en su trabajo y mostrarme el camino a seguir. La colaboración realizada desde los primeros momentos con el grupo de reconocimiento de formas de la Universidad Politécnica de Valencia, ha sido fundamental, siendo destacable la dedicación de Francisco Casacuberta y Enrique Vidal con sus sugerencias en las líneas a seguir. La posterior colaboración con la Universidad Politécnica de Barcelona, en especial con José B. Mariño y Antonio Bonafonte, nos ha permitido disponer de las bases de datos necesarias para llevar a buen termino la experimentación presentada.

Imposible olvidarme de mi compañero de despacho y de fatigas, Luisja. Creo que siempre ha estado ahí. Empezamos prácticamente juntos y desde entonces hemos ido cambiando de localización física, pero siempre he podido y puedo contar con él ante cualquier duda o problema que me surge. Ni por supuesto de Karmele no sólo compañera sino amiga, con quien he compartido a lo largo de los últimos años muchos momentos buenos y malos.

Pero aparte de los apoyos investigadores, se necesitan apoyos personales para poder llevar a buen puerto un trabajo tan largo como éste y que necesita tanta dedicación al margen de la actividad docente. El Departamento de Electricidad y Electrónica con todas las personas que lo constituyen me han dado el cobijo espacial y personal necesario para realizar ese trabajo. En especial debo dar las gracias al empeño de Ester Alonso quien siempre tiene una palabra de aliento cuando estoy un poco cansada y quien también sabe dar un toque de atención en los momentos de desidia. No puedo olvidarme tampoco de Iñaki Gomendiourritia, quien ha sido fundamental en los últimos y peores días antes de entregar la memoria, ayudándome con la estética de la misma.

Y por supuesto el apoyo de mis padres que siempre han sabido encontrar las palabras de aliento que he necesitado. También Oscar ha tenido que aguantar a lo largo de los últimos años momentos de conversación monotemática y de obsesivo trabajo.

PRÓLOGO

Los sistemas de Reconocimiento Automático del Habla (RAH) tratan de reconocer la secuencia de palabras pronunciadas por el locutor. Estos sistemas pueden tener aplicación en muchos campos de la vida cotidiana y por ello su desarrollo ha sido extraordinario en las últimas décadas. Los sistemas actuales tienden al reconocimiento de discurso continuo con grandes vocabularios y los resultados que se están logrando son realmente espectaculares.

Una de las etapas fundamentales en un sistema de RAH es la incorporación de conocimiento sintáctico y/o semántico mediante un Modelo de Lenguaje (ML). Tradicionalmente se han utilizado modelos estadísticos (N-gramas) obtenidos a partir de grandes bases de datos de texto en vez de modelos gramaticales capaces de recoger mejor la estructura sintáctica del lenguaje. Este hecho se ha justificado en la dificultad de integración de los modelos gramaticales con los modelos acústicos en los sistemas de RAH. En este trabajo se pretende evaluar una aproximación gramatical al modelado de lenguaje, alternativa al uso de la aproximación puramente estadística. Está basada en la utilización de una subclase de las gramáticas regulares, las gramáticas k-Explorables en Sentido Estricto (k-EE). El principal objetivo del trabajo es demostrar la adecuada integración de modelos k-EE en el sistema de RAH. Además la evaluación de cualquier modificación practicada a los ML se realizará en base a resultados de reconocimiento, que se comparan con las medidas de evaluación clásicas: perplejidad del corpus de test sobre texto.

Con la integración directa de las probabilidades acústicas y de ML no se logran los resultados de decodificación óptimos. En este trabajo se van a estudiar y evaluar los heurísticos que se utilizan tradicionalmente (ponderación lineal y exponencial sobre las probabilidades del acústicas y de ML) para mejorar las tasas del sistema, junto con nuevas propuestas (aplicación de una función escalón, combinación de heurísticos, etc).

Uno de los principales problemas a la hora de evaluar un ML es la falta de cobertura. Siempre hay una gran cantidad de combinaciones de palabras que no aparecen en las muestras de entrenamiento y a las que el modelo asigna una probabilidad igual a cero. Para evitar este problema se recurre a aplicación de las técnicas de suavizado. Se trata de descontar una cierta masa de probabilidad sobre los eventos vistos para que sea repartida entre los eventos no vistos. En este trabajo se van a evaluar varios descuentos utilizando el suavizado por back-off sintáctico.

Para reducir el tamaño de los modelos se va a evaluar también el efecto que tiene en decodificación la poda de los modelos k-EE suavizados con los distintos descuentos evaluados previamente. Se comparará el diferente comportamiento de los modelos podados frente a los no podados.

El esquema de la presente memoria es como sigue:

En el capítulo 1 se hace una introducción al problema general del reconocimiento automático del habla. En este trabajo tanto los modelos acústicos como de lenguaje se

desarrollan dentro de la aproximación inductiva (aprendizaje automático a partir de muestras) y en este capítulo se presentan las ecuaciones que los relacionan (regla de Bayes).

En el capítulo 2 se realiza una revisión bibliográfica de los modelos de lenguaje estadísticos y gramaticales aplicados al RAH. Dentro de la aproximación gramatical se presentan los conceptos básicos que se utilizan a lo largo de la memoria, introduciendo la extensión estocástica. Además, se presenta una discusión sobre las ventajas e inconvenientes de la aproximación gramatical frente a la aproximación puramente estadística. Finalmente, se introduce la perplejidad como medida de bondad de los ML, en una evaluación externa al sistema de RAH y se presentan las técnicas de suavizado básicas necesarias para que el ML no asigne una probabilidad igual a cero a ninguna secuencia de palabras.

El capítulo 3 está dedicado al estudio y desarrollo de los modelos k-Explorables en Sentido Estricto (k-EE) que serán evaluados a lo largo de la memoria. En primer lugar, se presenta la definición e inferencia de los lenguajes k-EE (García & Vidal, 1999), junto con la extensión estocástica (Segarra, 1993) que permite su utilización en el sistema de RAH. A continuación se define y construye el autómata k-EE de estados finitos determinista. Estos modelos se pueden construir fácilmente a partir de las muestras de entrenamiento pero para poder ser integrados en el sistema de RAH es necesario el uso de alguna de las técnicas de suavizado presentadas en el capítulo 2.

Finalmente, se formula la integración en un sólo modelo de K autómatas k-EE deterministas (k=1..K) y la técnica de suavizado (back-off sintáctico utilizando el descuento de Witten-Bell (Bordel et al, 1994)) en un esquema de construcción recursivo. Además se presenta una forma eficiente de representación del modelo completo para que pueda ser fácilmente integrado con los modelos acústicos en el sistema de RAH. Los modelos integrados obtenidos se evalúan según el coste de memoria necesario para su representación, observándose que el uso de modelos con valores altos de K no supone un gasto prohibitivo. En este capítulo los modelos se evalúan de forma independiente al sistema de RAH utilizando como medida la perplejidad.

En el capítulo 4 se presenta el prototipo de RAH TORNASOL desarrollado en la UPV/EHU integrando los modelos acústicos (Modelos Ocultos de Markov discretos y semicontinuos de unidades subléxicas incontextuales, los fonemas, y contextuales obtenidas a partir de árboles de decisión) y los Modelos de Lenguaje (gramáticas k-Explorables en Sentido Estricto). Los experimentos se han realizado fundamentalmente utilizado dos corpus: el primero corresponde a una tarea muy sencilla de 29 palabras (Feldman et al, 1990) y el segundo corresponde a una tarea (BDGEO) de vocabulario medio (1208 palabras) (Díaz et al, 1993).

En primer lugar se presentan los resultados de decodificación acústico fonética y de palabras aisladas para establecer las condiciones iniciales del sistema antes de evaluar el modelo integrado con la incorporación del conocimiento aportado por el ML. Para lograr la adecuada integración de los modelos acústicos y de lenguaje es necesario la aplicación de heurísticos (Bourlard et al, 1996) (Rubio et al, 1997), que modifiquen la contribución acústica y/o la del ML, lo cual no es una solución muy ortodoxa desde el punto de vista de la teoría estadística pero

permite paliar el problema derivado de la distinta naturaleza de las distribuciones de probabilidad que se intentan combinar. En este trabajo se abordará este problema también en el capítulo 4 estudiando el efecto que produce sobre los resultados de decodificación el uso de diferentes funciones de ponderación sobre las probabilidades acústicas y/o de ML. En concreto se evalúan distintas funciones (lineales, exponenciales, escalones, etc.) sobre una o ambas distribuciones de probabilidad.

En general, la necesidad real de la redistribución de las probabilidades del ML, de una u otra manera, pone de manifiesto que la distribución de probabilidad obtenida tras el suavizado de los modelos utilizando el descuento de Witten-Bell no es la más adecuada (se sobrestiman las probabilidades de los eventos no vistos en entrenamiento) a pesar de haber sido elegido por proporcionar los mejores resultados de perplejidad (Bordel, 1996).

En el capítulo 5, se evalúan y comparan los resultados de decodificación obtenidos aplicando diversas técnicas de descuento. Propuestas clásicas: descuento de Witten-Bell, Absoluto y Lineal (Ney et al 1997) (Clarkson & Rosenfeld, 1997) y nuevas propuestas: descuento Sencillo y Acotado.

El estudio nos ha permitido probar experimentalmente la fuerte relación existente entre las distribuciones de probabilidad del ML suavizadas, (dependientes de las técnicas de descuento) y el factor de escalado de las probabilidades del ML necesario para alcanzar la máxima eficacia del sistema. De hecho, cuanta mayor dispersión de las probabilidades del ML se obtengan, lo cual implica un modelo menos suavizado, se necesita un valor menor del factor de reescalado para lograr el óptimo. Para todos los descuentos estudiados las diferencias del porcentaje de error obtenidas alrededor del óptimo, no son significativas. Además se puede observar que no siempre el mejor suavizado de los ML, evaluado en perplejidad, resulta ser el mejor suavizado para conseguir el rendimiento óptimo del sistema. Estos resultados permiten cuestionar que la perplejidad pueda en todos los casos predecir el comportamiento del ML en el sistema integrado ya que para conseguir unos buenos resultados parece importante trabajar con el valor del parámetro de ponderación de las probabilidades del ML adecuado.

En el capítulo 6 se ha evaluado el efecto de la poda de los modelos en el sistema de RAH para reducir el coste espacial y temporal. A pesar de que los modelos gramaticales desarrollados en la presente memoria necesitan relativamente poca cantidad de memoria para ser representados, los requisitos espaciales con valores altos de k empiezan a ser significativamente altos y puede ser recomendable aplicar una técnica de poda para reducir el tamaño de los modelos tratando de mantener su efectividad en el sistema de RAH. La técnica de poda que se ha utilizado es muy simple y consiste en eliminar estados del autómata tras exigir un número mínimo de ocurrencia de la combinación de palabras que representa el estado. Este criterio hace que se eliminen muchos estados asociados a historias específicas que aparecen muy pocas veces y que sólo se modelen las historias que aparecen con una relativa regularidad.

Como en los modelos podados las distribuciones de probabilidad son completamente distintas dependiendo de la técnica de descuento que se emplee, se ha evaluado el

comportamiento del sistema en función del parámetro de ponderación de las probabilidades del ML, mostrando cúales son los descuentos que menos degradan el comportamiento de los modelos podados en el sistema de RAH. En este caso, las diferencias en el porcentaje de error alrededor del óptimo si que son significativas. De nuevo se manifiesta que la perplejidad no siempre es capaz de determinar cual es el suavizado que mejores resultados proporciona al sistema completo de reconocimiento.

Finalmente en un capítulo de conclusiones se presenta un resumen del trabajo, indicando las aportaciones realizadas y se bosquejan las líneas de futuras investigaciones relacionadas con el trabajo presentado. La memoria finaliza con la relación de bibliografía y tres apéndices. El primero de ellos está dedicado a presentar los corpus de voz y texto utilizados junto con sus principales características. En el segundo se muestra la representación interna de la estructura del modelo de lenguaje integrado. El tercero presenta los resultados de decodificación obtenidos cuando se evalúan los distintos métodos de descuento sobre dos conjuntos de frases de voz muy diferentes; por un lado se evalúan las frases "fáciles" que corresponden a aquellas en las que la mayoría de combinaciones de palabras se han visto en entrenamiento y las frases "difíciles" en las que muchas combinaciones de palabras no se han visto en entrenamiento.

Capítulo 1

Introducción y objetivos

Un sistema de Reconocimiento Automático del Habla (RAH) puede ser definido como cualquier mecanismo, distinto del sistema auditivo humano, capaz de decodificar la señal acústica producida por el aparato fonador humano en una secuencia de unidades lingüísticas que contiene el mensaje que ese locutor desea comunicar.

Los sistemas de RAH han evolucionado mucho en las dos últimas décadas. Se partía del reconocimiento de palabras aisladas con vocabularios pequeños y desde hace unos pocos años se logra, con relativa facilidad, el reconocimiento de discurso continuo de tareas no excesivamente restringidas (Huang et al, 1993) (Ney et al, 1994). Las nuevas tendencias llevan al reconocimiento de lenguajes seudonaturales con gran cantidad de vocabulario. Estos sistemas, lógicamente, necesitan hacer uso de métodos que no supongan costes (espaciales y/o computacionales) inabordables y que se desarrollen en tiempo real.

El habla es una señal compuesta, jerárquicamente organizada, de forma que los patrones más simples en un determinado nivel, se combinan de forma definida para formar patrones más complejos en el siguiente nivel. En (Levinson, 1985) se propone la construcción de los sistemas de reconocimiento de discurso continuo de acuerdo a una jerarquía abstracta compuesta de tres niveles. Esta organización es la base de la mayoría de los sistemas actuales de RAH, en los que se definen los siguientes niveles: subléxico, léxico, y sintáctico. En el nivel subléxico se representan las unidades subléxicas escogidas (fonemas, sílabas, semisílabas, etc.) mediante modelos estadísticos (por ejemplo, Modelos Ocultos de Markov). En el nivel léxico, estos modelos acústicos previamente obtenidos, se concatenan para formar palabras siguiendo las reglas léxicas definidas. Finalmente los modelos resultantes de palabras se concatenan para formar frases, siguiendo las reglas sintácticas y/o semánticas de concatenación de palabras (típicamente, haciendo uso de modelos estadísticos o gramaticales). En este tercer nivel se podría añadir el conocimiento pragmático, que proporciona la información suficiente para dar a las frases sintácticamente correctas un significado. A los mecanismos para definir y controlar las posibles concatenaciones de las unidades en el tercer nivel (palabras) es a lo que se llama Modelos de Lenguaje.

Dentro de un sistema de RAH, el modelado acústico y de lenguaje, así como los mecanismos necesarios para su integración son temas de vital importancia. Es por ello, que la

comunidad científica internacional está volcada en encontrar o mejorar técnicas, que supongan alguna aportación teórico y/o práctica al sistema de reconocimiento, sin degradar el tiempo y/o el espacio necesario para su correcto funcionamiento.

Cuando se intenta abordar la construcción de un sistema de RAH, nos encontramos con dos aproximaciones principalmente:

deductiva: la información se le aporta al sistema a partir de los conocimientos previos adquiridos por expertos en el tema (ejemplos típicos de uso de la aproximación deductiva pueden ser la elección del número de unidades subléxicas, estructura de los modelos acústicos, etc.).

inductiva: la información se obtiene a partir de un conjunto de muestras de entrenamiento (ejemplos típicos de uso de la aproximación inductiva pueden ser la obtención de la distribución de las probabilidades de los modelos acústicos y de lenguaje, etc.).

Los sistemas con una gran carga en la aproximación deductiva presentan principalmente dos problemas: la incorporación del conocimiento a priori del sistema se hace mucho más difícil a medida que va aumentando la complejidad de la tarea a abordar y la adaptación a cada variación de la tarea supone un duro trabajo por parte del diseñador del sistema.

Por ello, parece más interesante el uso de la aproximación inductiva. No obstante, los sistemas normalmente se construyen combinando ambas aproximaciones, ya que en algunos puntos se puede recurrir al conocimiento adquirido por un experto, por ejemplo para fijar la topología de los modelos acústicos, como se verá en el capítulo 4.

El objetivo de este trabajo está totalmente vinculado a la construcción de un sistema de reconocimiento automático del habla de discurso continuo, en el que tanto el nivel acústico como de ML se desarrollen dentro de la vía del aprendizaje automático. El trabajo se enmarca dentro del área del Reconocimiento de Formas y se trata de asignar modelos acústicos a ciertas unidades lingüísticas, e imponer restricciones a las posibles concatenaciones de dichas unidades a partir del conocimiento extraído del conjunto de muestras de aprendizaje.

Desde un punto de vista general, una técnica de Reconocimiento de Formas (RF) puede verse como un traductor que convierte los objetos dados como entrada, mediante unas adecuadas representaciones, en mensajes que permiten la identificación de tales objetos. En el campo del RF se encuentran metodologías procedentes de diferentes campos: teoría de la señal, teoría de lenguajes formales, modelos estocásticos, etc. Sin embargo, se puede establecer una clasificación entre la llamada aproximación estadística o geométrica y la aproximación sintáctica.

Normalmente la aproximación estadística se adopta para los problemas en los que las formas vienen representadas en forma de vectores en un espacio multidimensional de características. Esta aproximación se basa típicamente en la utilización de la regla de decisión de Bayes para minimizar las tasas de error (Duda, 1973) y se habla de la aproximación basada en la teoría de decisión. Sin embargo, hay otros problemas en los que la estructura de las formas juega un papel muy importante y se necesita un formalismo que maneje información estructurada. En

estos casos se trabaja con la aproximación sintáctica al RF, que recurre usualmente a conceptos de la teoría de lenguajes formales (Hopcroft & Ullman, 1979). En esta aproximación los unidades se representan como composición de objetos más elementales. La descripción estructural de una clase de formas viene dada a través de una gramática formal y el reconocimiento puede llevarse a cabo mediante procedimientos basados típicamente en técnicas de análisis sintáctico (Fu, 1992).

La obtención de las reglas que guían el análisis sintáctico, sean gramáticas o máquinas como los autómatas finitos puede ser de forma manual, sin embargo, con el objeto de mecanizar al máximo el trabajo en el sistema de RF, es deseable que sean obtenidas por medio de procesos de aprendizaje inductivo. La teoría de los lenguajes formales proporciona los mecanismos básicos que permiten tanto el aprendizaje como el reconocimiento en un sistema de RF sintáctico. Generalmente los fenómenos de ruido y distorsión que se presentan en los objetos reales, hace necesaria la incorporación de técnicas que permitan la modelización del error (Casacuberta & Vidal, 1987) como por ejemplo las gramáticas estocásticas, cuyas reglas tienen asociados unos valores de probabilidad (Segarra, 1993).

1.1.- El sistema de reconocimiento

El objetivo final de un sistema de reconocimiento automático del habla es el de obtener la secuencia de unidades léxicas (clásicamente palabras) que han sido pronunciadas por el locutor.

Formalmente, el Reconocimiento Automático bajo la aproximación estadística (Jelinek, 1976) consiste en encontrar aquella secuencia de L palabras $\Omega = w_1 ... w_L$ que con mayor probabilidad corresponden a una observación acústica dada A. O lo que es lo mismo, que maximice la probabilidad a posteriori de Ω dada A. En definitiva, se trata de encontrar la secuencia de palabras Ω que maximice P(Ω/A).

$$\Omega = \arg \max_{\forall \Omega'} P(\Omega' / A) \tag{1.1}$$

Aplicando la **regla de Bayes** la ecuación 1.1 se convierte en **:**

$$P(\Omega / A) = \frac{P(A / \Omega) P(\Omega)}{P(A)} \tag{1.2}$$

donde $P(A) = \sum_{\forall \Omega} P(A / \Omega) P(\Omega)$ es una cantidad constante, por lo que el problema se reduce a maximizar el producto P(A/Ω) P(Ω).

$$\Omega = \arg \max_{\forall \Omega'} P(\Omega') P(A / \Omega') \tag{1.3}$$

La probabilidad de A dada Ω, P(A/Ω), es la probabilidad acústica y se puede estimar usando algún modelo estadístico. Quizá, los modelos más comúnmente utilizados sean los Modelos Ocultos de Markov (MOM), aunque existen varias alternativas (Castro, 1999). En los

sistemas de vocabularios medianos y grandes, se utiliza un conjunto de unidades subléxicas (por ejemplo fonemas) para construir mediante concatenación el modelo de cada palabra w_i. La probabilidad de Ω, P(Ω), es la probabilidad de la secuencia de palabras pronunciada. Se obtiene a partir de la probabilidad a priori dada por un Modelo de Lenguaje (ML) que trata de integrar las restricciones sintácticas y/o semánticas de un lenguaje. En definitiva, este término representa el conocimiento que tiene el sistema sobre cómo formar frases con las palabras del léxico.

La integración de las probabilidades acústicas y de lenguaje se puede realizar en un sólo paso o en dos como se explica en detalle en el capítulo 4. En este punto nos vamos a fijar en el esquema de reconocimiento en un sólo paso que consiste en una integración completa de los modelos acústicos y de Lenguaje. Si el ML es muy simple, la integración será más "barata" en coste computacional, pero los resultados de decodificación obtenidos no serán muy buenos. Por el contrario, si el ML es muy complicado, los resultados de decodificación serán óptimos pero puede resultar un coste computacional inabordable. Por supuesto, habrá que llegar a un compromiso.

La figura 1.1 muestra el esquema general de un sistema integrado de RAH continua. De hecho, este es el esquema del prototipo TORNASOL (ver capítulo 4) desarrollado en la Universidad del País Vasco /Euskal Herriko Unibertsitatea (Rodríguez et al, 1999). El prototipo se ha realizado dentro del proyecto SENGLAR (TIC95-0884-C04-03) titulado "Construcción de sistemas robustos de reconocimiento de habla continua y gran vocabulario mediante técnicas de aprendizaje automático".

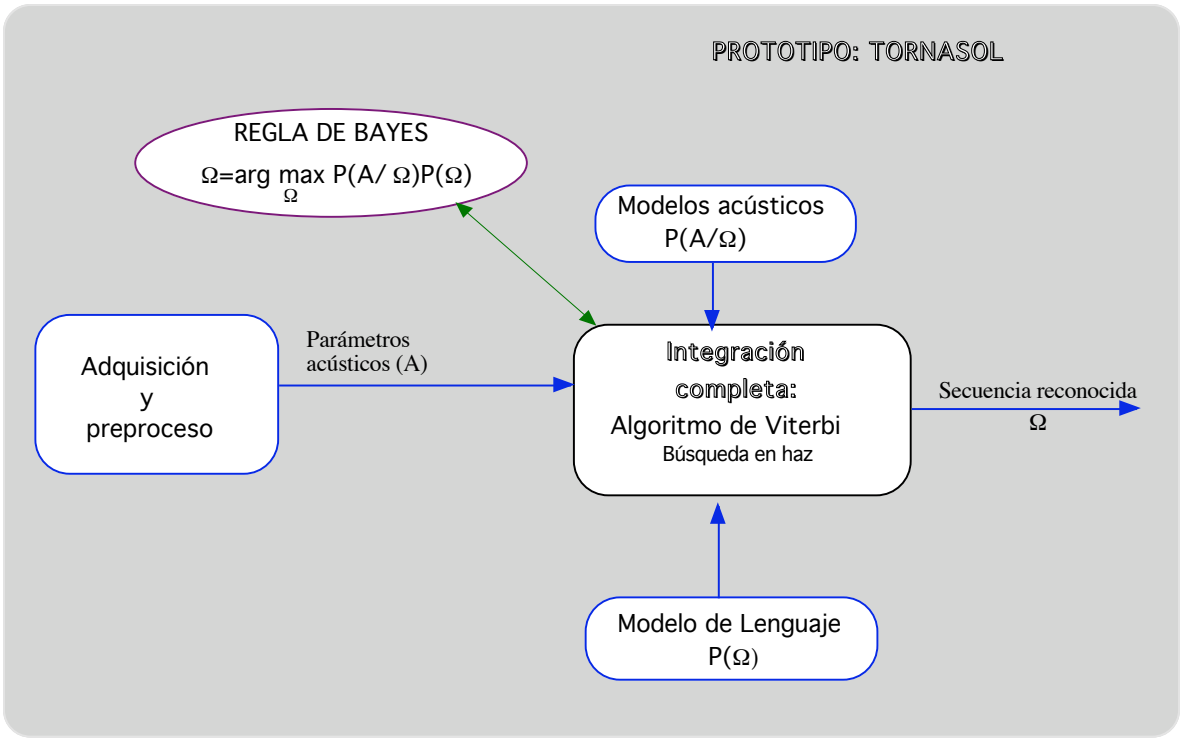


Figura 1.1.- Esquema general de un sistema de Reconocimiento Automático del Habla.

Tras la etapa de adquisición y preproceso de la señal de voz, se obtiene la secuencia de vectores de características acústicas (A). La secuencia finalmente decodificada se puede calcular directamente integrando las probabilidades proporcionadas por el ML y los modelos acústicos según muestra la ecuación 1.3. Cada uno de los estados del ML se sustituye por un autómata que representa la concatenaciones de los modelos (Modelos Ocultos de Markov) individuales que constituyen la transcripción fonética de las palabras (los detalles se pueden ver en el capítulo 4).

Dentro del prototipo TORNASOL, se va a trabajar con el algoritmo de Viterbi como algoritmo de decodificación (ver en detalle en el capítulo 4). Para reducir el coste computacional asociado a la decodificación de cada sentencia se pueden utilizar varios heurísticos como por ejemplo la búsqueda en haz que también se explicará en el capitulo 4. Finalmente la secuencia de palabras Ω es reconocida.

1.2.- Presentación de objetivos

A continuación se presentan los principales objetivos pretendidos en el desarrollo del presente trabajo:

1. Una de las etapas fundamentales en un sistema de RAH es la incorporación de conocimiento sintáctico y/o semántico mediante un Modelo de Lenguaje (ML). Para el desarrollo de tal ML se pueden utilizar dos aproximaciones: estadística o gramatical. Tradicionalmente se han utilizado modelos estadísticos (N-gramas) porque el uso de gramáticas ha sido considerado como una aproximación muy difícil para ser integrada junto con los modelos acústicos en los sistemas de RAH (Segarra, 1993). En este trabajo se va a aplicar una aproximación gramatical al modelado de lenguaje, basada en la utilización de una subclase de las gramáticas regulares, conocidas como gramáticas k-Explorables en Sentido Estricto (k-EE).
2. Además, la construcción y evaluación de los ML se ha realizado tradicionalmente de forma independiente al sistema de decodificación para el que habían sido construidos. En este trabajo la evaluación de cualquier modificación practicada a los ML se realiza en base a resultados de reconocimiento, que se comparan con las medidas de evaluación clásicas: perplejidad del corpus de test sobre texto.
3. Con la integración directa de las probabilidades acústicas y de ML no se logran los resultados de decodificación óptimos. En este trabajo se van a estudiar y evaluar los heurísticos que se utilizan tradicionalmente (ponderación lineal y exponencial sobre las probabilidades del acústicas y de ML) para mejorar el comportamiento del sistema, junto con nuevas propuestas (aplicación de una función escalón, combinación de heurísticos, etc.).

4. El uso de las gramáticas k-EE permite integrar en un sólo modelo K autómatas k-EE deterministas (k=1...K) y la técnica de suavizado por back-off. La formalización de este modelo es también objetivo del trabajo.
5. El suavizado de los ML es fundamental para evitar las probabilidades igual a cero para las combinaciones de palabras que no aparecen en las muestras de entrenamiento. En este trabajo se van a evaluar, dentro del sistema de RAH, varios descuentos usando la técnica de back-off sintáctico: propuestas clásicas (Witten-Bell, Absoluto y Lineal) y nuevas propuestas (Sencillo y Acotado).
6. Para reducir el tamaño de los modelos se van a evaluar también el efecto que tiene en decodificación la poda de los modelos k-EE suavizados con los distintos descuentos evaluados previamente. Se comparará el diferente comportamiento de los modelos podados frente a los no podados.

1.3.- Entorno de trabajo

El trabajo presentado en esta memoria y muchos más se han podido desarrollar a partir de la colaboración entre varias universidades españolas dedicadas a sacar adelante el mismo objetivo común. En primer lugar y dentro del proyecto Albaycin (TIC91-1488) se diseñó la tarea BDGEO (tarea de consulta a una base de datos geográfica, ver apéndice A) sobre la que se han realizado la mayor parte de los experimentos de esta memoria. Este trabajo fue realizado por un consorcio formado por 5 universidades españolas: Universidad Autónoma de Barcelona, Universidad de Granada, Universidad Politécnica de Barcelona, Universidad Politécnica de Valencia, Universidad Politécnica de Madrid.

El desarrollo del sistema de reconocimiento (prototipo TORNASOL) se ha realizado dentro del proyecto SENGLAR (TIC95-0884-C04-03) con una participación muy importante de la autora de la presente memoria. El proyecto de tres años se titulaba “Construcción de sistemas robustos de reconocimiento de habla continua y gran vocabulario mediante técnicas de aprendizaje automático” y ha sido desarrollado por cuatro universidades La Universidad Politécnica de Valencia, Universidad Politécnica de Barcelona, Universidad del País Vasco y Universidad de Zaragoza.

El proyecto ha culminado con éxito, dando como resultado cuatro prototipos que funcionan *on-line* (Bonafonte et al, 1996) (Llorents et al, 1999) (Rodriguez et al, 1999 a,b) y una evaluación positiva por parte de la CICYT.

Capítulo 2

El modelo de lenguaje

Como se ha explicado en el capítulo anterior, dentro de un sistema de Reconocimiento Automático del Habla, el Modelo de Lenguaje (ML), se sitúa en un nivel superior al de la formación de palabras y establece las restricciones y relaciones existentes en la combinación de las mismas.

Debido a que la distribución de las combinaciones de palabras no es uniforme en el lenguaje humano, es conveniente utilizar la extensión estocástica de los ML, que proporciona una distribución de probabilidad de las combinaciones de palabras y que por tanto permite conocer cual es la palabra más probable tras una secuencia previa.

La probabilidades del ML, $P(\Omega)$, se combinan con las probabilidades acústicas, $P(A/\Omega)$, utilizando la regla de Bayes presentada en la ecuación 1.3. Si el decodificador acústico tuviera una certeza del 100% en el reconocimiento, la inclusión del ML no sería necesaria, pero ésto está bastante lejos de la realidad, por lo que cuanto mejores sean los ML (modelos con mayor información), mayor ayuda podrán proporcionarle al decodificador acústico y se obtendrán mejores decodificaciones.

Una vez vista la fundamental importancia los ML en el sistema de RAH de discurso continuo y teniendo en cuenta que en este trabajo se pretende la evaluación e integración de una clase concreta de los ML (ver capítulo 3) se va a presentar en primer lugar las diferentes propuestas de construcción de los modelos de lenguaje (aproximación estadística y gramatical).

Dentro de la aproximación gramatical se va a presentar el concepto de gramática, lenguaje y autómata bajo la aproximación estocástica. A continuación se muestra la estimación de las probabilidades asociadas al ML, donde se incluye una definición de los principales métodos de suavizado. Finalmente, se presenta la evaluación de los ML, de forma independiente a la inclusión de los mismos en el sistema de RAH, haciendo uso de medidas de perplejidad.

2.1.- Métodos para la modelización del lenguaje

En las aplicaciones que precisan del conocimiento relativo al lenguaje, cabe distinguir también dos alternativas muy diferentes:

a) aproximación deductiva: Modelización del Lenguaje basada en el conocimiento (del inglés "Knowledge based"). Se trata de plasmar en el sistema los conocimientos de expertos en el tema. Entre ellos podemos encontrar los primeros métodos en los que se imponían las restricciones sintáctico-semánticas de forma manual, lógicamente para tareas bastante restringidas (Bahl et al, 1983).

b) Aproximación inductiva: obtención de los modelos de lenguaje a partir de la información contenida en muestras de entrenamiento. Es decir, que el conocimiento necesario sea adquirido por el sistema, sin aportación de conocimiento humano.

Se trata de dos tendencias claramente identificadas. No obstante, no tienen porqué ser exclusivas y se puede pensar en una vía alternativa en la que no todo el conocimiento fuera extraído de manera automática sino que una parte fuera aportada por expertos humanos.

En lo que sigue, se pretende dar una visión general del enfoque basado en el aprendizaje automático a partir de muestras de entrenamiento. La idea básica del aprendizaje a partir de muestras es que cuando el lenguaje presenta regularidades, es posible obtener un modelo a partir del estudio de un subconjunto finito de sentencias en las que se manifiesten esas regularidades. Dentro de esta aproximación cabe distinguir fundamentalmente entre dos tipos de modelos: estadísticos o gramaticales.

Los modelos estadísticos asignan las probabilidades "a priori" a las secuencias de palabras a partir de la cuenta del número de veces que aparece una determinada palabra tras la o las palabras que le preceden. En el área del procesamiento del lenguaje natural las primeras técnicas utilizadas fueron las estimaciones de las probabilidades basadas en N-gramas (Jelinek, 1991).

Teniendo en cuenta que las restricciones del lenguaje se pueden modelar mejor bajo una aproximación sintáctica, también es posible usar formalismos gramaticales para desarrollar los ML. Usando por ejemplo: gramáticas de contexto libre (Jurafsky et al, 1995), (Sánchez & Benedí, 1999) o gramáticas regulares en cuyo caso se trabaja con autómatas de estados finitos estocásticos (Riccardi et al, 1995), (Bordel, 1996). Sin embargo, el uso de gramáticas no está muy extendido debido a los problemas computacionales. De hecho, la integración completa con los modelos acústicos en una tarea de RAH ha sido considerada durante mucho tiempo como una tarea muy difícil bajo estos formalismos (Jelinek, 1991), (Segarra, 1993).

2.1.1.- Modelos estadísticos

La aproximación estadística básicamente consiste en obtener las frecuencias de aparición de cadenas (concatenación de unidades léxicas) en las frases tomadas como muestras de entrenamiento. El ML estocástico debe proporcionar estimaciones de las probabilidades a priori $P(\Omega \equiv w_1,...w_L)$ de que se produzcan las distintas secuencias de palabras w_1, w_2, ..,w_L de longitud L. Usando la definición de la probabilidad condicional, se obtiene la descomposición en

el producto de la probabilidad de cada palabra, una vez que se han dado las que le preceden en la frase:

$$P(\Omega \equiv w_1,...w_L) = P(w_1)\prod_{i=2}^{L} P(w_i / w_1,...w_{i-1}) = P(w_1)\prod_{i=2}^{L} P(w_i / w_1^{i-1}) \qquad (2.1)$$

Se va a adoptar la notación w_j^i para una cadena $w_j,...w_i$ y $P(w_i / w_1^{i-1})$ es la probabilidad de ocurrencia de la palabra w_i tras la secuencia w_1^{i-1}.

Por tanto, la obtención de la probabilidad de cada frase del lenguaje puede basarse en la estimación de la probabilidad que cada palabra que forma la frase tiene de aparecer detrás de las que le preceden. De la expresión anterior se deduce que el problema de la estimación de parámetros del ML queda reducido a la estimación de un conjunto de probabilidades, una para cada combinación posible de palabras w_1^i. Esto en la práctica resulta imposible ya que el número de estas combinaciones crece exponencialmente con la longitud de las frases. Basta ver lo que sucede con un lenguaje cuyo vocabulario sea de 100 palabras: cada distribución constará de 100 valores, y habrá 100 distribuciones relacionadas con cadenas previas de longitud 1, 10.000 relacionadas con cadenas de longitud 2, 1.000.000 con cadenas de longitud 3, etc (Jelinek, 1993).

Con objeto de reducir la dimensionalidad del problema, es práctica común reducir el número de distribuciones, asociando la misma probabilidad a todas las cadenas previas que coincidan en un determinado número de palabras. En otras palabras, se equiparan todas las distribuciones para aquellas cadenas de palabras que sólo difieran en palabras que distan de la analizada más de un determinado umbral. Es decir:

$$P(w_1^{i-1}) \cong P(w_{i-N+1}^{i-1}) \qquad (2.2)$$

donde P son distribuciones de probabilidad asociadas a las cadenas, w_1^{i-1} y w_{i-N+1}^{i-1} respectivamente, N determina que el número de palabras significativas en las cadenas previas sea de N-1. Es por ello que se habla de N-gramas, ya que cada valor de probabilidad se asigna a una cadena de longitud N.

Al considerar los N-gramas la ecuación 2.1 pasa a ser:

$$P(\Omega) = P(w_1)\prod_{i=2}^{L} P(w_i / w_{i-(n-1)}^{i-1}) \quad \text{con n=max}\{\text{N, i}\} \qquad (2.3)$$

Tradicionalmente el valor de N se toma igual a 2 y se habla de bigramas o igual a 3 y se habla de trigramas. En principio para valores más altos de N no se han conseguido mejoras significativas y sí se aprecia un importante aumento del tamaño de los modelos (Jelinek, 1991) (Placeway et al, 1993).

Una estimación de los valores de las distribuciones de probabilidad puede llevarse a cabo mediante el simple recuento del número de veces que se da cada situación en el conjunto de entrenamiento siguiendo el criterio de Máxima Verosimilitud (Duda, 1993):

$$P(w_i / w_{i-(n-1)}^{i-1}) = \frac{n(w_i / w_{i-(n-1)}^{i-1})}{n(w_{i-(n-1)}^{i})} \tag{2.4}$$

donde $n(w_i / w_{i-(n-1)}^{i-1})$ representa la cuenta del número de veces (frecuencia) que se da la palabra w_i tras la secuencia $w_{i-(n-1)}^{i-1}$ y $n(w_{i-(n-1)}^{i})$ representa el número de palabras totales que se han observado tras la secuencia $w_{i-(n-1)}^{i-1}$.

El modelo de los unigramas consiste en una sola distribución de probabilidad (básicamente la frecuencia de cada palabra en el lenguaje) ya que no se condiciona a ninguna cadena previa (la longitud de la cadena previa es N-1=0).

La formulación anterior da lugar a que el modelo no se ajuste a las muestras de aprendizaje. Efectivamente, una cadena w_1^{i-1} nunca vista en la muestra de aprendizaje recibirá una probabilidad distinta de cero en el caso de que la subcadena $w_{i-(N-1)}^{i-1}$ si que haya sido vista. Este comportamiento no es negativo ya que, la muestra de aprendizaje siempre será un subconjunto de las muestras de lenguaje. Es positivo siempre que la generalización realizada dé lugar a la aceptación de frases que se ajusten a las regularidades del lenguaje.

Sin embargo, no es posible contar con bases de datos de texto suficientemente grandes que contengan todas las combinaciones posibles de unidades léxicas o palabras. Se hace por ello necesaria la utilización de técnicas de suavizado que resten una cierta masa de probabilidad sobre las situaciones vistas y la redistribuyan entre las situaciones no vistas. La generalización del modelo es consecuencia directa de esta modificación de la distribución total de la probabilidad. La forma en que se haga esta modificación es determinante para la captura de las regularidades del lenguaje, siendo de vital importancia en la calidad del modelo final. Las técnicas de suavizado permiten controlar esta modificación de la distribución llegando a determinar probabilidades distintas de cero para todas las frases de la aplicación. Estas técnicas de suavizado están por supuesto, forzadas a mantener el cumplimiento de la condición estocástica, es decir, la suma de las probabilidades extendida al conjunto total de las frases debe ser la unidad. La distribución de probabilidad frente a las frases de muestra para un modelo que se ajuste al corpus de entrenamiento tendrá un aspecto extremadamente "anguloso", mientras que para un modelo que generalice, el perfil será más suave.

Un problema importante asociado de los modelos de N-gramas es la gran cantidad de muestras necesarias para el aprendizaje. Para tratar de limitar el problema se puede optar por intentar modelar lenguajes con restricción en la temática de manera que las frases se refieren sólo a una determinada tarea y en consecuencia el vocabulario se encuentre suficientemente acotado. A estos lenguajes se les suele llamar "lenguajes pseudo-naturales".

A pesar de que el modelo de N-gramas es un modelo muy sencillo, puede recoger un porcentaje importante de las restricciones de un lenguaje pseudo-natural con valores relativamente pequeños de N. No obstante, esta capacidad de captura de la información con valores bajos de N depende en cierta medida del idioma. Por ejemplo las concordancias en

Inglés suelen producirse en distancias más cortas que en Castellano por lo que los valores bajos de N recogen porcentualmente más información.

En la mayor parte de los sistemas, se han estado utilizando y se siguen utilizando trigramas. En general, utilizar valores mayores de N supone un alto coste computacional, sin que las mejoras sean apreciablemente mejores (Jelinek, 1990). Sin embargo, los trigramas no son capaces de recoger dependencias a mayor distancia de tres en los lenguajes naturales. Aproximaciones más recientes tratan de evitar estas limitaciones. Entre ellas se pueden encontrar:

a) Modelos que recogen relaciones a corta distancia

Modelos de lenguaje basados en categorías de palabras (Ney et al, 1994): La solución es agrupar las palabras de baja frecuencia (y por tanto mal estimadas con el método clásico de N-gramas) en clases equivalentes y usar los N-gramas para estimar las probabilidades entre las clases de palabras. Cuanto menor sea el número de grupos formados, más sencillo resultará el modelo, al reducirse el vocabulario y por tanto el número de parámetros que deben ser aprendidos. Aunque en un menor grado existe también la misma problemática que en los modelos de los N-gramas en lo que se refiere a la cobertura y por tanto también se usan los métodos de suavizado.

En los primeros algoritmos se trabajaba con bigramas y cada palabra se limitaba a estar incluida en una sola clase. Por ejemplo, en (Paeseler et al, 1989) la obtención de las clases se hacía de forma manual. En estos momentos la principal línea de actuación es la de elegir correctamente las clases de palabras de forma automática. El agrupamiento se puede hacer de diversas formas. Por ejemplo, se puede hacer un agrupamiento por conocimiento lingüístico. Las reglas que ligan las palabras no se establecen en relación con la identidad concreta de cada palabra, sino con su función dentro de la frase. Esta forma de agrupamiento, quizá sea la más extendida. Estos grupos se suelen denominar "Parts Of Speech" (POS) (Jelinek, 1990).

En (Dumouchel et al, 1988) se estudiaban diferentes modelos de N-gramas y N-POS, llegando a la conclusión de que las ventajas de los modelos N-POS se hacen patentes en su uso con tareas cuya temática sea muy amplia. Para lenguajes de temática restringida es muy posible que sea preferible utilizar el modelo de N-gramas. En (Samuelsson & Reichl, 1999) se presenta un método automático de división del conjunto de palabras basado en el comportamiento sintáctico de la palabra dentro de la frase. Cada palabra pertenece a una sola clase. Se prueba sobre bigramas y trigramas obteniéndose muy buenos resultados.

Esta aproximación presenta algunas desventajas según (Rosenfeld, 1994).

1) Hay palabras que pueden pertenecer a más de una clase.

2) Hay muchas clasificaciones posibles y distintas aunque se aplique el mismo conocimiento lingüístico.

3) La clasificación tiene sentido desde el punto de vista lingüístico pero no tiene porqué ser óptima desde el punto de vista del modelado de lenguaje.

En (Ney et al, 1994) y (Martin et al, 1998) se presenta un agrupamiento distinto conducido por los datos. Se utiliza una aproximación automática basada en dividir las palabras cuando se minimiza una cierta medida del conjunto de entrenamiento (la perplejidad que se estudiará más adelante).

En (Jardino, 1996) se dice que los resultados proporcionados por los métodos automáticos dependen del orden en que se haya llevado a cabo el proceso de clasificación. En este trabajo la clasificación se llevó a cabo con el método de "simulated annealing".

También es posible el agrupamiento por dominio de conocimiento. Por ejemplo, si la tarea es de reserva de vuelos, tiene sentido agrupar por nombres de compañías aéreas, nombres de aeropuertos, de ciudades, etc (Rosenfeld, 1994).

b) Modelos que recogen relaciones a distancia intermedia.

i) *Los N-gramas a distancia X*, (del inglés "*distance-X N-grams"*) (Rosenfeld, 1994). El modelo trata de capturar directamente la dependencia con las palabras representadas por el N-1 grama que se encuentra a alguna distancia X. Un caso especial son los N-gramas a distancia 1 que son los N-gramas convencionales. En (Huang et al, 1993) se estudió la magnitud de la información relevante de los bigramas a larga distancia y se concluyó que sólo hay información significativa en una distancia inferior a 5.

En (Martin et al, 1999) se prueban también los trigramas a distancia 2 (Distance-2 trigrams). Se trata de tripletas de palabras con huecos entre las mismas, es decir, las palabras que se consideran no son consecutivas. En realidad utilizan dos tipos de trigramas a distancia (w_{n-3},,w_{n-1},w_n), en las que el hueco está entre las dos primeras palabras del trigrama, o (w_{n-3},w_{n-2},,w_n), en las que el hueco está en las dos últimas palabras del trigrama.

ii) *Modelos de N-gramas de longitud variable* (Deligne et al, 1995): ajustan la longitud de los contextos dinámicamente dependiendo de la cantidad de datos de entrenamiento disponible. En el toolkit de la CMU (Clarkson & Rosenfeld, 1997) se pueden construir este tipo de modelos. Los modelos de N-gramas pueden incluir N-gramas, N-1-gramas, N-2-gramas ... hasta 1-grama. En (Bonafonte & Mariño, 1996) se presenta también la idea del X-grama pero asociándolo con la idea de gramática. En (Niesler & Woodland, 1999) se presenta un ejemplo de este modelo trabajando con grupos de palabras (categorías). Para mantener el modelo con un tamaño adecuado, sólo se introducen N-gramas a mayor distancia de tres, cuando la mejora de los resultados es significativa.

c) Modelos que recogen relaciones a larga distancia

i) *modelos Triggers*: Dentro del lenguaje natural es fácil encontrar relaciones entre pares de palabras o secuencias de palabras. Ejemplos típicos del castellano son "no sólo/sino que" o "médico/enfermedad". Existen experimentos sicológicos que indican que es más importante y más rápida la reacción humana ante combinaciones de palabras fuertemente relacionadas. La relación entre palabras se puede expandir desde la corta a la larga distancia. El uso de pares de Trigger (*Triggers pair*) (Rosenfeld, 1994) se usa como el concepto básico para representar la

información asociada al par de palabras. Si una palabra A esta fuertemente relacionada con otra B entonces (A→B) es un "trigger pair". A y B se pueden extender a secuencias de palabras. Hay dos problemas básicos a resolver a la hora de construir los triggers. ¿Cómo seleccionar los triggers y como medir su importancia?. Para ello se puede aplicar el criterio de información mutua.

En (Rosenfeld, 1994) se obtienen las siguientes conclusiones. Diferentes tipos de triggers presentan diferente comportamiento y por tanto deben ser modelados de diferente manera. Triggers de la forma (A→A) son particularmente útiles. La mayor parte de los triggers se concentran en las palabras de alta frecuencia. Se pueden utilizar también triggers negativos. Uno de los últimos trabajos presentados utilizando esta aproximación es (GuoDong & KimTeng, 1999).

ii) *Modelos Cache* (Kuhn et al, 92) (Clarkson, et al, 1997b) incrementan la probabilidad de las palabras ya observadas por medio de un factor que decae exponencialmente con la distancia. La principal motivación detrás del unigram-cache es que, una vez que una palabra ocurre en un texto, su probabilidad de volver a ocurrir es típicamente alta. Pero este fenómeno está relacionado por la frecuencia "a priori" de las palabras y es más pronunciado para las palabras raras. Por ejemplo la información asociada a la aparición de la palabra "el" (articulo muy utilizado en Castellano) es muy poco significativa. Pero si ocurre otra palabra rara la información es mucho mayor.

d) Modelos de Máxima entropía: Este tipo de modelos se sitúa a parte de la clasificación anterior, ya que incluyen combinaciones de algunos modelos ya presentados anteriormente. Tradicionalmente, cada una de las fuentes del conocimiento se usa separadamente para construir los respectivos modelos y después se combinan. Bajo la aproximación de máxima entropía los modelos no se construyen por separado, sino que se construye un modelo único tratando de capturar toda la información aportada por las diferentes fuentes de conocimiento.

Si la intersección de todas las restricciones impuestas por las diferentes fuentes de conocimiento no es vacía, contiene un conjunto con un gran número (incluso infinito) de funciones de probabilidad que deben ser consistentes con todas las fuentes de conocimiento. La aplicación del principio de máxima entropía consiste en elegir de entre todo el conjunto de funciones aquella que proporciona la más alta entropía. En otras palabras, una vez que se incorpora todas las fuentes de conocimiento, no se asume nada más acerca de los datos. Si las restricciones de todas las fuentes son coherentes se garantiza la existencia de una solución de forma iterativa, pero no se sabe es el número de iteraciones necesarias, lo que puede ocasionar un coste computacional bastante alto.

En (Rosenfeld, 1994) se utiliza el principio de máxima entropía para crear un modelo que integra varias fuentes de conocimiento. En concreto de utiliza con N-gramas y triggers. En (Khudanpur & Wu, 1999) también se utiliza el principio de máxima entropía explorando las dependencias a larga distancia con modelos dinámicos en función del tema de conversación. En (Martin et al, 1999) se combinan N-gramas y trigramas a distancia 2.

2.1.2.- Modelos gramaticales

La obtención de los N-gramas tradicionalmente ha sido relegada a obtener información de los segmentos locales de palabras a partir de la probabilidad "a priori". Esto hace imposible modelar relaciones a larga distancia. Por ejemplo un lenguaje con sentencias tan simples como $(ab^+c)+(cb^+a)$ hace que un modelo de N-gramas sea inadecuado dado que la relación "si comienza por a termina por c y viceversa" no es representable por cadenas de longitud determinada (Vidal et al, 1995). Sin embargo, este tipo de relaciones no presenta dificultades importantes en la especificación de los lenguajes formales mediante gramáticas, por lo que resulta interesante aplicar los conocimientos de la Teoría de los Lenguajes Formales a la modelización automática del lenguaje. Sin embargo, las gramáticas no se han usado extensivamente porque se dice que presentan graves problemas computacionales. De hecho, la integración completa con los modelos acústicos en una tarea de RAH han sido consideradas como aplicaciones muy difíciles bajo estos formalismos (Segarra, 1993).

En (Jelinek, 1991) se disertaba sobre las alternativas a los trigramas. En este trabajo se presenta como una posible alternativa al modelado de lenguaje la aproximación gramatical. Como limitación se destaca la dificultad de su aprendizaje a partir de texto.

En (Rosenfeld, 1994) se postula que la información contenida en las restricciones sintácticas es difícil de medir cuantitativamente, sin embargo, serían muy beneficiosas porque esta fuente de conocimiento parece ser complementaria a las fuentes de conocimiento estadísticas. De hecho, muchas de los errores de los reconocedores de voz, los humanos los pueden identificar porque violan las restricciones sintácticas.

Por otro lado, la Teoría de Lenguajes Formales es un disciplina bien asentada, con una formulación rigurosa que permite desarrollos en un marco puramente teórico (Fu & Jung, 1972) (Hopcrof & Ullman, 1979). La conexión de esta disciplina con las técnicas de modelización del lenguaje a partir de muestras se encuentra en la Inferencia Gramatical (IG), denominación que se aplica a la obtención de gramáticas para lenguajes a partir de la observación de muestras de los mismos. En (Vidal et al, 1995) se muestran un gran número de algoritmos de inferencia para determinadas clases de lenguajes.

Cuando se plantea la modelización del lenguaje desde el punto de vista gramatical es necesario hacerlo también desde el punto de vista estocástico. En (González & Thomason, 1978) o (Fu & Juang, 1972) se pueden encontrar las bases de la teoría de lenguajes en su extensión estocástica (se tendrá una probabilidad asociada a cada regla de producción). En (Gold, 1967) se encuentra formalizado el problema de la determinación de las reglas de la gramática asociada al lenguaje a partir de muestras, así como un estudio de los límites del aprendizaje automático. Estos límites son diferentes en función de si se dispone de una muestra completa (muestras que pertenecen y no pertenecen al lenguaje) o únicamente una muestra positiva (sólo frases del lenguaje). Para el caso de una muestra completa, cualquier lenguaje de la clase de lenguajes dependientes del contexto o más simple, puede ser identificado en el límite (es decir, cuando el

tamaño de la muestra crece, la distancia entre el lenguaje definido por el modelo y el lenguaje objetivo disminuye de forma que, en el límite -cuando se aporten todos los elementos del lenguaje objetivo- la identificación sea total). Pero es habitual no disponer de muestras negativas y en este caso surge la imposibilidad de identificación (ni siquiera en el límite) de ninguna clase superfinita (clase que contiene todos los lenguajes finitos y al menos uno infinito) de lenguajes. En consecuencia, dado que la clase de los lenguajes regulares es superfinita, ni tan siquiera ésta es identificable.

Dada la limitación anterior, la inferencia de los lenguajes se aborda en dos direcciones diferentes: por una parte se intenta caracterizar clases de lenguajes no superfinitos que puedan presentar algún interés y por otra lado se establecen heurísticos basados en el conocimiento a priori de la tarea a abordar y que dan lugar a clases de lenguajes no caracterizadas (se desconocen sus propiedades). El desarrollo de la Inferencia Gramatical para clases dentro de los lenguajes regulares es bastante amplio debido a su utilidad en muchos problemas prácticos. Incluso lenguajes que no son regulares pueden aproximarse con arbitraria precisión mediante gramáticas regulares estocásticas. Los métodos de inferencia desarrollados son numerosos ya que cada unos de ellos determina una subclase en particular de los lenguajes Regulares.

El siguiente nivel en dificultad es el de los Lenguajes Independientes del Contexto. Estos suponen un fuerte incremento en la complejidad de los métodos a aplicar y en los requerimientos computacionales necesarios frente a los anteriores. También en este caso los métodos desarrollados infieren, en general, subclases particulares. En (Vidal et al, 1995) se hace una exposición de las técnicas de aprendizaje sintáctico.

2.1.2.1.- Concepto de gramática, lenguaje y autómata

Gramática: una gramática es una formulación compacta de un lenguaje, puesto que de ella pueden derivarse las sentencias del mismo. Una gramática queda definida por 4 elementos (Σ,N,P,S) donde:

Σ: es el vocabulario y a sus elementos se les denomina Terminales (puede representar el léxico).

N: es el conjunto de símbolos auxiliares y a sus elementos se les denomina no terminales o variables (puede representar categorías sintácticas).

P: es el conjunto de reglas o derivaciones que combinan los elementos Σ y N. P es un conjunto finito de pares (α,β) con $\alpha\in(N\cup\Sigma)^*N(N\cup\Sigma)^*$ y $\beta\in(N\cup\Sigma)^*$, que se suele denotar por $\alpha\rightarrow\beta$.

S: es el axioma de la gramática. Es un elemento de N que se toma como origen de toda derivación de las frases del lenguaje.

La jerarquía de Chomsky (Hopcroft & Ullman, 1979) establece una clasificación de las gramáticas formales, que refleja el poder generativo o expresivo de las mismas. Las gramáticas se clasifican según las restricciones impuestas a sus reglas.

Tipo 0: sin restricciones en las reglas.

Tipo 1: dependientes de contexto. Las reglas son de la forma $\alpha X\beta \rightarrow \alpha\gamma\beta$ donde $\alpha,\beta \in (N \cup \Sigma)^*$, $X \in N$ y $\gamma \in (N \cup \Sigma)^+$

Tipo 2: incontextuales. Las reglas son de la forma $X \rightarrow \alpha$, donde $X \in N$ y $\alpha \in (N \cup \Sigma)^*$

Tipo 3: regulares. Las reglas, o bien, son lineales por la izquierda, o bien, lineales por la derecha. Sean $X, Y \in N$, $a \in (\Sigma \cup \{\varepsilon\})$. Lineales por la derecha: $X \rightarrow aY$; $X \rightarrow a$

Lineales por la izquierda: $X \rightarrow Ya$; $X \rightarrow a$

Lenguaje: para i=0, 1, 2, 3 un lenguaje es de tipo i si y sólo si, es generado por una gramática de tipo i. La clase de los lenguajes de tipo i se representa por L_i. Esta clasificación define conjuntos cada vez más restrictivos de los lenguajes, de forma que los del tipo i incluyen a los del tipo i+1. Se puede afirmar por tanto que: $L_3 \subset L_2 \subset L_1 \subset L_0$. Esta clasificación representa también una clasificación de la dificultad de su aprendizaje automático. Los lenguajes más restrictivos, los Lenguajes Regulares, son los que representan una menor dificultad y además son de una gran utilidad como se verá a lo largo de la memoria.

Autómata: un autómata es un modelo analítico que permite reconocer las frases de un determinado lenguaje (ver figura 2.1). Los autómatas que permiten reconocer los lenguajes sin restricciones (tipo 0) reciben el nombre de máquinas de Turing. Aquellos que permiten reconocer lenguajes dependientes de contexto (tipo 1) reciben el nombre de autómatas linealmente acotados. Los que reconocen lenguajes independientes de contexto (tipo 2) reciben el nombre de autómatas a pila y finalmente los lenguajes regulares (tipo 3) se reconocen usando los autómatas de estados finitos.

Un autómata de estados finitos deterministas es una quíntupla $(\Sigma, Q, \delta, q_0, F)$ donde Σ, es un conjunto finito de símbolos o alfabeto de entrada, Q es un conjunto finito de estados, δ es la función de transición $\delta: Q \times \Sigma \rightarrow Q$, $q0 \in Q$ es el estado inicial, $F \subseteq Q$ es el conjunto de estados finales o aceptores.

Se dice que una cadena de entrada $x \in \Sigma^*$ es aceptada por un autómata A si y sólo si $\delta(q_0,x) \in F$. El lenguaje aceptado por una autómata de estados finitos A se define como $L(A)=\{x \in \Sigma^* | \delta(q_0,x) \in F\}$.

Existen otros tipos de autómatas de estados finitos, los autómatas de estados finitos no deterministas se definen de forma similar a los no deterministas, pero ampliando la función a δ: $Q \times \Sigma \rightarrow 2^Q$, de forma que dado un estado y un símbolo de entrada se permite transitar a un conjunto de estados.

Se puede demostrar la equivalencia entre los lenguajes aceptados por autómatas de estados finitos y los lenguajes generados por gramáticas de tipo 3 o regulares. La demostración proporciona un mecanismo para la obtención de una gramática regular equivalente dado un autómata finito no determinista y viceversa. En consecuencia, el manejo de lenguajes regulares se puede llevar a cabo con gramáticas o autómatas indistintamente.

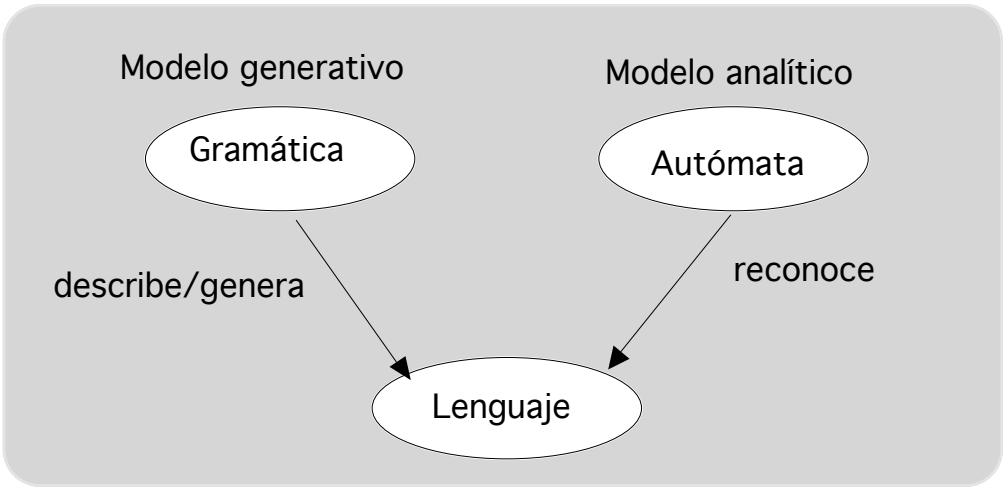


Figura 2.1.- Relación entre gramática, lenguaje y autómata. Una gramática permite representar (describir frases) de un lenguaje. El autómata es un dispositivo teórico que manipula la información y que reconoce las frases del lenguaje.

Gramáticas y autómatas estocásticos

En la práctica, para trabajar con un sistema de reconocimiento de discurso continuo, algunas cadenas del lenguaje ocurren con más frecuencia que otras, por lo que conviene aportar el conocimiento a priori de las frecuencias de ocurrencias de tales cadenas en las propias gramáticas. Además, el ruido y la distorsión causan errores en el reconocimiento de las primitivas o palabras del lenguaje, por lo que las gramáticas deben introducir probabilidades en el proceso de clasificación y se recurre a los conceptos de gramática y autómata estocástico.

Una gramática estocástica se define como una quíntupla (Σ,N,P,S,D) donde Σ es el vocabulario, N es el conjunto de símbolos auxiliares, P es el conjunto de reglas que combinan los elementos Σ y N y S es un elemento de N que se toma como origen de toda derivación de las frases del lenguaje y finalmente D es el conjunto de probabilidades asignadas a las reglas de P.

En la demostración de equivalencia entre gramáticas regulares y autómatas finitos no deterministas, se establece una correspondencia entre las producciones en la gramática y las transiciones entre estados del autómata. Esta correspondencia puede ser extendida a los modelos estocásticos, haciendo probabilísticas las transiciones entre estados de la misma forma que se asignan probabilidades a las reglas de la gramática.

Un autómata finito estocástico queda definido por la quíntupla $(\Sigma, Q, \delta, q_0, \{qf\})$ donde, donde Σ, es un conjunto finito de símbolos o alfabeto de entrada, Q es un conjunto finito de estados, $q_0 \in Q$ es el estado inicial, $F \subseteq Q$ es el conjunto de estados finales o aceptores y δ es la función de transición δ: $Q \times \Sigma \rightarrow Q \times [0...1]$ que da como resultado un estado destino con una probabilidad asociada[1].

[1] Otra posibilidad sería definir el autómata como un 6-tupla $(\Sigma, Q, \delta, q_0, \{qf\}, D)$ donde Σ, es un conjunto finito de símbolos o alfabeto de entrada, Q es un conjunto finito de estados, $q0 \in Q$ es el estado inicial, $F \subseteq Q$ es el conjunto de estados finales o aceptores, δ es la función de transición δ: $Q \times \Sigma \rightarrow Q$ *y D* es un conjunto de probabilidades entre cero y la unidad asociadas a las transiciones del autómata.

El autómata estocástico definido previamente es determinista y opera de la siguiente forma: se sitúa en el estado inicial y empieza a leer la cadena de entrada, suponiendo que en un tiempo dado, está en el estado q y lee el símbolo de entrada a, entonces $\delta(q,a)$ proporciona el estado destino junto con una probabilidad de transición entre 0 y 1. Si se trata de un autómata no determinista el resultado sería el conjunto de estados potenciales 2^Q a los que se puede transitar con una probabilidad asociada de transición por cada estado. De forma análoga a una gramática estocástica, el autómata estocástico proporciona un reconocimiento probabilístico.

La capacidad de los autómatas estocásticos como aceptores probabilísticos viene asegurada por el resultado siguiente: si un lenguaje estocástico L es generado por una gramática regular estocástica entonces L=L(A) para algún autómata finito estocástico A. Este resultado proporciona un mecanismo para obtener un autómata finito estocástico equivalente, a partir de una gramática regular estocástica.

2.1.2.2.- Gramáticas incontextuales en RAH

Las gramáticas incontextuales se utilizan en modelización del Lenguaje como una herencia del procesamiento del Lenguaje Natural (Segarra, 1993). Para poder ser utilizados en reconocimiento de discurso continuo, el ML debe asignar probabilidades a las cadenas que genera o acepta, es decir, hay que introducir estas gramáticas dentro de un marco probabilístico (Dupont, 1993), (Jurafsky et al, 1995).

El principal problema que se debe solucionar es que dado un conjunto de reglas que especifican la gramática incontextual. ¿Cómo estimar las probabilidades de las mismas?. Este problema se resuelve con el conocido algoritmo "*Inside-Outside*" (Backer, 1979). Existen otras alternativas como por ejemplo, un algoritmo basado en la aproximación de Viterbi que logra una convergencia más rápida (Ney, 1992b) aunque los resultados son peores porque las probabilidades se obtienen a partir de la mejor derivación. Otra aproximación posterior utiliza las m mejores derivaciones obteniéndose mejores resultados (Sánchez & Benedí 1997). La adaptación de estos modelos de lenguaje para la integración en tareas reales de RAH es un problema aún por resolver (Sánchez & Benedí 1999).

2.1.2.3.- Gramáticas regulares en RAH

Las gramáticas regulares dentro de los sistemas de reconocimiento se han utilizado desde los primeros sistemas: Dragon (Baker, 1975) y Harpy (Lowerre, 1976). De hecho, los cuatro sistemas del proyecto ARPA-SUR[2] utilizaban formalismos gramaticales. La principal dificultad era la falta de transportabilidad entre tareas.

---

[2] "Advanced Research Projects Agency- Speech Understanding Research". Proyecto de investigación con propósitos muy optimistas para la época, lanzado en EE.UU. en 1971 con un presupuesto de quince millones de dólares y una duración de 5 años.

Estos formalismos en su primera concepción no evaluaban la hipótesis sino que indicaban sólo las frases que eran válidas. Esto no permitía una buena combinación con la salida del decodificador que no eran ni son 100% certeros en sus decisiones.

Una de las primeras aproximaciones que modelaron el lenguaje de la tarea mediante una gramática regular, aunque no estocástica se utilizó en el sistema Harpy. Sin embargo, las tareas que podía abarcar este tipo de sistemas eran muy limitadas. En este tipo de tareas, el lenguaje que se utilizaba se definía *a priori*, solía ser de vocabulario no muy grande y por tanto se podía modelar sin problemas en una red finita, diseñada a mano.

En un primer paso, se dividen las palabras en diferentes categorías sintácticas, y se diseña la gramática de forma que contenga todas las posibles secuencias de categorías sintácticas de palabras que pueden aparecer en las frases del lenguaje. El gran inconveniente de estas gramáticas recae, en la dificultad de construirlas a mano.

Hoy por hoy y gracias a los algoritmos de inferencia es posible el uso de gramáticas estocásticas aplicadas al reconocimiento sintáctico de formas, basándose en la teoría de los lenguajes formales.

La inferencia gramatical trata de extraer una regla general a partir de ejemplos. Hay básicamente dos formas de abordar el diseño de un algoritmo de inferencia a partir de datos positivos (Segarra, 1993):

- Mediante un heurístico que calcule información relevante contenida en la muestra de aprendizaje, por ejemplo utilizando el algoritmo de "*Error Correcting Grammatical Inference*" (ECGI) (Rulot, 1992).

- Mediante métodos capaces de identificar determinadas clases de lenguajes identificables a partir de muestras positivas. Por ejemplo, el algoritmo de inferencia de los lenguajes k-Reversibles, el de los lenguajes k-Contextuales y el de los lenguajes k-Explorables en Sentido Estricto (García & Vidal, 1990).

Por otra parte se puede encontrar la inferencia de lenguajes regulares utilizando la metodología de inferencia gramatical mediante generadores mórficos, "*Morphic Generator Grammatical Inference*" (MGGI), que constituye un marco general que puede identificarse con los métodos caracterizables, al estar basada en un método de inferencia de los lenguajes locales (los lenguajes *k*-Explorables en Sentido Estricto para *k*=2) y con los métodos heurísticos ya que contienen el mecanismo para aportar en cada aplicación el conocimiento sobre la misma.

A pesar de ser consideradas como una tarea difícil, hay algunos trabajos en los que el uso de autómatas de estados finitos ha sido presentado con éxito. En (Placeway et al, 1993) se presentó la topología de los autómatas a nivel de bigramas. En (Riccardi et al, 1995) se introdujo una generalización logrando buenos resultados. Sin embargo, los autómatas propuestos, (Zhao et al., 1993) y (Riccardi et al., 1996) a nivel de trigramas, son mecanismos finitos no deterministas donde el reconocedor debe explorar varios caminos alternativos para cada palabra. En el grupo de reconocimiento automático del habla de la Universidad del País Vasco se

estudiaron los modelos de lenguaje *k*-Explorables en Sentido Estricto para ser aplicados al reconocimiento de discurso continuo, aunque el estudio se realizó sobre texto (Bordel, 96).

2.1.3.- Enfoque estadístico y/o gramatical

Los ML estadísticos se han utilizado extensivamente en RAH. Hay diversas razones para este hecho: a) por la sencillez de su formulación en contraposición con los modelos gramaticales. b) porque su obtención es muy fácil a partir de ejemplos.

A pesar de sus ventajas han recibido críticas por parte de los defensores de la metodología gramatical. La más importante es que no hay una teoría sólida y establecida detrás de esta técnica (Segarra, 1993). Además, como por motivos prácticos la memoria debía ser no excesivamente grande, sólo se realizan modelizaciones locales. No se modelaban restricciones semánticas, sintácticas o morfológicas que muchas veces trascienden más allá de la memoria del formalismo.

Los problemas principales que se le han asociado tradicionalmente a la aproximación gramatical son los siguientes: a) difícil integración con los modelos acústicos en un sistema de RAH. b) posibles dificultades en la introducción de extensiones estocásticas. c) dificultades para ser aprendidos automáticamente a partir de un corpus de ejemplos (Jelinek, 1991).

Por supuesto, hay varios trabajos que han atacado estos problemas con éxito, como se ha presentado en el apartado anterior. En especial, se pueden destacar las diferentes aproximaciones que consisten en modelar el lenguaje utilizando herramientas similares a las empleadas en los niveles acústicos, es decir, redes finitas con extensiones estocásticas (Riccarti et al, 1996) (Bordel, 1996).

Tras los últimos avances en ambas aproximaciones: estadística y gramatical, ambas son fácilmente transportables a distintas tareas y ponderan cada una de las frases candidatas. Cumplen por tanto los requisitos básicos exigidos al modelo de lenguaje.

También es posible usar Modelos combinados (estadísticos y gramaticales) con intención de explotar las ventajas de cada tipo de aproximación:

(Derouault et al, 1986): Se combinan las probabilidades que proporcionan un modelo de triPOS y una gramática regular mediante la operación producto.

(Wright et al, 1992): Se combinan bigramas, gramáticas incontextuales estocásticas e información sobre las dependencias de primer orden de las reglas de la gramática.

(Meeter et al, 1993): Se combinan modelos de N-gramas con gramáticas de contexto libre.

(Wright et al, 1994): Combina bigramas extendidos (para tener en cuenta las dependencias entre palabras a larga distancia) y una gramática de contexto libre para tener en cuenta la estructura sintáctica.

(Schmid, 1994). En este trabajo se combina información simbólica procedente de una gramática lingüística e información estocástica obtenida a partir de un modelo de lenguaje estadístico. Se obtiene un grafo de palabras a partir de un algoritmo de búsqueda de dos pasos. En primer lugar el algoritmo de Viterbi proporciona las probabilidades exactas. Después un algoritmo A*, con una búsqueda en forma de árbol usa la información gramatical para encontrar una secuencia sintácticamente correcta a través de la salida del primer decodificador.

(Kilian et al, 1995): Representan una gramática de estados finitos como si fuera un modelo de bigramas. Parten del sistema que trabaja con bigramas y no quieren modificar la estructura. Al modificar la gramática incrementan notablemente las repeticiones del léxico

(Brugadna et al, 1997): Trabaja con Modelos de lenguaje dinámicos: Se presenta un sistema para construir modelos de lenguaje dependientes del contexto, en un sistema de dialogo. Se pueden combinar: gramáticas, clases de palabras y modelos estadísticos.

(Nasr et al, 1999). Se trata de modelar mediante gramáticas, distintas clases de secuencias de palabras. Es muy probable que una combinación de palabras en el conjunto de entrenamiento aparezca relativamente pocas veces y sin embargo dentro de una clase, la probabilidad puede ser muy alta. Se trabaja con un "*tree-structured weighted finite state autómata*" para modelar las combinaciones de palabras de longitud variable y por encima se aplica un modelo de bigrama.

(Kilian et al, 1999). En esta aproximación se combinan restricciones sintácticas y estadísticas en un sólo modelo ("*Syntactical bigram with trigram*"). Trabaja con modelos de lenguaje dinámicos dependientes del contexto en un sistema de dialogo. Se trata de deletrear los nombres de ciudades alemanas. Una parte inicial de las palabras (por ejemplo, las tres primeras letras) se modelan explícitamente teniendo en cuenta las restricciones sintácticas, el resto de las palabras se generalizan utilizando n-gramas.

2.2.- Problemática de la estimación de las distribuciones de probabilidad

En la construcción de un ML uno de los principales problemas que debe resolverse es la falta de cobertura, es decir, debe estar contemplada la posibilidad de encontrar en la etapa de test o de funcionamiento normal dentro de un sistema de RAH, situaciones que no han aparecido en la etapa de aprendizaje.

La estructura del lenguaje natural -o de uno próximo- hace imposible contar con muestras de entrenamiento suficientes que permitan estimar correctamente todas las probabilidades. Además, la distribución de probabilidad de las palabras (más aún en el caso de cadenas de palabras) es tal que un número significativo de ellas aparecen tan esporádicamente que la estimación de su probabilidad no puede ser fiable estadísticamente. El problema de la falta de cobertura puede verse como un caso extremo de este fenómeno, ya que el entrenamiento

siempre será insuficiente para llegar a observar todos los fenómenos del lenguaje (concatenación de unidades léxicas o palabras).

Es un problema inherente a la construcción de un ML, independientemente que se trabaje con grandes o pequeños vocabularios y será más acuciante cuanto mayor sea el valor de N en el caso de trabajar con los N-gramas, ya que cuanto más alto sea ese valor, mayor será el número de situaciones que no son vistas en entrenamiento (peor cobertura). El modelo del 1-grama presenta cobertura del 100%. A medida que aumenta el valor de N se logra un modelo con mayor número de distribuciones de probabilidad, por lo tanto más preciso, de forma que cobertura y precisión son intereses contrapuestos.

Supongamos que *h* representa una historia, secuencia concreta de N-1 palabras y Σ representa al vocabulario de la aplicación. Entonces el conjunto de palabras para las que hay transición a partir de h, $\Sigma(h)$ ($\Sigma(h) \subset \Sigma$), será el vocabulario asociado a la historia *h*.

La estimación de la probabilidad dada por el criterio de máxima verosimilitud consiste en asignar a un determinado evento -la aparición de una determinada palabra una vez observada la historia *h*- el valor de la frecuencia de esa situación en entrenamiento. Esta frecuencia es el cociente del número de veces que se ha observado la palabra *w* tras *h*, N(*w*/*h*), entre el número total de unidades que se han observado tras la historia previa *h,* N(*h*).

En general dada la historia *h*, se pueden distinguir tres situaciones (ver figura 2.2):

1. - Cuentas N(*w*/*h*) altas que darán lugar a estimaciones que pueden considerarse fiables.
2. - Cuentas N(*w*/*h*) bajas que pueden dar lugar a estimaciones de probabilidad poco fiables
3. - Palabras $w \in \Sigma$ no vistas tras la historia *h*

La falta de cobertura consiste en que si en el uso del ML, a partir de una historia observada *h*, la próxima palabra no es una de las pertenecientes a $\Sigma(h)$, la expresión de la probabilidad por el criterio de máxima verosimilitud daría una probabilidad cero. Para resolver este problema se utilizan diferentes estrategias que van desde heurísticos "ad-hoc" hasta recetas estrictamente bayesianas.

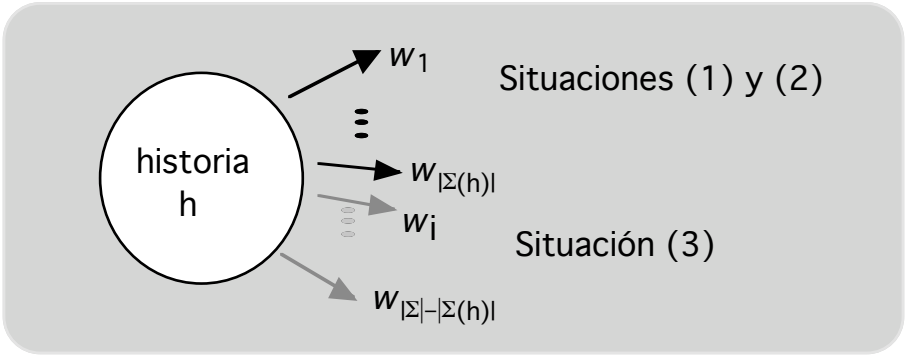


Figura 2.2.- Representa las transiciones a partir de un nodo que representa una determinada historia h, mediante arcos para las transiciones distintas vistas $|\Sigma(h)|$ y no vistas $|\Sigma|-|\Sigma(h)|$ en el entrenamiento.

La estimación de la distribución de probabilidad en la situación 1 es la más adecuada, pero las situaciones 2 y 3 son inevitables. La situación 2 da lugar a una asignación de probabilidades inadecuadas, sin embargo el problema más apremiante es la situación 3 ya que un modelo que no les de solución puede verse ante sentencias a las que es incapaz de asignar una probabilidad. Una forma de salvar esta situación consiste en reservar una cierta cantidad de probabilidad para los arcos inexistentes, de forma que las estimaciones de los casos 1 y 2 se vean "disminuidas" de alguna forma que es necesario determinar. Una vez hecho esto, hay que especificar también como repartir esta probabilidad entre las unidades no vistas tras la historia *h*.

En resumen hay dos cuestiones a las que hay que dar solución. A) ¿qué cantidad se debe descontar a cada arco visto?, B) ¿cómo repartir la probabilidad descontada entre los arcos no vistos?.

La solución a la situación 2, que de momento habíamos considerado menos apremiante, suele relacionarse con la pregunta A), ya que un buen método de descuento contemplara de alguna manera el nivel de fiabilidad de las estimaciones. En el siguiente punto se presentan las técnicas de suavizado más significativas relatadas en la literatura: el método de back-off y el de interpolación. Se llaman técnicas de suavizado porque tienen el efecto mencionado de "suavizar" la distribución de probabilidad total del modelo.

En el caso de estar realizando el análisis de un texto escrito, la certeza sobre la identidad de las palabras nos asegura que estamos ante un problema de falta de cobertura del ML. Por el contrario, en el caso de analizar las unidades-hipótesis dadas por un decodificador acústico no se dispone de certeza absoluta sobre la identidad de la palabra, pudiendo estar también ante un error del propio decodificador. Podría plantearse un análisis según un modelo de error que considere posibles sustituciones, borrados e inserciones además de considerar la falta de cobertura. Esto implicaría la perdida del determinismo en el modelo con la consiguiente necesidad de aplicación de técnicas más costosas para el análisis de las frases.

2.2.1.- Métodos convencionales para el suavizado

Para estimar las probabilidades de un modelo se debe establecer un "descuento" en los valores que se obtendrían para las probabilidades de los eventos vistos en la fase de aprendizaje. El método más sencillo que se puede encontrar es el de establecer el descuento con un único parámetro para todos las historias del modelo. Este parámetro se fija tras un estudio de la calidad obtenida en función del valor fijado. Este proceder implica el disponer de un conjunto de muestras para entrenar el modelo y otro para estimar el descuento. Por extensión del método anterior, se puede establecer un parámetro diferente para cada historia h observada, lo que lleva a un mayor esfuerzo a la hora de estimación de los parámetros, haciéndose necesaria la utilización de algoritmos de estimación de parámetros (tipo Viterbi o Baum-Welch).

Una vez realizado el descuento, el reparto de dicha cantidad entre los elementos no vistos puede realizarse de forma uniforme (suavizado plano). Pero, normalmente la obtención de un

buen suavizado pasa en general por la utilización de otros modelos que aporten diferentes distribuciones de probabilidad.

Los métodos de suavizado más comúnmente utilizados son los de back-off e interpolación. Las diferencias entre ambos métodos recae fundamentalmente en los desarrollos matemáticos necesarios para obtener los modelos, proporcionando resultados muy similares (Ney et al, 97). Para realizar el suavizado la interpolación no lineal, a diferencia de la interpolación lineal, pretende tener en cuenta la "calidad" de los valores aprendidos. Por el otro lado, también el método de back-off realiza el descuento en función de la fiabilidad de los parámetros aprendidos. Se basa en una estimación alternativa a la Máxima Verosimilitud mediante la aplicación de la formula de Turing. Esta fórmula determina tanto la cantidad total a descontar como la forma en que este descuento afecta a cada evento visto.

2.2.1.1.- Interpolación

Lineal

En (Jelinek & Mercer, 1980) se propone un método de suavizado para ser aplicado a los Modelos Ocultos de Markov que posteriormente se ha aplicado a los ML (Jelinek, 1985) y (Ney & Essen, 1991).

El valor de la probabilidad del modelo suavizado $P^S(w/h)$ se obtiene por combinación lineal de la probabilidad del propio modelo a suavizar $P(w/h)$ y el suavizador $\beta(w/h^*)$ (distribución de probabilidad más general, h^* representa una historia de palabras de longitud inferior a h).

$$P^s(w/h) = \lambda P(w/h) + (1-\lambda)\beta(w/h^*) \qquad (2.5)$$

Con λ próximo a uno se considera que el valor de $P(w/h)$ es fiable y si es próximo a cero será poco fiable. Los dos modelos que forman parte de la interpolación se pueden entrenar con las mismas muestras. Sin embargo, deben ser diferentes a las utilizadas para ajustar los valores de λ. Para disponer de todo el conjunto de entrenamiento, tanto para estimar $P(w/h)$ y $\beta(w/h^*)$ como λ puede utilizarse una técnica de validación cruzada o leaving-k-out realizándose el entrenamiento mediante el algoritmo fordward-backward (Baum-Welch) (Ney & Essen, 1993).

Aunque la formulación parezca a primera vista sencilla, no es trivial llevarlo a cabo con modelos cualesquiera. El modelo suavizador debe ser tal que una vez observada la historia h en el modelo a suavizar, siempre se considere la misma historia h^* en el modelo suavizador. Si no es así, no podríamos disponer de los parámetros λ, sino que habría que establecer un parámetro diferente para cada par de historias observadas en cada modelo. Lo que sí que ocurrirá en general es que varias historias del modelo a suavizar se corresponderán con la misma historia en el modelo suavizador.

En (Jelinek, 1985) se presenta una modificación del fordward-backward que incluye en sí mismo la técnica del *leaving-k-out* y que toma el nombre de *deleted interpolation*. Este algoritmo permite llevar a cabo el entrenamiento de todos los parámetros simultáneamente.

No lineal

Parece aconsejable que la contribución al descuento de los eventos con cuentas más altas sea menor que las aportadas por los eventos con cuentas menores. Esto es así porque la probabilidad obtenida por el criterio de Máxima Verosimilitud es más correcta cuanto mayor sea la cuenta. De este hecho surge la idea de aplicar un descuento no lineal.

Una posible interpolación no lineal es aquella en la que el descuento es constante (y por tanto porcentualmente menos significativo cuanto mayor es la cuenta de un evento). La cantidad total descontada sobre los eventos vistos tras la historia *h* se reparte entre todos los eventos incluyendo a los no vistos:

$$P^s(w/h) = max\left\{\frac{N(w/h) - d}{N(h)}, 0\right\} + \frac{dn_+(h)}{N(h)}\beta(w/h^*) \qquad (2.6)$$

Siendo N(*w*/*h*) el número de veces que aparece *w* tras la historia *h*, N(*h*) el número de eventos totales que se observan tras la historia *h* y $n_+(h)$ el número de eventos distintos (*h*,*w*) en el conjunto de entrenamiento.

Este método implica la determinación de un único parámetro *d*, y con la ventaja de actuar de forma más especifica sobre los eventos "supuestamente" peor entrenados. Un valor *d*<1 puede interpretarse como una corrección al ruido de discretización en la estimación de las cuentas discretas N(*w*/*h*). Para valores mayores o iguales a uno será necesario modificar el segundo término de la ecuación 2.6 ya que el factor $dn_+(h)/N(h)$ es la probabilidad total descontada si *d*<1, caso en que la función max() solo afecta a los eventos no vistos.

Modulada por la calidad

En (Bordel et al, 1995) se presenta una nueva propuesta conocida como (QWI) *Quality weighted Interpolation*- Interpolación modulada por la calidad. La propuesta consiste en realizar en cada paso una interpolación lineal de los valores de probabilidad dados por los diferentes modelos, utilizando como factores de peso unos valores proporcionales a la calidad de los mismos. De acuerdo con este método, dados dos modelos M1 y M2, la probabilidad de la palabra w se calculará como:

$$P^s(w/h) = \mu P_1(w/\alpha) + (1-\mu)P_2(w/\beta) \qquad (2.7)$$

donde P_1 y P_2 son las probabilidades asignadas a los modelos M1 y M2 respectivamente y α y β representan la información de cada uno de los modelos respecto a la historia *h*. Los detalles se pueden encontrar en (Bordel, 1996).

2.2.1.2.- Back-off

El método de Back-off fue introducido por (Katz, 1987) y estaba planteado para modelos de N-gramas en base a dos ideas principalmente: la utilización de una estimación alternativa a la de Máxima Verosimilitud que sumara una probabilidad total inferior a la unidad para todos los N-gramas vistos (formula de Turing) y la distribución de la cantidad de probabilidad restante a los eventos no vistos de forma recursiva mediante el modelo inmediatamente inferior ((N-1) gramas).

El valor de la probabilidad del modelo suavizado $P^S(w/h)$ se calcula a partir de la probabilidad del propio modelo a suavizar $P(w/h)=N(w/h)/N(h)$ y el suavizador $\beta(w/h^*)=N(w/h^*)/N(h^*)$ (distribución de probabilidad más general, donde h^* representa una historia de palabras de longitud inferior a h), teniendo en cuenta que $N(w/h)$ es el número de veces que se observa la historia h tras la palabra w y, N(h) es el número de eventos distintos que se observan tras la historia h, es decir $N(h) = \sum_{\forall w' / N(w'/h)>0} N(w'/h)$:

$$p^s(w/h) = \begin{cases} \left[1-\lambda\right]\dfrac{N(w/h)}{N(h)} & N(w/h) > 0 \\ \left(\sum_{\forall w' / N(w'/h)>0}\left[\lambda\dfrac{N(w'/h)}{N(h)}\right]\right)\dfrac{\beta(w/h^*)}{\sum_{\forall w' / N(w/h)=0}\beta(w'/h^*)} & N(w/h) = 0 \end{cases} \quad (2.8)$$

Se sustrae una cierta cantidad de la probabilidad dada por el criterio de Máxima Verosimilitud de la probabilidad $P(w/h)=N(w/h)/N(h)$ dependiendo del valor de un parámetro λ que a su vez puede depender del número de veces que se haya observado la palabra w tras la secuencia h, $N(w/h)$. La probabilidad sobre los eventos no vistos se obtiene repartiendo la probabilidad que se ha sustraído a los eventos vistos en función de $\beta(w/h^*)$ (probabilidad del modelo suavizador, si h está etiquetado como $w_{i-(N-1)}^{i-1}$ entonces $h^*= w_{i-(N-2)}^{i-1}$).

El modelo inferior $\beta(w/h^*)$, del que se tiene en cuenta la probabilidad llamada de back-off, debe asignar probabilidad a todas las unidades, bien por ser un modelo base con cobertura del monoide libre o por haber sido suavizado previamente.

La probabilidad para sucesos que no se encuentran en el corpus de entrenamiento deben de descontarse de los sucesos que probablemente estén mal estimados, es decir, de aquellos que aparecen pocas veces en las muestras de entrenamiento. La técnica de suavizado por back-off se estudia en el capítulo 3 aplicada a los modelos de lenguaje *k*-Explorables un sentido Estricto y en mayor profundidad en el capítulo 5.

2.3.- La perplejidad como medida de evaluación de los modelos de lenguaje

La construcción de un ML supone la generalización del conjunto de muestras de entrenamiento extrayendo las regularidades que se detecten, de forma que también se aceptan

como frases pertenecientes al ML, frases no vistas en entrenamiento. Por tanto la "calidad" del modelo depende en gran medida del tipo de generalización que realice. Una buena generalización conseguirá incluir prácticamente todas las frases del lenguaje (buena cobertura) y no considerará validas otras que no pertenecen a él (no sobregeneralizará).

Siendo totalmente estrictos la evaluación de un ML debe venir necesariamente dada en términos de su efectividad en la ejecución de la tarea para la que ha sido diseñado. En el caso del RAH hay que disponer de todas las funciones necesarias. Esto supone una dependencia de los resultados obtenidos en función del resto del sistema, con lo que la evaluación presenta una dificultad añadida. Sin embargo, si se pretende comparar variantes del ML o del suavizado aplicado, resultará el método más adecuado manteniendo fijos el resto de parámetros del sistema.

Sin embargo, como primera aproximación se recurre a la Teoría de la Información (Shannon, 1964). Se trata de aislar los dos tipos de probabilidades que aparecen en la regla de Bayes y obtener la probabilidad que el ML asigna a una secuencia de palabras $P(\Omega \equiv w_1,...w_L)$. Al normalizar esta probabilidad a priori con respecto al número L de palabras, se obtiene la inversa de la raíz L-ésima y de ese modo se obtiene la llamada perplejidad del corpus (Bahl, 1993).

$$PP = \left(P(\Omega \equiv w_1^L\right)^{\frac{-1}{L}} \tag{2.9}$$

Si insertamos la descomposición en las probabilidades condicionales en la ecuación anterior y obtenemos el logaritmo se obtiene:

$$\log PP = \frac{-1}{L}\sum_{i=1}^{L}\log\left(P(w_i / w_{i-(n-1)}^{i})\right) \text{ con n=max\{N,i\}} \tag{2.10}$$

que corresponde a la llamada entropía H (proveniente de la mecánica estadística). Por tanto la perplejidad (PP) es:

$$PP = 2^{H} = 2^{\frac{-1}{L}\sum_{i=1}^{N}\log_2(p(w_i))} \tag{2.11}$$

Este es el valor que se suele aparecer en la literatura para dar la medida de la calidad de los modelos (Jelinek, 1977).

Este valor resulta ser el inverso de la media geométrica de las probabilidades que el modelo asigna a los elementos del conjunto de la muestra de entrenamiento.

$$PP = 2^{\frac{-1}{L}\sum_{i=1}^{L}\log_2(p(w_i))} = 2^{\frac{-1}{L}\log_2(\prod_{i=1}^{L}p(w_i))} = 2^{\log_2(\prod_{i=1}^{L}p(w_i))^{\frac{-1}{L}}} = \frac{1}{\sqrt[L]{\prod_{i=1}^{L}p(w_i)}} \tag{2.12}$$

Tenemos en definitiva, que el término empleado para medir la "calidad" de los modelos en experimentos comparativos, la perplejidad, es un valor que nos da una idea del "factor de ramificación medio". En caso de que todos los elementos sean equiprobables presentará el valor

máximo, que coincide con la talla del vocabulario (en cada punto del análisis cualquier palabra es igualmente probable y por tanto las ramificaciones "más probables" son todas).

La ecuación 2.12 muestra que la perplejidad es la media geométrica de la probabilidad recíproca sobre las N palabras. Excepto por el factor -1/L, para un corpus dado, la perplejidad es idéntica a la probabilidad por máxima verosimilitud. Por tanto minimizar la perplejidad del corpus es lo mismo que maximizar la probabilidad acumulada, lo cual se corresponde con el método clásico de estimación estadística (Duda, 1973).

Como primera aproximación se dice que la perplejidad mide la dificultad de la tarea de reconocimiento ya que mide el grado de limitaciones expresadas por el ML. Desde el punto de vista del reconocimiento de tareas, se puede decir que el modelo de lenguaje reduce el número de elecciones de distintas palabras durante el proceso de reconocimiento. En general, cuanto más baja es la perplejidad mejor reconocimiento (menor porcentaje de error). Por ejemplo, dependiendo de la aplicación y del ML, un sistema de reconocimiento con un vocabulario de 1000 palabras puede tener un grado de restricciones tan grande que la tarea de reconocimiento sea más fácil que la de los dígitos. Hay un aspecto muy importante en la definición de la perplejidad: si a una palabra el ML se le asigna una probabilidad cero, la perplejidad será infinitamente alta. Esa es una de las características del ML: la predicción de la próxima palabra debe ser tan buena como sea posible, sin excluir ninguna de las palabras del vocabulario.

Dentro de la comunidad científica internacional es el parámetro de medida de los MLs más comúnmente utilizado pero existen cada vez más trabajos en los que se logra un modelo de lenguaje mucho mejor que el modelo base evaluado en perplejidad y que a la hora de ser trasladado al sistema de reconocimiento proporciona peores resultados. Por ejemplo, dentro de los últimos trabajos presentados, estos resultados se pueden encontrar en (Gilde & Hofmam, 1999), (Adda et al, 1999).

En (Codogno et al, 1987) se presenta un ejemplo de un ML basado en segmentos de voz (*parts of speech*), que tiene una perplejidad mayor que un modelo de trigramas pero logra un mejor porcentaje de reconocimiento en un test pronunciado por un sólo locutor.

En (Samuelsson & Reichl, 1999) se presenta un ejemplo en el que la reducción de perplejidad no es muy significativa pero si la disminución del error del sistema (además el ML se introduce directamente sin sacar las N-mejores).

En (Rosenfeld, 1994) cuando se habla de la relación entre la perplejidad y el porcentaje de error del sistema de RAH se dice que sus experimentos daban una reducción de la perplejidad del 32% y sin embargo la reducción del error fue del 14% en las condiciones más favorables. Lo que se observa experimentalmente les confirma, la regla de la "raíz cuadrada", que dice que la mejora del error es aproximadamente la raíz cuadrada de la mejora de la perplejidad. La pregunta que se plantea inmediatamente, es ¿porqué el impacto es tan pequeño?.

Se podría decir que la perplejidad presenta algunas deficiencias, principalmente dos:

a) Por un lado, la perplejidad es una medida dependiente del corpus del que se obtiene el modelo. No es una evaluación del tipo de modelo generado, sino de un representante concreto de dicho tipo. La comparación de la efectividad de dos métodos de modelización mediante la obtención de perplejidades pasa por la utilización de los mismos conjuntos de aprendizaje y de prueba. No hacerlo así da lugar a que en la literatura se encuentren valores de perplejidad que difieren en varios ordenes de magnitud, ya que las comparaciones entre unos métodos y otros no sean del todo realistas.

b) Por otra parte, no debe olvidarse que es incorrecto identificar la capacidad de obtener un modelo para un lenguaje con la capacidad de reconocimiento automático de dicho lenguaje. El uso de la perplejidad como medida de la dificultad de una tarea de Reconocimiento es incorrecta, ya que la evaluación se lleva a cabo frente a un conjunto de prueba que no ha sufrido las deformaciones que introduce el procesador acústico. La distribución de probabilidad del conjunto de frases obtenidas por aplicación de un procesamiento acústico presentará algunas diferencias sistemáticas con la distribución de probabilidad real del lenguaje debidas a la estructura fonética de dicho lenguaje. Por tanto la dificultad de la tarea de reconocimiento tiene que ver con la dificultad de "neutralizar" dichas diferencias, lo que no queda correctamente reflejado en caso de utilizar un conjunto de prueba de texto. En (Rubio et al, 1993) se muestra gráficamente como, para diferentes tareas de reconocimiento del discurso continuo, la correlación entre la perplejidad del ML y las tasas de Error es deficiente. En dicho trabajo se define una medida de la redistribución de probabilidad impuesta por el modelo acústico para ser incluida también en una nueva expresión modificada para la perplejidad. Esto lleva finalmente a lo que denominan RWTD (*Real Weighted Task Dificulty*), medida que presenta una correlación con las tasas de error aparentemente muy superior a la dada por la perplejidad.

En general, se dice que la perplejidad es una buena medida para predecir el comportamiento de un modelo de lenguaje "per se", pero si lo que se pretende es medir la contribución del modelo de lenguaje al decodificador de voz, hay que tener en cuenta que la perplejidad no tiene en cuenta de ninguna manera la relación con el modelo acústico. Por tanto no puede asumirse directamente, que en general, un decremento en perplejidad resulte necesariamente en una mejor eficiencia del sistema. Nuevas medidas como la aperplejidad descrita en (Jelinek, 1986) tratan de tener en cuenta la interacción entre los modelos acústicos y de lenguaje. La aperplejidad es la perplejidad media dada por un ML cuando su obtención se limita a un subconjunto del vocabulario (obtenido de los conjuntos de palabras que son acústicamente más similares a la palabra correcta). Los resultados que se presentaron para una base de datos italiana (6,000 palabras) sugieren que si se asume que la aperplejidad mide suficientemente bien la interacción entre los modelos acústicos y de lenguaje, en ese caso concreto, la interacción debía ser prácticamente irrelevante y la perplejidad era tan buena como la aperplejidad.

En (Ferreti, 1990) se propone una nueva medida llamada SDE (*Speech Decoder Entropy*). Se dice que es mas cara computacionalmente hablando que las medidas de perplejidad

y aperplejidad pero proporciona una mejor predicción de la interacción entre los modelos acústicos y de lenguaje. Dicen que es particularmente útil para comparar modelos de lenguaje con estructuras muy diferentes.

En (Ito et al, 1999) se dice que la perplejidad puede resultar útil para la evaluación de pequeñas tareas, pero a medida que se va ampliando el número de datos cada vez va peor. La nueva propuesta se basa en medir la diferencia en probabilidad del ML en texto y el de la palabra que de la máxima probabilidad en el contexto. La medida necesita unos umbrales que se obtienen en el propio proceso de decodificación.

En (Clarkson et al, 1999) se presenta un nuevo método basado en tener en cuenta la probabilidad sobre el resto de palabras diferentes a la verdadera que pueden darse en el proceso de decodificación. Se parte de la construcción y posterior evaluación en perplejidad y error de decodificación para 50 modelos de lenguaje. Se probaron distintas alternativas: perplejidad, entropía, *rank* (se evalúa el ML de acuerdo a la porción de palabras que tienen una probabilidad más alta que la palabra que aparece realmente en el test). Se concluye que habría que combinar la información de las fuentes que en cierto sentido es ortogonal. Se usa la interpolación lineal para esa combinación.

Capítulo 3

Modelo de lenguaje k-Explorable en Sentido Estricto (k-EE)

Los modelos de lenguaje se suelen clasificar en dos categorías: modelos estadísticos y gramaticales. Los modelos de lenguaje estadísticos se basan en la estimación de las probabilidades de observar las N-1 palabras anteriores a una dada y reciben el nombre de N-gramas. Los modelos gramaticales, recogen en su estructura la naturaleza sintáctica del lenguaje, lo cual permite modelar mejor el lenguaje pero su uso no está muy extendido ya que se suele asociar con costes computacionales excesivamente altos. Además, durante muchos años, sólo se podían integrar en los sistemas de reconocimiento gramáticas regulares construidas a mano (Segarra, 1993) porque no estaba resuelto el aprendizaje del modelo a partir de muestras de entrenamiento.

Por otro lado, existen varios trabajos en los que se representan los modelos de lenguaje estadísticos mediante autómatas estocásticos que se integran en el sistema de reconocimiento (Riccardi et al, 1996) (Bonafonte & Mariño 1998) (Suzuki & Aso, 1999). Ésto es debido a que esta aproximación se puede integrar muy fácilmente junto con los modelos acústicos en una única red finita.

Este trabajo se enmarca dentro del formalismo gramatical y el principal objetivo consiste en mostrar la viabilidad de la integración de los ML *k*-EE junto con los modelos acústicos en un sistema de RAH (Varona & Torres, 1999a). Los modelos de lenguaje *k*-EE se pueden inferir a partir de muestras de entrenamiento mediante un algoritmo de inferencia (García & Vidal, 1990). El uso de gramáticas *k*-EE permite obtener un Autómata de Estados Finitos y Estocástico asociado a cada modelo *k*-EE (AEFE *k*-EE) (Bordel, 1996) (Bordel et al, 1997) que, a diferencia de los propuestos en (Zhao et al, 1993) y (Riccardi et al, 1996), es una red finita determinista. Este autómata debe establecer una probabilidad para toda combinación de *k* palabras pero no es posible obtenerlas todas a partir del conjunto de entrenamiento por grandes que sean las bases de datos utilizadas. Como consecuencia se hace necesario el uso de una técnica de suavizado que sustraiga una cierta cantidad de probabilidad de las combinaciones de palabras vistas en entrenamiento y la reparta entre las combinaciones no vistas.

En este capítulo, en primer lugar se presenta la definición e inferencia de los lenguajes *k*-EE. A continuación se desarrollan los ML *k*-EE Suavizados capaces de otorgar una probabilidad

para cualquier combinación de *k* palabras pertenecientes a la tarea representada por el ML. El resultado será un único modelo suavizado, desarrollado mediante un esquema de construcción recursivo, que integra K modelos *k*-EE (*k*=1...K) y la técnica de suavizado (back-off bajo el formalismo sintáctico). En tercer lugar, se muestra una representación eficiente del modelo integrado, en una estructura compacta de tipo array en la que sólo están representadas de forma explícita las transiciones vistas en entrenamiento. Finalmente se muestran los resultados obtenidos al evaluar los modelos integrados de orden K (K=1...6) externamente al sistema de reconocimiento (perplejidad y memoria requerida para su utilización).

3.1.- Definición e inferencia de los lenguajes k-EE

Dado un alfabeto Σ, un lenguaje *k*-EE $L_{k-EE}(\sum, I_k, F_k, T_k)$ es un subconjunto de Σ* formado por todas las cadenas con segmentos iniciales pertenecientes a un determinado conjunto I_k, segmentos finales pertenecientes a otro conjunto F_k y que no contienen ningún segmento de un tercer conjunto T_k. Los elementos de los conjuntos F_k e I_k son cadenas de elementos de Σ de longitud menor de *k*, y los elementos del conjunto T_k son cadenas de elementos de Σ de longitud *k* (Zalcstein, 1972). Un lenguaje *k*-EE está definido por la siguiente expresión regular (García & Vidal, 1990) (Segarra, 1993):

$$L_{k-EE}(\sum, I_k, F_k, T_k) \cong (I_k \sum^* \cap \sum^* F_k) - \sum^* T_k \sum^* \tag{3.1}$$

donde Σ es el alfabeto $I_k, F_k \subseteq \bigcup_{i=1} \sum^i$, $T_k \subseteq \sum^k$, siendo Σ^i una cadena de elementos de Σ de longitud *i*.

En (García & Vidal, 1990) se demuestra que la clase de los lenguajes *k*-Explorables en Sentido Estricto es identificable a partir de un conjunto de muestras positivas. Dado un conjunto de datos positivos R^+ (muestra de aprendizaje) sobre un determinado alfabeto Σ, para cada $k \geq 1$, es posible obtener una cuádrupla $Z_k(R^+)=(\Sigma_k(R^+), I_k(R^+), F_k(R^+), T_k(R^+))$ que define el menor lenguaje *k*-EE que contiene a R^+. $Z_k(R^+)$ se obtiene de R^+ como sigue:

$\Sigma_k(R^+)$ es el conjunto de símbolos que ocurren en las frases de R^+

$$I_k(R^+)=\{u|uv \in R^+, |u|=k\text{-}1, v \in \Sigma_k(R^+)^*\} \cup \{x \in R^+, |x|<k\text{-}1\} \tag{3.2}$$

(contiene todos los segmentos iniciales de a lo sumo *k*-1 de las muestras)

$$F_k(R^+)=\{v|uv \in R^+, |v|=k\text{-}1, u \in \Sigma_k(R^+)^*\} \cup \{x \in R^+, |x|<k\text{-}1\} \tag{3.3}$$

(contiene todos los segmentos finales de a lo sumo *k*-1 de las muestras)

$$T'_k(R^+)=\{v|uvw \in R^+, |v|=k, u, w \in \Sigma_k(R^+)^*\} \tag{3.4}$$

(contiene todos los segmentos de longitud *k* de la muestra)

$$T_k(R^+)= \Sigma_k (R^+)^k - T'_k(R^+) \tag{3.5}$$

Además, dada una muestra de entrenamiento, valores crecientes de *k* producen cada vez lenguajes más restringidos $L_{k\text{-EE}} \subseteq L_{(k\text{-}1)\text{-EE}} \subseteq ... \subseteq L_{1\text{-EE}}$. Vamos a ver a continuación un ejemplo de inferencia de lenguajes *k*-EE (*k*=1...3) a partir de un conjunto de muestras positivas $R^+ \equiv$ {aba, aabbaa, abbaaa}. En la figura 3.1 se muestran los grafos asociados a cada lenguaje *k*-EE inferido. Los nodos se etiquetan con las cadenas de R^+ de longitud hasta *k*-1. Los arcos se etiquetan con los símbolos del alfabeto que aparecen en el corpus de entrenamiento al final de la cadena representada en el nodo fuente. El estado inicial se etiqueta con el símbolo *nil* y los estados finales se representan con un doble círculo.

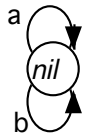
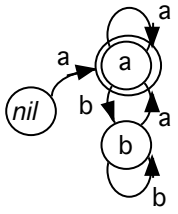
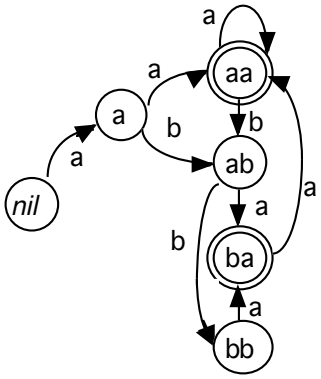
Para *k*=1: I_1, segmentos iniciales de longitud 0: Ø F_1, segmentos finales de longitud 0: Ø T_1, segmentos prohibidos de longitud 1: Ø $L_{1_EE(R^+)} \equiv (\{a,b\}, Ø, Ø, Ø) \equiv (a\|b)^* \equiv \Sigma^*$	
Para *k*=2: I_2, segmentos iniciales de longitud 1: {a} F_2, segmentos finales de longitud 1: {a} T_2, segmentos prohibidos de longitud 2: Ø $L_{2_EE(R^+)} \equiv (\{a,b\}, \{a\}, \{a\}, Ø) \equiv a^+(b^*a^+)^*$	
Para *k*=3: I_3, segmentos iniciales de longitud 1 y 2: {a, aa, ab} F_3, segmentos finales de longitud 1 y 2: {a, aa, ba} T_3, segmentos prohibidos de longitud 3: {bab, bbb} $L_{3_EE(R^+)} \equiv (\{a,b\}, \{aa, ab\}, \{aa, ba\}, \{bab, bbb\}) \equiv$ $(a\|ab\|abb)a^+((b\|bb)a^+)^*$	

Figura 3.1.- $L_{1\text{-EE}}$, $L_{2\text{-EE}}$, $L_{3\text{-EE}}$ inferidos a partir de un conjunto de muestras positivas R^+={aba, aabbaa, abbaaa} y alfabeto Σ={a,b}; I_k son los segmentos iniciales de longitud hasta *k*-1, F_k son los segmentos finales de longitud hasta *k*-1. El modelo *k*-EE inferido está contenido en *k*-1, $L_{3\text{-EE}} \subseteq L_{2\text{-EE}} \subseteq .L_{1\text{-EE.}}$ El $L_{1\text{-EE}}$ es igual a Σ^*. $L_{5\text{-EE}}$ consiste exactamente en las tres muestras.

En la figura 3.1 se ve cómo a medida que aumenta el valor de *k* el lenguaje inferido es cada vez más reducido, como consecuencia la capacidad de generalización se reduce a medida que aumenta el valor de *k*. Con este ejemplo para *k*=5 el modelos identifica exclusivamente las tres frases del conjunto de entrenamiento. El modelo k-EE inferido está contenido en el (k-1)-EE: $L_{3\text{-EE}} \subseteq L_{2\text{-EE}} \subseteq .L_{1\text{-EE.}}$

Aplicando el algoritmo de inferencia (García & Vidal, 1999) se obtiene un autómata de estados finitos determinista que acepta el menor lenguaje *k*-EE representado en la muestra de entrenamiento. Pero, para su aplicación a la modelización del lenguaje, en (Segarra, 1993) se

presenta una extensión estocástica. El autómata estocástico contiene un conjunto de probabilidades asociadas a las transiciones entre estados que se estiman a partir de la frecuencia relativa de ocurrencia de las transiciones en las muestras de aprendizaje.

3.1.1.- El Autómata de Estados Finitos Estocástico k-EE (AEFE *k*-EE)

El autómata determinista se obtiene directamente del conjunto de muestras de entrenamiento y queda definido por la quíntupla $(\Sigma, Q^k, \delta^k, q_0, F)$ donde:

- $\Sigma = \{w_j\}, j = 1...|\Sigma|$, es el vocabulario de la aplicación o lo que es lo mismo el conjunto de palabras que aparecen en el corpus de entrenamiento. Su tamaño se representa como $|\Sigma|$.

- Q^k es el conjunto de estados del autómata asociado a un modelo de orden k. Cada estado representa una cadena de palabras $w_{i-(k-1)}...w_{i-1}$, de longitud k-1, donde i es un índice cualquiera de la secuencia $w_1...w_i...$ que aparece en el corpus de entrenamiento. Este estado se etiqueta como $w_{i-(k-1)}^{i-1}$.

Los estados que representan subcadenas iniciales de longitud inferior a k-1 pertenecientes al conjunto I_k del $L_{k\text{-EE}}$ correspondiente se etiquetan como w_{i-j}^{i-1}, donde j=1...k-2. El valor j=1 se asocia a los $|\Sigma|$ estados etiquetados con una única palabra w_j=1,..., $|\Sigma|$.

- F es el conjunto de estados finales del autómata y cada estado representa una subcadena de longitud k-1. Estas subcadenas se corresponden con las subcadenas finales de longitud k-1, que pertenecen al conjunto F_k en el $L_{k\text{-EE}}$ correspondiente.

- δ^k es la función de transición δ^k: $Q^k \times \Sigma \rightarrow Q^k \times [0...1]$. $\delta^k(q, w_i)=(q_d, P(w_i/q)$ define un estado destino $q_d \in Q^k$ y una probabilidad $P(w_i/q) \in [0...1]$ de transición por la palabra w_i a partir del estado q, que será asignada a cada elemento $(q, w_i) \in Q^k \times \Sigma$. Cada transición entre dos estados representa un cadena de k palabras siendo la etiqueta del estado origen la cadena de las (k-1) primeras palabras $w_{i-(k-1)}^{i-1}$ y la del estado destino la cadena de las (k-1) últimas palabras $w_{i-(k-1)+1}^{i}$ (ver figura 3.2b).

$$\delta^k(w_{i-(k-1)}^{i-1}, w_i) = (w_{i-(k-1)+1}^{i}, P(w_i / w_{i-(k-1)}^{i-1})) \tag{3.6}$$

Casos especiales aparecen al comienzo de las secuencias de palabras porque se parte del estado inicial etiquetado por *nil,* que corresponde a una secuencia vacía de palabras (ver figura 3.2a), para el que se definen $|\Sigma|$ transiciones y cuyos estados destino están etiquetados con subcadenas de longitud 1.

$$\delta^k(nil, w_j) = (w_j, P(w_j / nil)) = (w_j, P(w_j)) \quad j = 1,..., |\sum| \tag{3.7}$$

Y en las transiciones entre las subcadenas iniciales de longitud inferior a k-1 representadas en los estados w_{i-j}^{i-1} donde j=1...k-2.

$$\delta^k(w_{i-j}^{i-1}, w_i) = (w_{i-j}^{i}, P(w_i / w_{i-j}^{i-1})) \text{ con j=1..}k\text{–2} \tag{3.8}$$

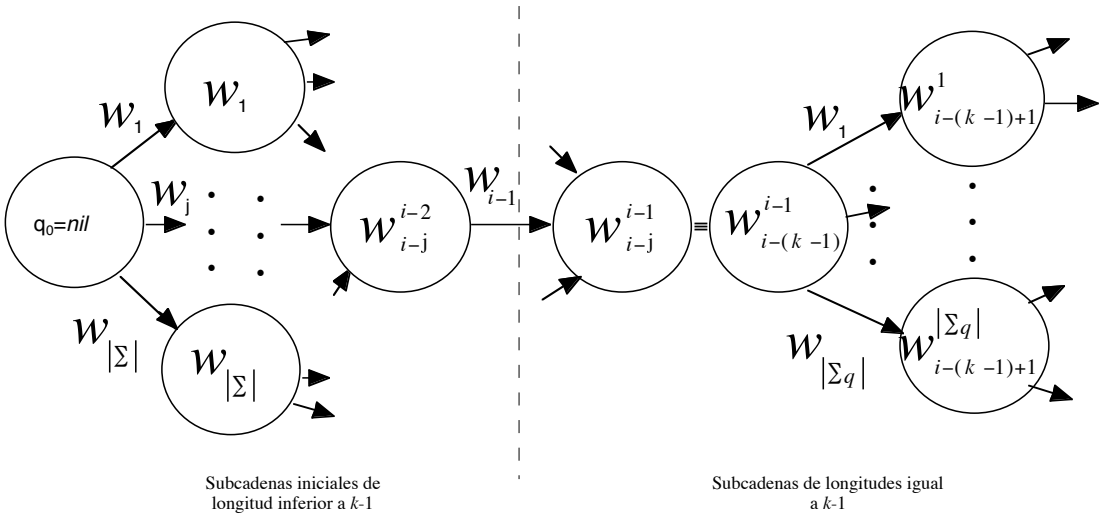


Figura 3.2.- a) Transiciones a partir del estado inicial etiquetado como *nil* y entre estados representando secuencias de palabras de longitud inferior a *k-1 (*j=1...*k-2*). b) Transiciones entre estados representando secuencias de palabras de longitud *k-1*.

La condición de contorno estocástica que se debe cumplir es la siguiente:

$$\sum_{\forall w_j \in \Sigma} P(w_j / q) = 1 \quad \forall q \in Q \tag{3.9}$$

Bajo este formalismo el modelo de 1-EE consiste en un único estado, el etiquetado por *nil*. Este estado es el origen y el destino de un número de transiciones igual al número de palabras del vocabulario y la probabilidad asociada a cada transición representa la frecuencia de aparición de la palabra correspondiente en el corpus de entrenamiento.

$$\delta^{k=1}(nil,\mathrm{w_j})=(nil,\mathrm{P(w_j/}nil\mathrm{)})=(nil,\mathrm{P(w_j)})\ \forall \mathrm{w_j}\in\Sigma \tag{3.10}$$

Dada la naturaleza determinista y como consecuencia no ambigua del autómata, sólo se alcanza un estado final para cada frase. La probabilidad de la sentencia de $\Omega \equiv \mathrm{w_1...w_L}$ de longitud L, es decir, la probabilidad de que Ω sea aceptada por el autómata, se obtiene como el producto de la probabilidad asignada a las transiciones utilizadas para alcanzar dicho estado final:

$$P(\Omega) = P(w_1 / nil)P(w_2 / w_1)...P(w_{k-1} / w_1^{k-2})\prod_{i=k}^{L} P(w_i / w_{i-(k-1)}^{i-1}) \tag{3.11}$$

La no ambigüedad del autómata permite obtener una simple estimación por el criterio de máxima verosimilitud de la probabilidad de cada transición $\delta^k(w_{i-(k-1)}^{i-1}, w_i)$ como (Fu & Huang, 1972):

$$P(w_i / w_{i-(k-1)}^{i-1}) = \frac{N(w_i / w_{i-(k-1)}^{i-1})}{\sum_{\forall w_j \in \Sigma} N(w_j / w_{i-(k-1)}^{i-1})} \tag{3.12}$$

donde $N(w_i / w_{i-(k-1)}^{i-1})$ es el número de veces que la palabra w_i aparece al final del *k*-grama $w_{i-(k-1)}...w_{i-1}w_i$,, o lo que es lo mismo, la cuenta asociada a la transición etiquetada por w_i viniendo del estado etiquetado como $w_{i-(k-1)}^{i-1}$.

Este autómata tal y como está definido, sólo es capaz de representar las probabilidades para los eventos observados en el corpus de entrenamiento. En el siguiente punto veremos las modificaciones necesarias para otorgar una probabilidad distinta de cero a los eventos no vistos.

En la figura 3.4 se muestran algunos ejemplos de autómatas *k*-EE utilizando como corpus de entrenamiento un poema del famoso poeta español Miguel Hernández. En la figura 3.3 se muestra el poema, junto con el conjunto de entrenamiento R^+ y el vocabulario Σ.

llego con tres heridas
la del amor
la de la muerte
la de la vida
con tres heridas viene
la de la vida
la del amor
la de la muerte
con tres heridas yo
la de la vida
la de la muerte
la del amor

Miguel Hernández

$$R^+ = \left\{ \begin{array}{l} \textit{llego con tres heridas, la del amor} \\ \textit{la de la muerte, la de la vida} \\ \textit{con tres heridas viene ,la de la vida} \\ \textit{la del amor, la de la muerte} \\ \textit{con tres heridas yo, la de la vida} \\ \textit{la de la muerte, la del amor.} \end{array} \right\}$$

$$\Sigma = \left\{ \begin{array}{l} \textit{la, de, del, amor, muerte, vida,} \\ \textit{llego, con, tres, heridas, viene, yo} \end{array} \right\}$$

Figura 3.3.- Corpus de entrenamiento seleccionado para realizar la construcción de los autómatas *k*-EE

3.2.- Lenguajes *k*-EE suavizados

Uno de los principales problemas que debe resolverse en la construcción de un ML es la falta de cobertura, es decir, debe estar contemplada la posibilidad de encontrar en la etapa de evaluación o de funcionamiento normal dentro de un sistema de RAH situaciones que no han aparecido en la etapa de aprendizaje. La forma de salvar esta situación consiste en utilizar las técnicas de suavizado en las que se reserva una cierta cantidad de probabilidad para las transiciones inexistentes, de forma que las estimaciones de las probabilidades de los eventos vistos se vean "disminuidas" de alguna forma. También, debe especificarse como repartir esta probabilidad entre las unidades no vistas (ver capítulo 2).

Una vez establecidas las características estructurales y estocásticas de los lenguajes *k*-EE y vista la necesidad del suavizado de las probabilidades de los modelos, se plantea la construcción de un ML de "tipo gramatical" que incluya el suavizado en su propia estructura (Bordel, 1996). El método de suavizado que facilita este planteamiento es el del back-off. Como ya se ha dicho en el capítulo 2 el método de back-off fue introducido por (Katz, 1987) en el modelado de lenguaje y estaba planteado para modelos de N-gramas en base a dos ideas principalmente: la utilización de una estimación alternativa a la de Máxima Verosimilitud que sumara una probabilidad total inferior a la unidad para todos los N-gramas vistos (formula de

Turing) y la distribución de la cantidad de probabilidad restante a los eventos no vistos de forma recursiva mediante el modelo inmediatamente inferior ((N-1)-gramas).

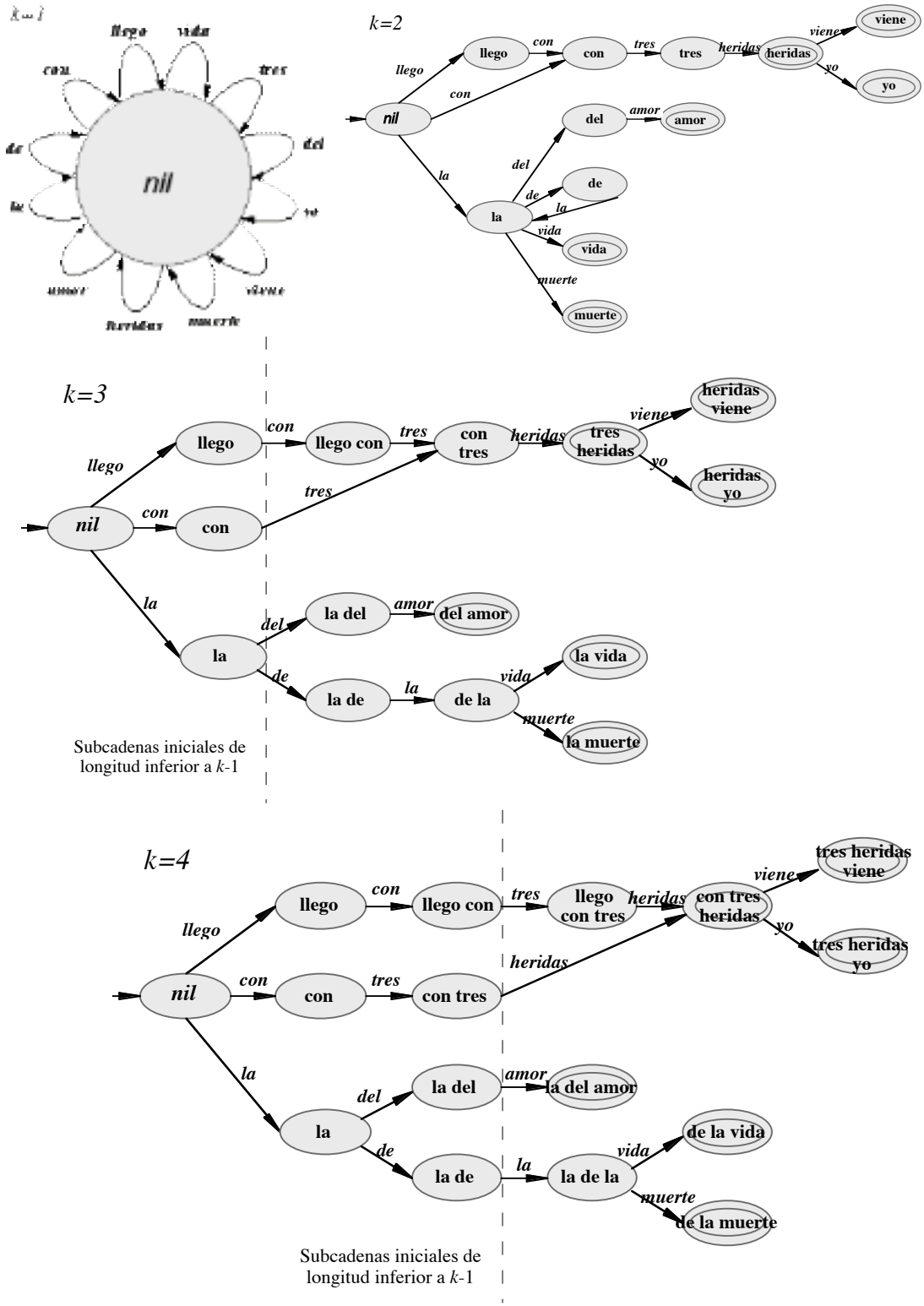


Figura 3.4.- Los autómatas 1-EE, 2-EE, 3-EE y 4-EE obtenidos de R^+ y Σ de la figura 3.3.

En (Bordel et al, 1994) se desarrolló la técnica de back-off bajo el enfoque sintáctico. Como se verá a continuación, este método de suavizado se integra muy fácilmente en la

"mecánica" de construcción recursiva de los autómatas de estados finitos a partir de muestras de entrenamiento.

3.2.1.- Aplicación del back-off sintáctico a los modelos *k*-EE

Utilizando la técnica del suavizado por back-off es posible la creación mediante un esquema de construcción recursivo de un único modelo que integra K modelos *k*-EE con *k*=1...K. En este esquema, las probabilidades que se asignan a los *k*-gramas no vistos se obtienen recursivamente de modelos más generales, *k*-1...1, ponderándolas con las probabilidades sustraídas a los eventos vistos. Dentro de la aproximación que utiliza el suavizado por back-off existen varías propuestas de descuento que se pueden aplicar sobre los eventos vistos (Ney et al, 1997) (Clarkson & Rosenfeld, 1997). En trabajos previos (Bordel et al, 1994) (Bordel, 1996) se desarrolló la técnica del back-off bajo el formalismo utilizando el descuento de Witten-Bell. Bajo este formalismo sintáctico la probabilidad $P(w/q)$ que se asocia a la transición $\delta^k(q, w)= (q_d, P(w/q))$ se estima de acuerdo a la siguiente ecuación:

$$P(w/q) = \begin{cases} \dfrac{N(w/q)}{N(q)+|\Sigma_q|} & w \in \Sigma_q \\ \dfrac{|\Sigma_q|}{N(q)+|\Sigma_q|} \dfrac{P(w/b_q)}{1-\sum\limits_{\forall w' \in \Sigma_q} P(w'/b_q)} & w \in \Sigma - \Sigma_q \end{cases} \qquad (3.13)$$

donde Σq es el vocabulario asociado al estado q y está formado por el conjunto de palabras que aparecen en el corpus de entrenamiento después de la cadena asociada al estado q, es decir, el conjunto de palabras que etiquetan transiciones desde el estado q; $N(w/q)$ es el número de veces que la palabra w aparece después de la cadena que etiqueta al estado q y $N(q) = \sum_{\forall w_i \in \Sigma_q} N(w_i/q)$ es igual al número total de palabras vistas en el estado q.

$P(w/ b_q)$ es la probabilidad estimada asociada el mismo evento en el submodelo (k-1)-EE; por tanto, si el estado q se etiqueta como $w_{i-(k-1)}^{i-1}$ con $q \equiv w_{i-(k-1)}^{i-1} \in Q^k$ entonces el estado b_q se etiqueta como $w_{i-(k-1)+1}^{i-1}$ con $b_q \equiv w_{i-(k-1)+1}^{i-1} \in Q^{k-1}$.

En la ecuación 3.13 las probabilidades que se asocian al conjunto de palabras que aparecen después de la cadena de palabras representada por el estado q en el corpus de entrenamiento -el vocabulario del estado Σq- se estiman explícitamente sustrayendo una pequeña probabilidad a la dada por el criterio de Máxima Verosimilitud para cada una de las palabras $|\Sigma q|$. Las probabilidades de las $|\Sigma| - |\Sigma_q|$ transiciones restantes que corresponden a los eventos que no están representados en el corpus de entrenamiento se estiman de acuerdo a distribuciones de probabilidades más generales en el modelo (*k-1)*-EE.

Utilizando la ecuación 3.13 en vez de la 3.12, se puede obtener la probabilidad asociada a cualquier secuencia de palabras independientemente de haber sido observada en las secuencias de entrenamiento. A partir de cada estado existen tantas transiciones como palabras en el

vocabulario $|\Sigma|$, por lo que mantener toda la información en memoria (el número de probabilidades y transiciones que habría que guardar, coincide con el número de estados por número de palabras distintas de la aplicación ($|Q^k| \times |\Sigma|$)) implica un coste espacial prohibitivo (Riccardi et al, 1996) (Bordel et al, 1997). Sin embargo, no es necesario que las transiciones no vistas durante la etapa de entrenamiento sean explícitamente estimadas ni representadas en cada estado. La estructura del autómata anteriormente definido junto con la técnica de suavizado permiten agruparlas en una única transición a un estado de back-off b_q, que puede ser encontrado en un submodelo más general. Como consecuencia, el número de parámetros que deben ser representados explícitamente se reduce de $|Q^k| \times |\Sigma|$ a $|Q^k| \times (|\Sigma_q|+1)$.

Para todos los estados al aplicar la ecuación 3.13 se debe cumplir la siguiente condición estocástica:

$$\sum_{\forall w_i \in \Sigma} P(w_i / q) = 1 \quad \forall q \in Q^k \tag{3.14}$$

por tanto, la probabilidad que se asigna a cada transición desde cada estado a su estado de back-off $P(b_q /q)$ puede ser eficientemente estimada a partir de las ecuaciones 3.13 y 3.14 de la siguiente manera:

$$P(b_q / q) = \frac{\left|\sum_q\right|}{N(q) + \left|\sum_q\right|} \frac{1}{1 - \sum_{\forall w' \in \Sigma_q} P(w' / b_q)} \tag{3.15}$$

En el modelo suavizado esta transición conecta cada estado q con su estado de back-off b_q, que representa el mismo evento en el modelo (k-1)-EE, por eso si el estado q se etiqueta como $w_{i-(k-1)}^{i-1}$ entonces el estado b_q se etiqueta como $w_{i-(k-1)+1}^{i-1}$. La probabilidad que se asocia a cada suceso no representado en el corpus de entrenamiento $P(w_j/q)$ $\forall w_j \in (\Sigma' - \Sigma_q)$ será estimada de acuerdo a la siguiente ecuación:

$$P(w_j / q) = P(b_q / q) P(w_j / b_q) \quad \forall w_j \in \left(\sum' - \sum_q\right) \tag{3.16}$$

La aproximación sintáctica nos permite obtener un único autómata que integra el conjunto de autómatas k-EE y la técnica de suavizado por back-off sintáctico. Utilizando el suavizado por back-off, la probabilidad que se asigna a los eventos no vistos se obtiene recursivamente a través del conjunto de modelos estocásticos, k-1, k-2,...,1. Este es el modelo que se define en la próxima sección.

3.2.2.- Modelo integrado y suavizado

El modelo integrado y suavizado es un modelo que integra K modelos k-Explorables en Sentido Estricto (k-EE), con k=1...K en un único modelo junto con la técnica de suavizado por back-off sintáctico que los relaciona. El modelo de orden K se define con una quíntupla (Σ', Q^K, q_0, q_f, δ^K) donde:

- $\Sigma' = \Sigma \cup \{\$\}$ siendo $\Sigma = \{w_j\}$, $j = 1...|\Sigma|$, el vocabulario o lo que es lo mismo el conjunto de palabras que aparecen en el corpus de entrenamiento. El símbolo \$ se ha incluido en el corpus de entrenamiento para aislar cada sentencia de las demás. Este símbolo se puede considerar como la primera y la última "palabra" de cada frase.

- Q^K es el conjunto de estados del modelo. Cada estado representa una cadena de palabras $w_{i-k}\, w_{i-(k-1)}...w_{i-1}$, $k = 1...K-1$, con una longitud máxima de K-1, donde i es un índice cualesquiera de la secuencia $w_1...w_i...$ que aparece en el corpus de entrenamiento[3]. Este estado se etiqueta como w_{i-k}^{i-1}. Los estados que representa el comienzo de las frases se etiquetan como $\$w_{i-k}^{i-1}$ donde $k = 1...K-2$ para garantizar la longitud máxima K-1. Un estado especial etiquetado como *nil* representa la secuencia vacía de palabras.

- El autómata tiene un único estado inicial y final $q_0 \equiv q_f \in Q^K$ el cual se etiqueta como \$. Este elemento permite analizar consecutivamente cada una de las sentencias sin tener en cuenta la información contextual entre ellas. El estado inicial es diferente del estado etiquetado con *nil* ya que la probabilidad $P(w_i/nil)$ es la probabilidad estimada $P(w_i)$, mientras $P(w_i/\$)$ es la probabilidad de que w_i sea la primera palabra de la sentencia. La única excepción se encuentra en el modelo 1-EE con un único estado *nil*, que también corresponde al estado inicial y final.

- δ^K es la función de transición $\delta^K : Q^K \times (\Sigma \cup \{\$\}) \rightarrow Q^K \times [0...1]$. $\delta^K(q, w_i) = (q_d, P(w_i/q))$ define un estado destino $q_d \in Q^K$ y una probabilidad $P(w_i/q) \in [0...1]$ que será asignada a cada elemento $(q, w_i) \in Q^K \times (\Sigma \cup \{\$\})$. Cada transición representa a un k-grama, $k = 1...K$, se etiqueta por la última palabra w_i y conecta dos estados etiquetados hasta un máximo de K-1 palabras. Se pueden encontrar varias clases de transiciones en el modelo:

- a) Transiciones desde el estado especial etiquetado como *nil*, que representa la secuencia vacía de palabras, a los $|\Sigma|$ estados etiquetados con cada una de las palabras del vocabulario $w_j \in \Sigma$, $j = 1...|\Sigma|$. La probabilidad asociada a cada transición representa la frecuencia de aparición de la correspondiente palabra en el corpus de entrenamiento:

$$\delta^K(nil, w_j) = (w_j, P(w_j/nil)) = (w_j, P(w_j)) \quad j = 1,...,|\sum| \tag{3.17}$$

- b) Transiciones desde el estado inicial etiquetado como \$ correspondientes al conjunto de palabras que aparecen en la primera posición de cada frase:

$$\delta^K(\$, w_i) = (\$w_i, P(w_i/\$)) \tag{3.18}$$

- c) Transiciones entre secuencias de palabras de longitud inferior a K conectando estados asociados a longitudes de palabras como máximo de K-2:

$$\begin{cases} \delta^K(w_{i-k}^{i-1}, w_i) = (w_{i-k}^{i}, P(w_i/w_{i-k}^{i-1})) & k = 1,...,K-2 \\ \delta^K(\$w_{i-k}^{i-1}, w_i) = (\$w_{i-k}^{i}, P(w_i/\$w_{i-k}^{i-1})) & k = 1,...,K-3 \end{cases} \tag{3.19}$$

Esas transiciones proceden de los K-1 autómatas correspondientes a modelos k-EE con k<K.

---

[3] En este punto, k representa un índice que permite identificar la longitud de las secuencias de palabras que representan los distintos estados. La longitud de las secuencias puede tomar valores desde k=1 hasta k=K-1, puesto que el modelo de orden K integra estados de K modelos k-EE con k=1..K.

- d) Transiciones en el nivel máximo que corresponden a secuencias de palabras de longitud K conectando estados asociados a secuencias de palabras de longitud K-1:

$$\begin{cases} \delta^K(w_{i-(K-1)}^{i-1}, w_i) = (w_{i-(K-1)+1}^{i}, P(w_i / w_{i-(K-1)}^{i-1})) \\ \delta^K(\$w_{i-(K-2)}^{i-1}, w_i) = (w_{i-(K-2)}^{i}, P(w_i / \$w_{i-(K-2)}^{i-1})) \end{cases} \tag{3.20}$$

- e) Transiciones al estado final q_f ($\equiv q_0$) etiquetado por \$:

$$\begin{cases} \delta^K(w_{i-k}^{i-1}, \$) = (q_f, P(\$/ w_{i-k}^{i-1})) & k = 1,..., K-1 \\ \delta^K(\$w_{i-k}^{i-1}, \$) = (q_f, P(\$/ \$w_{i-k}^{i-1})) & k = 1,..., K-2 \end{cases} \tag{3.21}$$

- f) transiciones al estado de back-off. Tal y como se ha visto, la probabilidad $P(w_i/q)$ asociada a cada transición $\delta^K(q, w_i)$ se estima según la ecuación 3.13. Por tanto, para cada estado del autómata $q \in Q^K$, excepto para el estado etiquetado como *nil*, se debe añadir una nueva transición a su estado de back-off b_q.

$$\delta^K(q, U) = (b_q, P(b_q / q)) \tag{3.22}$$

donde: el símbolo U representa a todos los eventos no vistos en entrenamiento $w_j \in (\Sigma' - \Sigma_q)$ asociados al estado q. El estado de back-off b_q asociado a cada estado q se puede encontrar en el submodelo (K-1).

Por eso, cuando los estados están asociados a cadenas de longitud hasta un máximo de K-2, se cumple que:

$$\begin{aligned} q \equiv w_{i-k}^{i-1} \quad &\Rightarrow \quad b_q \equiv w_{i-k+1}^{i-1} \quad k = 1,..., K-2 \\ q \equiv \$w_{i-k}^{i-1} \quad &\Rightarrow \quad b_q \equiv \$w_{i-k+1}^{i-1} \quad k = 1,..., K-3 \end{aligned} \tag{3.23}$$

y para los estados en el nivel superior asociados a secuencias de palabras de longitud igual a K-1:

$$\begin{aligned} q \equiv w_{i-(K-1)}^{i-1} \quad &\Rightarrow \quad b_q \equiv w_{i-(K-1)+1}^{i-1} \\ q \equiv \$w_{i-(K-2)}^{i-1} \quad &\Rightarrow \quad b_q \equiv \$w_{i-(K-2)+1}^{i-1} \end{aligned} \tag{3.24}$$

Finalmente, el estado de back-off asociado al estado etiquetado por \$ es el estado especial etiquetado como *nil*. El estado *nil* no tiene estado de back-off asociado puesto que representa la cadena vacía de palabras.

La figura 3.5 muestra esta estructura para un estado q etiquetado como $w_{i-(K-1)}^{i-1}$. Las transiciones etiquetadas para las $|\Sigma_q|$ palabras que se observan en la etapa de entrenamiento tras el K-grama representado en el estado q se conectan a otros estados en el mismo nivel K. Sin embargo, las transiciones etiquetadas por U conectan el estado q a su estado de back-off, $w_{i-(K-1)+1}^{i-1}$en el submodelo (K-1).

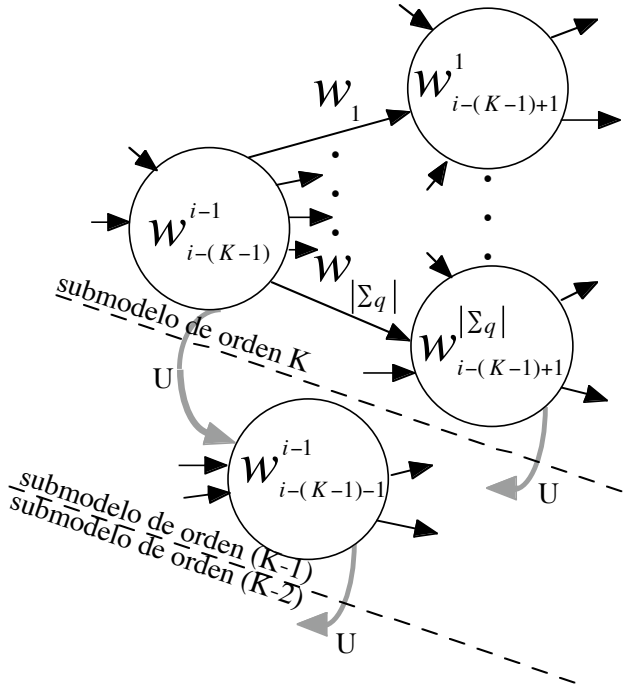


Figura 3.5.- Las transiciones entre eventos vistos en un estado q ($w_j \in \Sigma_q$) conectan el estado con estados en el mismo submodelo K-EE. Las transiciones para los elementos no vistos (etiquetadas por U) conectan con su estado de back-off en el modelo (K-1)-EE.

En la figura 3.6 se muestra la red finita representando la estructura del modelo integrado y suavizado para K=4 obtenido para el conjunto de entrenamiento de la figura 3.3 (se ha añadido el símbolo $ como separador de sentencias). Este modelo supone la integración en una única estructura de los 4 modelos presentados en la figura 3.4. En la figura 3.6 cada nodo de la red representa un estado del autómata. Las transiciones a partir del estado especial etiquetado como *nil* (que corresponde la secuencia vacía de palabras) están representadas como enlaces entre el nodo raíz y sus nodos hijos. Los enlaces a partir del nodo etiquetado por $ representan las transiciones a partir del estado inicial. Las transiciones que corresponden a secuencias de palabras de longitud inferior a K están representadas por $|\Sigma_q|$ enlaces conectando cada nodo del nivel k=1...K-2 con sus $|\Sigma_q|$ nodos hijos. Los nodos del nivel K-1 representan los estados asociados a secuencias de palabras de longitud igual a K-1 etiquetadas como $w_{i-(K-1)}^{i-1}$. Por claridad, no se han representado las transiciones al estado final que existen para todos los nodos cada vez que el símbolo $ aparece en las muestras de entrenamiento y en su lugar estos nodos aparecen marcados.

La ventaja principal de esta formulación sintáctica es que lleva a una representación muy eficiente de los parámetros del modelos durante la etapa de entrenamiento, es decir, distribución de la probabilidad y estructura del modelo. Bajo esta representación, el estado de back-off asociado a cada estado puede ser fácilmente definido como un enlace en una tabla y calcular directamente su probabilidad asociada. Finalmente la función δ^K se podrá representar fácilmente como una función de búsqueda recursiva como se verá en la próxima sección.

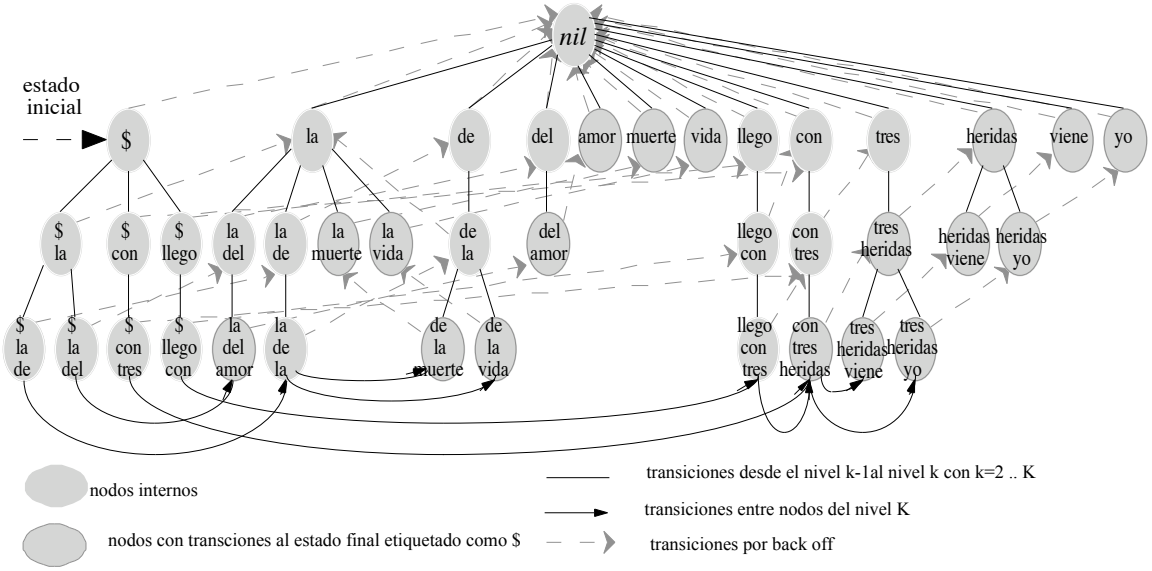


Figura 3.6.- La red completa representando la estructura del modelo integrado y suavizado (K=4) obtenido a partir del corpus de entrenamiento de la figura 3.3 cuando las marcas de separación (comas y punto) se sustituyen por el símbolo especial $.

3.2.3.- Modelos de lenguaje *k*-EE y N-gramas

Los modelos de lenguaje estadísticos se puedan representar mediante redes de estados finitos estocásticas bajo la aproximación sintáctica (Bonafonte & Mariño, 1998) (Suzuki & Aso, 1999), pero ésto no significa que ambas aproximaciones coincidan. Tanto en (Segarra, 1993) como en (Riccardi et al, 1996) se pone de manifiesto que la distribución de probabilidades de los N-gramas y de los mecanismos gramaticales utilizando gramáticas *k*-Explorables en sentido Estricto, en la práctica es la misma, suponiendo que N y *k* coinciden en valor si en ambos casos se utiliza el criterio de máxima verosimilitud para obtener las probabilidades asociadas a las transiciones. Sin embargo, las distribuciones de probabilidad de los N-gramas se obtienen sobre longitudes de frases fijas mientras que con los lenguajes *k*-EE, pueden ser definidas sobre Σ^*, por lo que está aún por establecer una similitud formal.

Los N-gramas de longitud variable (ver capítulo 2) permiten simular mejor el comportamiento de las gramáticas *k*-EE utilizando también símbolos especiales y una única transición para todos los eventos no vistos (Young et al, 1997). De hecho, la distribución de probabilidad que resulta al representar los N-gramas de longitud variable (obtenidos con toolkit de la CMU (Clarkson & Rosenfeld, 1997)) y los autómatas de estados finitos deterministas que reconocen los lenguajes K-EE sobre cadenas de longitud variable, es exactamente la misma. La diferencia recae en que los modelos obtenidos bajo la aproximación gramatical llevan en la expresión del ML, la propia estructura del autómata y no sólo la distribución de probabilidad. Además la aproximación gramatical está respaldada por una teoría ampliamente estudiada y bien consolidada tras muchos años de estudio (Fu & Juang, 1972) (Hopcroft & Ullman, 1979).

3.3.- Representación eficiente del autómata de estados finitos determinista K-EE suavizado

La formulación presentada permite una representación muy eficiente de los parámetros del modelo: distribución de la probabilidad y estructura del modelo. En la figura 3.6 se ha presentado la estructura del modelo que integra K modelos *k*-EE suavizado (*k*=1...K) mediante una red finita. Esta red se construye a partir de una estructura inicial en forma de árbol construido a partir de un conjunto de muestras de entrenamiento. El procedimiento para obtener dicha representación se muestra en el apéndice B. La naturaleza de la estructura de tipo árbol permite almacenar tanto la estructura del modelo suavizado como la distribución de la probabilidad en un simple array (ver sección 3.3.1). Para completar la representación del autómata es necesaria la implementación de la función de transición δ^K previamente definida mediante un esquema recursivo (búsqueda a través de una estructura array) que permita obtener una respuesta para cada estado $q \in Q^K$ y para cada $w \in \Sigma'$. Este objetivo se consigue a través de un función de búsqueda simple a través del array (ver sección 3.3.2).

3.3.1.- Representación de los modelos K-EE en una estructura tipo array

La red finita que se obtiene (figura 3.6) se puede representar de una forma muy eficiente y compacta, usado un array que contenga todos los parámetros del modelo. Cada estado del autómata se representa por $|\Sigma_q|+1$ posiciones, cada una de las cuales representa una transición de salida, a excepción del estado *nil* que ocupa $|\Sigma|$ posiciones puesto que no hay ningún estado de back-off asociado al estado *nil*. Cada posición del array representa un par (q,w) donde $q \in Q^K$ y $w \in \Sigma_q \cup \{U\}$ y se compone de cuatro elementos:

- un *short integer* que representa una transición por cada palabra $w \in \Sigma_q$ o por el símbolo *U* para cualquiera de los eventos no vistos $w_i \notin \Sigma_q$

- Un *double* que representa la probabilidad $P(w_i/q)\ \forall w_i \in \Sigma_q$ o $P(b_q/q)\ \forall w_i \notin \Sigma_q$

- Un *short integer* que representa el valor de $|\Sigma_q|$, es decir, el número de eventos distintos vistos en cada estado.

- Un *short integer* que representa el enlace explícito al primer nodo hijo de *q* o a su estado de back-off b_q.

La figura 3.7 muestra la tabla (array de cuatro elementos) que representa al modelo integrado y suavizado con K=4 de la figura 3.6. En la figura 3.7 se ha incluido información adicional para clarificar el significado de cada componente del array: el índice que ocupa en la tabla cada posición del array, el numeral correspondiente a cada estado #q del autómata y la etiqueta asociada para cada palabra $w_i \in \Sigma$. Sin embargo, sólo los cuatro elementos previamente citados deben ser realmente incluidos: *#w*, $P(w/q)$, $|\Sigma_q|$ y el índice de la posición del estado destino.

Por motivos de claridad, cada transición se representa por la etiqueta de la última palabra de la secuencia de palabras que se esté considerando, (etiqueta(*w*)), en vez de utilizar su índice numérico (*#w*). En esta columna el símbolo U representa cualquier evento no observado en el corpus de entrenamiento asociado el estado q que se etiqueta con la palabra $w_j \in (\Sigma - \Sigma_q)$. El destino de esta transición es el estado de back-off correspondiente b_q y la probabilidad asociada $P(b_q/q)$ se calcula a partir de la ecuación 3.15.

El array completo consta de $|Q^K| + \times (|\Sigma_q| + 1)$ posiciones. Cada modelo *k*-EE, *k*=1...K, es decir, cada nivel del bi-trie original, se representa por un conjunto de estados igual al número de *k*-gramas que aparecen en el corpus de entrenamiento. Por eso, la estructura que muestra la figura 3.6 está formada por un estado representado al modelo 1-EE, trece representado a los modelos 2-EE, etc. Cada estado se representa con $|\Sigma_q|+1$ posiciones del array, cada una de ellas representado una transición de salida.

El procedimiento para crear esta estructura a partir de muestras de entrenamiento consta de tres pasos que aparecen detallados en el apéndice B.

indice	#q	etiqueta(q)	$\|\Sigma_q\|$	etiqueta(w)	P(w/q)	destino	#*k*-EE
1	1	*[nil]*	**12**	**la**	**0.3333**	**17**	
2				**de**	**0.1333**	**22**	
3				**con**	**0.0667**	**24**	
4				**tres**	**0.0667**	**26**	
5				**heridas**	**0.0667**	**28**	
6				**del**	**0.0667**	**32**	
7				**amor**	**0.0667**	**34**	
8				**muerte**	**0.0667**	**36**	
9				**vida**	**0.0667**	**38**	
10				**viene**	**0.0222**	**40**	
11				**yo**	**0.0222**	**42**	
12				**llego**	**0.0222**	**44**	K=1
13	2	[$]	**3**	**la**	**0.6000**	**46**	
14				**con**	**0.1333**	**49**	
15				**llego**	**0.0667**	**51**	
16				***U***	**0.3461**	**1**	
17	3	[la]	**4**	**de**	**0.3158**	**55**	
18				**del**	**0.1579**	**53**	
19				**muerte**	**0.1579**	**57**	
20				**vida**	**0.1579**	**59**	
21				***U***	**0.3158**	**1**	
22	4	[de]	**1**	**la**	**0.8571**	**61**	
23				***U***	**0.2143**	**1**	
24	5	[con]	**1**	**tres**	**0.7500**	**64**	
25				***U***	**0.2679**	**1**	
26	6	[tres]	**1**	**heridas**	**0.7500**	**66**	
27				***U***	**0.2679**	**1**	
28	7	[heridas]	**3**	**$**	**0.1667**	**13**	
29				**viene**	**0.1667**	**70**	
30				**yo**	**0.1667**	**72**	
31				***U***	**0.5232**	**1**	
32	8	[del]	**1**	**amor**	**0.7500**	**74**	
33				***U***	**0.2679**	**1**	
34	9	[amor]	**1**	**$**	**0.7500**	**13**	
35				***U***	**0.2500**	**1**	
36	10	[muerte]	**1**	**$**	**0.7500**	**13**	
37				***U***	**0.2500**	**1**	
38	11	[vida]	**1**	**$**	**0.7500**	**13**	
39				***U***	**0.2500**	**1**	
40	12	[viene]	**1**	**$**	**0.7500**	**13**	
41				***U***	**0.2500**	**1**	
42	13	[yo]	**1**	**$**	**0.7500**	**13**	
43				***U***	**0.2500**	**1**	
44	14	[llego]	**1**	**con**	**0.5000**	**76**	
45				***U***	**0.5357**	**1**	K=2
46	15	[$la]	**2**	**de**	**0.5455**	**78**	
47				**del**	**0.2727**	**80**	
48				***U***	**0.3455**	**17**	
49	16	[$con]	**1**	**tres**	**0.6667**	**82**	
50				***U***	**1.3333**	**24**	
51	17	[$llego]	**1**	**con**	**0.5000**	**84**	
52				***U***	**1.0000**	**44**	
53	18	[la del]	**1**	**amor**	**0.7500**	**86**	
54				***U***	**1.0000**	**32**	
55	19	[la de]	**1**	**la**	**0.8571**	**88**	
56				***U***	**1.0000**	**22**	
57	20	[la muerte]	**1**	**$**	**0.7500**	**13**	
58				***U***	**1.0000**	**36**	
59	21	[la vida]	**1**	**$**	**0.7500**	**13**	
60				***U***	**1.0000**	**38**	
61	22	[de la]	**2**	**muerte**	**0.3750**	**91**	
62				**vida**	**0.3750**	**93**	
63				***U***	**0.3654**	**17**	
64	23	[con tres]	**1**	**heridas**	**0.7500**	**95**	
65				***U***	**1.0000**	**26**	
66	24	[tres heridas]	**3**	**$**	**0.1667**	**13**	
67				**viene**	**0.1667**	**99**	
68				**yo**	**0.1667**	**101**	
69				***U***	**1.0000**	**28**	
70	25	[heridas viene]	**1**	**$**	**0.5000**	**13**	
71				***U***	**1.0000**	**40**	
72	26	[heridas yo]	**1**	**$**	**0.5000**	**13**	
73				***U***	**1.0000**	**42**	
74	27	[del amor]	**1**	**$**	**0.7500**	**13**	
75				***U***	**1.0000**	**34**	
76	28	[llego con]	**1**	**tres**	**0.5000**	**103**	
77				***U***	**2.0000**	**24**	K=3
78	29	[$ la de]	**1**	**la**	**0.8571**	**88**	
79				***U***	**1.0000**	**55**	
80	30	[$ la del]	**1**	**amor**	**0.7500**	**86**	
81				***U***	**1.0000**	**53**	
82	31	[$ con tres]	**1**	**heridas**	**0.6667**	**95**	
83				***U***	**1.3333**	**64**	
84	32	[$ llego con]	**1**	**tres**	**0.5000**	**103**	
85				***U***	**1.0000**	**76**	
86	33	[la del amor]	**1**	**$**	**0.7500**	**13**	
87				***U***	**1.0000**	**74**	
88	34	[la de la]	**2**	**muerte**	**0.3750**	**91**	
89				**vida**	**0.3750**	**93**	
90				***U***	**1.0000**	**61**	
91	35	[de la muerte]	**1**	**$**	**0.7500**	**13**	
92				***U***	**1.0000**	**57**	
93	36	[de la vida]	**1**	**$**	**0.7500**	**13**	
94				***U***	**1.0000**	**59**	
95	37	[con tres heridas]	**3**	**$**	**0.1667**	**13**	
96				**viene**	**0.1667**	**99**	
97				**yo**	**0.1667**	**101**	
98				***U***	**1.0000**	**66**	
99	38	[tres heridas viene]	**1**	**$**	**0.5000**	**13**	
100				***U***	**1.0000**	**70**	
101	39	[tres heridas yo]	**1**	**$**	**0.5000**	**13**	
102				***U***	**1.0000**	**72**	
103	40	[llego con tres]	**1**	**heridas**	**0.5000**	**13**	
104				***U***	**2.0000**	**64**	K=4

Figura 3.7.- Se representan los parámetros del modelo integrado y suavizado de la figura 3.6 (una transición, una probabilidad, número de eventos distintos vistos en cada estado y un enlace al primer hijo o al estado *bq*).

3.3.2.- La función de transición δ a través de la representación interna del modelo K-EE

Para completar la representación del modelo K-EE integrado y suavizado, se debe representar la función de transición δ mediante una estructura de búsqueda a través de la estructura presentada (ver figura 3.8). La función de transición δ define un destino q_d y una probabilidad $P(w/q)$ asociada a cada par (q,w) $\forall q \in Q^K$ y $\forall w \in \Sigma'$:

$$\delta(q, w) = (q_d, P(w / q)) \qquad \forall q \in Q^K \wedge \forall w \in \sum' \qquad q_d \in Q^K \tag{3.24}$$

```
Funcion δ (q∈Q^K; w∈Σ): (d∈Q^K; P∈[0...1]);
var q_aux∈Q^K; P_aux∈[0...1];
principio
si w∈Σ_q entonces δ_d←array_dest[q,w]        (* eventos vistos *)
                  δ_P←array_prob[q,w]
        sino      P_aux←array_prob[q,U] (* eventos no vistos *)
                  q_aux←array_dest[q,U]
                  mientras w∉Σ_q_aux hacer
                     P_aux← P_aux × array_prob[q_aux,U]
                     q_aux←array_dest[q_aux,U]
                  fin_mientras
                  δ_P ← P_aux *array_prob[q_aux,w]
                  δ_d ←array_dest[q_aux,w]
fin_si
fin_δ.
```

Figura 3.8.- Función de búsqueda a través de la estructura presentada en la figura 3.7.

Cuando aparecen elementos vistos en entrenamiento, el estado de destino q_d se puede encontrar directamente como el índice destino de la posición del array (w,q) (ver figura 3.7). De la misma manera el valor de $P(w/q)$, obtenido a partir de la ecuación 3.13 para $w \in \Sigma_q$, se obtiene directamente como el valor de la probabilidad en la posición (w,q) (ver figura 3.7). Sin embargo, cuando aparece un evento que no ha sido observado en entrenamiento, ambos valores de q_d y $P(w/q)$ no se pueden encontrar directamente en la estructura array y es necesario manejar la información contenida en la misma mediante una función de búsqueda recursiva a través del mismo. Esta función (ver figura 3.8), busca a través de los estados de back-off, es decir, a partir de transiciones por el símbolo *U*, hasta que encuentra la palabra *w* como un evento visto en un estado q de nivel inferior (k<K), es decir, $w \in \Sigma_q$. El estado *q* será el estado destino buscado q_d. El valor $P(w/q)$ se debe calcular de acuerdo a la ecuación 3.16. Este procedimiento se describe a través de la función δ de la figura 3.8 donde: δ_d es el estado destino q_d y δ_P la probabilidad $P(w/q)$.

La función δ presentada en la figura 3.8 sólo es útil cuando se trabaja sobre un texto escrito ya que para cada estado q la función δ únicamente debe dar respuesta (estado destino y probabilidad) para una palabra concreta *w*. Sin embargo, si el modelo *k*-EE va a ser integrado junto con los modelos acústicos en un sistema de RAH, a partir de un estado para cada palabra

de la aplicación debe existir un estado destino y una probabilidad. Si se utilizara la función anterior habría que hacer en cada estado $|\Sigma|$ llamadas a la función y ésto implicaría un coste bastante alto. Por ello se ha desarrollado la función de la figura 3.9 cuya única diferencia es que la salida son dos arrays (uno de estados destinos y otro de probabilidades) para permitir explorar decodificaciones alternativas. Esta función recibe el nombre de δ_Viterbi ya que es la que se va a utilizar en el algoritmo de Viterbi (ver capítulo 4).

```
Función δ_Viterbi(q∈Q^K; w∈Σ): (d[w_1, ..,w_|Σ|]/d[w_i]∈Q^K; P[w_1, ..,w_|Σ|]/
P[w_i]∈[0...1]);
var q_aux∈Q^K; P_aux∈[0...1];
principio
para w=1 hasta |Σ| hacer
        δ_d[w] ←NUL
fin_para
para w=1 hasta |Σ_q| hacer
        δ_d[w] ← array_dest[q,w]
        δ_P[w] ← array_prob[q,w]
fin_para
P_aux ← array_prob[q,U]
q_aux ←array_dest[q,U]
repetir
   para w=1 hasta |Σ_q_aux| hacer
        si δ_d[w]=NUL
            δ_P[w] ← P_aux*array_prob[q_aux,w]
            δ_d[w] ← array_dest[q_aux,w]
        fin_si
   fin_para
   if q_aux≠nil entonces
        P_aux← P_aux × array_prob[q_aux,U]
        q_aux←array_dest[q_aux,U]
       sino    q_aux←NUL
   fin_si
hasta (q_aux=NUL)
fin δ_Viterbi
```

Figura 3.9.- Se representa la función de búsqueda extendida (δ_Viterbi) para dar como salida las probabilidades y estados destino para todas las palabras $|\Sigma|$ como posibles alternativas en decodificación, utilizando el algoritmo de Viterbi.

3.4.- Evaluación de los modelos *k*-EE

Una vez construido el ML que integra K modelos *k*-EE (*k*=1...K) y la técnica de suavizado en su estructura, es necesaria la evaluación de estos modelos de lenguaje para los distintos valores de K. Por supuesto, la mejor evaluación resulta tras observar su efectividad al ser integrados dentro del sistema de reconocimiento. Esta evaluación se realizará en profundidad en el capítulo 4. En este punto vamos a realizar una evaluación sin tener en cuenta el sistema de

RAH, primero midiendo la memoria necesaria para representarlos y usando después como medida de evaluación la perplejidad (ver capítulo 2).

3.4.1.- Memoria

El uso de una estructura de array supone un ahorro de memoria respecto a tener representados explícitamente todas las transiciones con sus probabilidades y estado destino asociados. La reducción en el número de parámetros es de $|Q^K|\times|\Sigma|$ a $|Q^K|\times(|\Sigma_q|+1)$. A continuación se presentan los resultados comparativos de ahorro de memoria con el desarrollo de la estructura array propuesta. En la tabla 3.1 se presentan los resultados obtenidos para la tarea MLA (29 palabras) y en la tabla 3.2 para la Tarea BDGEO (1208 palabras) (ver apéndice A). Con la estructura propuesta el número de parámetros asociado a cada estado es mayor (4 en vez de 2) pero hay muchos menos estados como se puede apreciar.

Tabla 3.1.- Número de estados y memoria necesaria para representar el autómata de la aplicación MLA para diferentes valores de K.

		Estructura Completa		Array propuesto	
K	$\|Q^K\|$	$\|Q^K\|\times\|\Sigma\|$	Memoria	$\|Q^K\|\times(\|\Sigma_q\|+1)$	Memoria
2	31	899	11.5 Kb	214	3.0 Kb
3	173	8590	64.3 Kb	850	11.9 Kb
4	643	18647	239.2 Kb	2542	35.6 Kb
5	1808	52432	672.5 Kb	6642	93.0 Kb
6	4518	131022	1680 Kb	15614	218.6 Kb

Con la representación de la estructura mediante un array se observa una fuerte reducción en la memoria necesaria para representar toda la información frente al desarrollo de la estructura completa que no se utiliza ya que es inviable para tareas incluso de vocabulario medio y valores de *k* no excesivamente altos. En la práctica, aunque se utilicen modelos estadísticos, se recurre también a transiciones comunes para todos los eventos no vistos (Young et al, 1997), aunque no sea algo implícito a su definición y normalmente no se explica claramente el procedimiento. El no proceder de esta manera llevaría a la propuesta de sistema inviables.

Tabla 3.2.- Número de estados y memoria necesaria para representar el autómata de la aplicación BDGEO para diferentes valores de K.

		Estructura Completa		Array propuesto	
K	$\|Q^K\|$	$\|Q^K\|\times\|\Sigma\|$	Memoria	$\|Q^K\|\times(\|\Sigma_q\|+1)$	Memoria
2	1,213	$14.60\ 10^5$	17.6 Mb	9285	0.13 Mb
3	7,479	$90.34\ 10^5$	108.8 Mb	30714	0.43 Mb
4	21,551	$26.03\ 10^6$	313.6 Mb	67857	0.95 Mb
5	42,849	$51.76\ 10^6$	623.7 Mb	120714	1.69 Mb
6	69,616	$84.09\ 10^6$	1013.3 Mb	182142	2.55 Mb

3.4.2.- Perplejidad

Vamos a ver los resultados que se obtienen de la evaluación de los modelos de lenguaje de forma independiente a los modelos acústicos, es decir, usando la perplejidad. La figura 3.10 muestra los resultados que se obtienen considerando los corpus de evaluación de la aplicación (Prueba-Texto que constan de 500 frases y 8397 palabras para la tarea MLA y de 1147 frases y 11823 palabras para la tarea BDGEO (ver apéndice A)) y los corpus que realmente se van a utilizar posteriormente para reconocimiento (Prueba_Voz que constan de 1600 frases y 25664 palabras para la tarea MLA y de 600 frases y 5655 palabras para la tarea BDGEO (ver de nuevo apéndice A)).

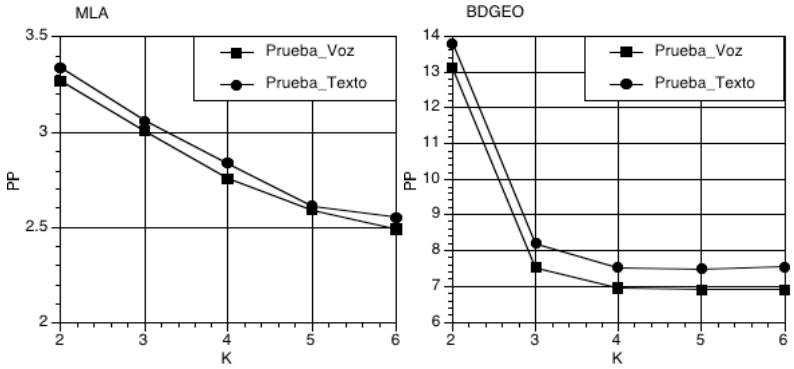


Figura 3.10.- Representa la perplejidad asociada a los corpus de evaluación del ML (Prueba_Texto) y de voz (Prueba_Voz) que se van a utilizar en el sistema de RAH para las tareas MLA y BDGEO respectivamente.

Por supuesto no son comparables los valores de la perplejidad obtenidos en números absolutos para ambas tareas como ya se ha explicado en el capítulo 2. En la tarea MLA se observa que la perplejidad sigue disminuyendo a medida que aumenta el valor de K, sin embargo para la tarea BDGEO se estabiliza a partir de K=4. Este resultado es lógico (ver apéndice A) ya que la tarea MLA es una tarea de baja complejidad (29 palabras) y las estructuras sintácticas de las frases se repiten mucho sobre todo para valores altos de K. Esto hace que los modelos K-EE con valores altos de K presenten probabilidades altas para esas combinaciones de palabras y que como consecuencia la perplejidad sea baja. Sin embargo, para la tarea BDGEO, las estructuras se repiten menos y sobre todo para valores altos de K habrá cada vez más situaciones menos probables, manteniéndose los valores de perplejidad.

Capítulo 4

Sistema de RAH: Prototipo TORNASOL

En este capítulo se presenta el prototipo completo de RAH que funciona en la Universidad del País Vasco/ Euskal Herriko Unibertsitatea (Rodríguez et al, 1999). El prototipo TORNASOL[4] se ha desarrollado dentro del proyecto SENGLAR (Casacuberta et al, 1995-98). El objetivo final era la obtención de un sistema que funcionara en tiempo real para una talla media de vocabulario sobre una aplicación concreta: la consulta a una base de datos geográfica del castellano (BDGEO) (ver apéndice A).

Se van a resumir cada una de las etapas que permiten el funcionamiento del sistema, pero lógicamente, la mayor parte de los resultados presentados en esta memoria hacen referencia al módulo que integra los modelos acústicos y de lenguaje. Se presentarán resultados del sistema funcionando *off-line* (evaluación del módulo de búsqueda del prototipo) para la base de datos BDGEO (1208 palabras), desarrollada como corpus de la tarea dentro del proyecto Albaycin. Para ello se cuenta con un subconjunto de frases grabadas (ver apéndice A). También y a efectos de comparación se presentarán los resultados para una base de datos de vocabulario menor llamada MLA (29 palabras) (ver de nuevo el apéndice A).

Según la regla de Bayes, el módulo de búsqueda debe integrar los conocimientos acústicos $P(A/\Omega)$ con los conocimientos léxicos, semánticos, etc $P(\Omega)$. Ambos tipos de conocimiento se aplican mediante modelos que deben ser estimados previamente a su integración en un sistema de RAH. El proceso de estimación de las probabilidades de los modelos se suele abordar en dos partes:

- Las del modelo de lenguaje: ya se ha hablado de ellas extensamente en los capítulos 2 y 3 introduciendo la aproximación gramatical. Este modelo, como ya se ha visto, es independiente de los datos acústicos y se obtiene a partir de bases de datos extensas de texto escrito.

- La estimación de las probabilidades acústicas, lógicamente depende de la información proporcionada por la secuencia acústica. En la siguiente sección vamos a ver en más detalle el modelado acústico. En este tipo de modelado no se va a introducir ninguna idea novedosa, sino que se va a hacer un resumen de las técnicas clásicas más utilizadas por la comunidad científica

[4] Este nombre fue elegido en honor al famoso científico sordo Silvestre Tornasol (professor Calculus en la versión inglesa) que aparece en la serie de comics TINTIN.

internacional y se va a presentar a continuación, la selección realizada para el prototipo desarrollado.

La organización de este capítulo es como sigue: en primer lugar se presentan los resultados de Decodificación Acústico-Fonética (DAF) obtenidos, lo que nos acerca a la problemática asociada a los modelos acústicos. Se presentan diferentes técnicas de modelado junto con diferentes conjuntos de unidades a modelar y se obtienen resultados experimentales de alguna de las propuestas. Se trata en definitiva de una puesta a punto de los modelos acústicos.

En segundo lugar se presentan los resultados de reconocimiento de palabras obtenidos con la incorporación del conocimiento léxico al sistema de RAH. En este punto, se presentan los resultados de reconocimiento obtenidos al desarrollar un sistema de RAH de palabras aisladas y que por tanto no hace uso de ningún conocimiento sintáctico. El objetivo es el de establecer la cota máxima de error que el sistema integrado debe reducir. En tercer lugar, se presenta el sistema de RAH integrado que combina los conocimientos acústicos, léxicos y sintácticos. Los modelos acústicos utilizados serán los elegidos tras el estudio previo y como modelo de lenguaje se utiliza una subclase de las gramáticas regulares, en concreto las gramáticas *k*-Explorables en Sentido Estricto (*k*-EE) presentadas en el capítulo anterior.

En cuarto lugar, se va a estudiar y evaluar la necesidad de la aplicación de heurísticos que modifiquen la contribución acústica y/o la del ML para lograr la adecuada integración de los modelos acústicos y de lenguaje. En concreto se evalúan los resultados de decodificación obtenidos tras aplicar distintas funciones (lineales, exponenciales, escalones, etc.) sobre una o ambas distribuciones de probabilidad.

Por último, se presenta la arquitectura del prototipo, desarrollada mediante una estructura cliente-servidor.

4.1.- Decodificación Acústico-Fonética

Como se ha mencionado anteriormente el módulo de búsqueda trata de obtener la secuencia de unidades léxicas (clásicamente palabras) pronunciadas por cada locutor. Para poder realizar una interpretación de la señal vocal es preciso un formalismo que relacione la secuencia de carácter acústico con su contenido fonético y léxico.

Los modelos de las unidades léxicas o palabras (w) se obtienen de acuerdo a una serie de reglas fonológicas y se forman por concatenación de las unidades subléxicas que las componen. Cada unidad subléxica se representa mediante un modelo dando lugar a un conjunto de modelos:

$$M = \{m^1, m^2, .. m^{|U|}\}$$

En la etapa de entrenamiento acústico se debe obtener este conjunto de modelos. Hay dos aspectos fundamentales que se deben tener en cuenta en el diseño del modelado acústico: unidades subléxicas a modelar y el tipo de modelado de estas unidades.

La etapa de decodificación acústico fonética consiste en obtener la secuencia de unidades subléxicas que han sido reconocidas. Esta evaluación nos dará una idea de la bondad de los modelos acústicos.

4.1.1.- Técnicas de modelado

El modelado acústico-fonético es un tema muy estudiado desde los inicios del RAH hasta la actualidad. Hay un gran número de publicaciones al respecto pero todas ellas se pueden clasificar fundamentalmente en cuatro conjuntos: Modelos Ocultos de Markov (método muy extendido y consolidado), Inferencia Gramatical, Métodos Estructurales-Conexionistas y modelado por patrones. En este trabajo sólo haremos una pequeña reseña de cada uno de los métodos. En (Rabiner & Juang, 1993), literatura básica de reconocimiento, se puede encontrar una descripción más detallada.

Modelos Ocultos de Markov (MOM): los MOM tratan el problema del modelado desde el punto de vista estadístico. Los MOM pretenden modelar las distribuciones de probabilidad de los elementos de cada clase (por ejemplo, fonemas). Son modelos paramétricos con limitaciones en su capacidad de representación, pero existen métodos consolidados para la estimación de sus parámetros y posterior utilización en clasificación. Además, son muy eficientes desde el punto de vista computacional, tanto en la memoria requerida como en el tiempo de decodificación necesario. El uso de MOM se puede considerar como un método clásico, bien establecido y parte fundamental en muchos sistemas de RAH. La teoría ha sido ampliamente difundida por varios autores (Lee, 1989) (Rabiner, 1989). Según este último autor la teoría fue desarrollada por Baum a finales de los 60 e introducida al RAH tanto por Baker (Baker, 1975) como por Jelinek (Jelinek, 1976) a principios de los 70.

Inferencia gramatical: los modelos basados en Inferencia Gramatical (IG) provienen del reconocimiento sintáctico de formas y conciben las secuencias acústicas del evento a modelar como frases de un lenguaje. Los modelos basados en IG no necesitan un conocimiento a priori de la topología del modelo (número de estados y conexiones entre ellos) sino que se infiere al mismo tiempo que el lenguaje. Tienen una información estructural muy rica (normalmente mayor que la de los MOM), por lo que el espacio de búsqueda puede ser muy grande. Por otra parte, si se utilizan versiones con probabilidades de observación continuas, las distribuciones de los eventos acústicos asociados a cada transición son relativamente sencillas, al ser muy especificas, y por tanto requieren funciones densidad de probabilidad menos complejas que los MOM. Este método ha sido ampliamente desarrollado dentro del Grupo de Reconocimiento de Habla de la Universidad Politécnica de Valencia (Segarra, 1993) (Sanchis, 1994).

Aproximación estructural-conexionista: la aproximación estructural conexionista consiste en aumentar la capacidad discriminativa de los modelos estructurales, ya sean los MOM o los obtenidos por IG. Para ello, se combinan con una red neuronal que se utiliza para estimar las probabilidades de emisión de un conjunto de modelos estructurales. En (Castro, 1999) y dentro del proyecto conjunto SENGLAR (Casacuberta et al, 1995-98) se han desarrollado

sistemas capaces de integrar un tipo de modelos conexionista, el perceptrom multicapa con distintos modelos estructurales, obteniéndose buenos resultados.

Modelado por patrones: está técnica proviene del reconocimiento geométrico de formas. Cada uno de los elementos a modelar, llamados clases, se suele representar mediante un conjunto de ejemplos. Para realizar la decodificación es necesario definir un medida de distancia y un criterio de decisión. Este tipo de modelado fue muy habitual al comienzo de la década de los 70 pero la popularización de los MOM, hace que actualmente haya pocas aplicaciones que los utilicen (Aibar, 1998).

4.1.2.- Selección de unidades

En la sección anterior se han presentado de forma resumida las posibilidades de modelado acústico que existen, pero se ha omitido una cuestión fundamental: ¿que debemos modelar?. La elección del tipo de unidades a utilizar es fundamental en el diseño de un decodificador Acústico-Fonético. Lo ideal sería encontrar una unidad independiente de contexto y que pueda ser modelada perfectamente a partir de un número finito de muestras. Desgraciadamente tal unidad aún no ha sido encontrada y lo que se hace a continuación es detallar los diferentes tipos de unidades propuestos en la bibliografía con sus ventajas y sus inconvenientes más generales.

4.1.2.1. Unidades independientes de contexto

Las **palabras** son las unidades más naturales en el lenguaje escrito, ya que son propiamente las palabras lo que queremos reconocer. Su principal ventaja es que son acústicamente estables, es decir capturan en su estructura la mayor parte de los fenómenos contextuales (por ejemplo, no es lo mismo una /n/ en 'trenza' que en 'cana', debido a los efectos coarticulatorios que se producen en la pronunciación) y la aplicación del nivel léxico no es necesario, lo cual simplifica bastante el proceso global de reconocimiento (Galiano, 1992) (Lee, 1988).

Como desventajas se puede destacar que las propiedades acústicas de los límites de las palabras (es decir el principio y fin de cada palabra) son fuertemente dependientes de la palabra anterior y posterior. Por ello para entrenar adecuadamente estos modelos es necesario tener un conjunto de entrenamiento tal que cada palabra en el vocabulario aparezca varias veces en cada contexto fonético (Lee et al, 1990). Para vocabularios pequeños, por ejemplo los dígitos del castellano, esta elección es la más adecuada y se han obtenido porcentajes de reconocimiento superiores al 99%. Pero si se pretende utilizar este modelo en sistemas de reconocimiento para tallas de vocabulario medias y grandes, el conjunto de entrenamiento que se necesita se hace prohibitivamente grande, además el coste computacional (básicamente espacial) crece linealmente con el número de palabras.

Por lo tanto, en la mayoría de las aplicaciones se hace necesario la elección de unidades subléxicas, estimables a partir de bases de datos fonéticas, independientes de la aplicación, y que permiten al combinarse, el modelado de cualquier palabra del idioma.

La primera propuesta de tipo subléxico son los **fonemas**, la propia fonología define los fonemas como la abstracción de los sonidos de una lengua. Así cuando en RAH se habla de unidad de tipo fonético se piensa en el conjunto de patrones acústicos que representan diferentes realidades acústicas (sonidos o fonos) de una misma unidad (abstracción).

La principal ventaja de los fonemas es que el número de componentes de este conjunto es pequeño (26 para el castellano), lo cual implica un alto porcentaje de cada unidad en el corpus de entrenamiento. Son independientes de contexto, lo que permite modelar por concatenación palabras nuevas que no están presentes en el corpus de entrenamiento. Hay que tener en cuenta que un sólo fonema, dependiendo de las modificaciones a que le dan lugar los sonidos que le rodean, puede variar su lugar de articulación dentro del aparato fonador humano dando lugar a diferentes sonidos que reciben el nombre de **alófonos** (Quillis, 1981).

4.1.2.2.- Unidades dependientes de contexto

Aunque decimos que cada palabra es una secuencia concatenada de fonemas, esos fonemas no se producen independientemente, porque las articulaciones humanas no se pueden mover instantáneamente de una posición a otra. Por eso, las características de un fonema están fuertemente relacionadas con los fonemas vecinos. Se recogen bajo la misma etiqueta o unidad, todas las variaciones alofónicas o sonidos asociados en la señal acústica a tal unidad. Por lo tanto, en ocasiones debido a la gran riqueza contextual se produce fácilmente una sobregeneralización del modelo representado.

Este problema se puede solucionar aumentando el número de modelos para cada unidad independiente de contexto o elegir unidades dependientes de contexto. En cualquier caso, debido a la excesiva cantidad de unidades existentes, nos encontramos con un grave problema a solucionar. Existen propuestas que llevan a reducir el número, basándose en la capacidad discriminante de las unidades y/o haciendo agrupaciones de aquellas que tienen características acústicas similares. En cualquier caso, es muy importante tener en cuenta la frecuencia con la que estas unidades aparecen en la lengua bajo estudio y en el corpus de entrenamiento. A continuación vamos a presentar una breve revisión de las unidades propuestas por los diversos autores:

Para incluir los efectos coarticulatorios se usan unidades más largas que los fonemas como pueden ser las **sílabas, demisílabas** que son las que comprenden el intervalo entre los puntos centrales (de máxima energía) de dos vocales consecutivas (Cravero et al, 1986).

Los **trifonemas** modelan cada fonema de una palabra en contexto. Un trifonema tiene en cuenta a sus vecinos más próximos por la derecha y por la izquierda, es decir si dos fonemas

tienen la misma identidad pero diferentes contextos por la derecha y/o por la izquierda son considerados trifonemas diferentes (Lee, 1988).

Otro tipo de unidad son los **fonemas con contexto.** A diferencia de los trifonemas que capturan el contexto por la derecha y por la izquierda, éstos sólo toman el contexto por la izquierda o por la derecha según sea más adecuado en cada caso.

Se puede hablar también de las **unidades que representan la zona transitoria del habla.** Puesto que lo que más nos preocupa son los efectos coarticulatorios que se producen en las transiciones entre los fonemas, una solución es modelar explícitamente estas transiciones por ejemplo utilizando difonemas. Los difonemas son el intervalo comprendido entre los puntos centrales de dos segmentos estacionarios de señal. Idealmente se asimilan a parejas de "semifonemas'" (mitades de fonemas). Los semifonemas han sido ampliamente estudiados en la Universidad Politécnica de Barcelona (Mariño et al, 1997) con muy buenos resultados.

Otra posible solución es utilizar los modelos estacionarios de los fonemas y a la vez modelar las transiciones propiamente dichas. Por cada N fonemas hay N^2 transiciones, por lo cual es conveniente tener en cuenta lo propuesto en (Cravero et al 1986). Es decir, la información incluida en la transición no es siempre significativa para la discriminación de la unidad en si misma. Por lo tanto, sí considerar explícitamente la transición ayuda al reconocimiento ese modelo es considerado, si no es así los modelos de los fonemas independientes de contexto se conectan directamente. En el Grupo de Reconocimiento Automático del Habla - Mintzo Ezagutza Automatikoaren Taldea (GRAH-MEAT) se obtuvieron algunos resultados con estas unidades (Varona et al, 1995).

La aproximación basada en **árboles de decisión** es una de las más usadas para la selección de unidades contextuales en RAH de los grandes sistemas actuales (Bahl et al, 1994) (Kuhn et al, 1995) (Odell, 1995). Los árboles de decisión combinan la ventaja de aplicar conocimiento fonético del lenguaje y un procedimiento de validación que se basa en la probabilidad de las muestras de voz con respecto a modelos probabilísticos. Además el conjunto de muestras obtenido a partir de arboles de decisión garantiza la cobertura de todas las dependencias contextuales.

4.1.3.- Condiciones experimentales

En los dos puntos previos se ha presentado una pequeña revisión de las posibles técnicas de modelado y de la selección del conjunto de unidades. En esta sección se presenta las características reales con las que se va a trabajar en la etapa de Decodificación Acústico Fonética. Estas características se mantendrán en el sistema completo junto con otras que se irán introduciendo a lo largo del capítulo. En primer lugar, se presenta la etapa de adquisición y preproceso de la secuencia acústica. Posteriormente se presentan los Modelos Ocultos de Markov (MOM) como elección de modelado acústico. En este punto se presentan las técnicas de aprendizaje de los parámetros de los modelos en general y se presenta el reconocimiento en

DAF. En el tercer punto se presentan las unidades subléxicas básicas elegidas: las unidades incontextuales de tipo fonético y unidades contextuales obtenidas a partir de árboles de decisión. Finalmente se presentan los criterios de evaluación elegidos para presentar los resultados de DAF.

4.1.3.1.- La secuencia acústica: adquisición y preproceso

La señal acústica se compone de una secuencia de vectores de parámetros indexados temporalmente. Cada elemento de la secuencia, el vector de parámetros, caracteriza un segmento de señal de voz con características estables. Las señales acústicas de las bases de datos utilizadas, han sido obtenidas en diferentes condiciones ambientales (diferentes lugares físicos: Universidad Politécnica de Barcelona, Universidad Politécnica de Valencia y Universidad del País Vasco (ver apéndice A)) y con distintos soportes de adquisición de audio. Sin embargo las características tanto de adquisición como de preproceso son comunes, no sólo a todos los corpus de voz, sino también la señal del prototipo TORNASOL.

- Frecuencia de muestreo de la señal: 16kHz

- Número de bits para cada muestra: 16

- Ancho de banda de la señal: 250Hz-7,2kHz

- Longitud de los tramos de análisis: 25 milisegundos

- Tipo de ventana: Hamming

- Frecuencia de obtención de características o sub-muestreo: 100Hz (10 milisegundos)

- Preénfasis: filtro paso alto de primer orden: $H(z)=1-0{,}97z^{-1}$

Para realizar una buena caracterización espectral se suele utilizar una estimación no paramétrica basada en un banco de filtros. Su implementación se realiza habitualmente a través de un módulo cuadrado de la transformada discreta de Fourier. Las frecuencias centrales de los filtros y el ancho de banda se diseñan según escalas logarítmicas (la escala Mel). Además es común la utilización de una transformación posterior sobre la información espectral, el cepstrum, calculado como el desarrollo en serie de Fourier del módulo del logaritmo del espectro de potencia de la señal (Lee, 1989). Para mejorar las prestaciones de muchos de los reconocedores, el conjunto de parámetros cespstrales se completa con otras informaciones derivadas de estos, habitualmente por una o más diferenciaciones de los parámetros (Partalo & Sijercic, 1989).

La representación final que se utiliza consta de 28 parámetros:

-12 parámetros cepstrales obtenidos a partir del espectro de potencia en escala Mel (número de filtros 22).

-12 parámetros correspondientes a la estimación de la derivada de dichos coeficientes respecto del tiempo. Para la estimación de la derivada se utilizan los parámetros cepstrales de los dos tramos anteriores y de los dos tramos posteriores.

-12 parámetros correspondientes a la estimación de la segunda derivada de los coeficientes cepstrales respecto del tiempo.

- Energía normalizada del tramo y su primera derivada obtenida de mismo modo que los parámetros cepstrales.

Los sistemas pueden trabajar directamente con estas secuencias de vectores de características, pero también es posible realizar una discretización de los mismos para reducir el coste. Esta discretización consiste en etiquetar cada una de las ventanas de análisis mediante símbolos ("*Codewords*") de un alfabeto asociado a un diccionario ("*Codebook*"). El etiquetado se obtiene mediante un proceso de Cuantificación Vectorial (CV) que particiona el espacio acústico en un conjunto de clases no solapadas, cada una de las cuales agrupará vectores de características semejantes (con una distorsión mínima). El algoritmo que se usa es de agrupamiento no supervisado de k-medias (Duda & Hart, 1993). Para trabajar con modelos discretos y semicontinuos se ha realizado una Cuantificación Vectorial a 256 codebooks para cada uno de los cuatro conjuntos: Coeficientes Cepstrales (CC), la primera derivada de los CC (DC), la segunda derivada de CC (DDC) y para la energía y la derivada de la energía (EDE). Se ha utilizado la regla de los k vecinos más próximos.

4.1.3.2.- MOM: aprendizaje y reconocimiento

En este trabajo, a nivel de modelado acústico se ha optado por utilizar, quizá, uno de los métodos más consolidados, los Modelos Ocultos de Markov (MOM) (Lee, 1989) (Rabiner, 1989). Esta decisión se ha basado fundamentalmente en dos puntos: a) probablemente, son los modelos acústicos más utilizados por la comunidad científica internacional, b) en nuestra propia experiencia como grupo de Reconocimiento Automático del Habla. En el GRAH-MEAT están desarrolladas desde hace algunos años las herramientas necesarias para realizar la estimación de los parámetros de los MOM tanto discretos (Torres, 1992) (Varona, 1993) como semicontinuos (Torres & Casacuberta, 1993) (Torres et al, 1994) y los algoritmos necesarios para su posterior utilización en reconocimiento.

Los MOM constan de dos partes muy bien diferenciadas: a) una cadena de Markov que consiste en un conjunto de estados, probabilidad de transición entre estados y probabilidades de los estados iniciales. b) cada uno de los estados tiene asociada una distribución de probabilidad que indica la probabilidad de que ocurra una determinada observación en ese estado.

Definición (Castro, 1999): un modelo oculto de Markov (MOM) se puede describir como una tupla de 6 elementos $(Q, q^F, A, P_I, P_T, P_E)$ donde:

- $Q=\{q^1, q^2, ...,q^{|Q|}\}$ es el conjunto de estados finito y no vacío. En cada instante discreto de tiempo t, el modelo se encuentra en un estado de Q, que denotaremos con q_t.

- q^F es un estado especial que representa la parada del modelo. Formalmente q^F es el estado de parada y no pertenece a Q.

- A (secuencia acústica) es el conjunto de elementos observables. En cada instante discreto de tiempo t, el modelo emitirá un elemento de A que denotaremos como a_t.

- P_I: $Q \rightarrow [0, 1]$, $\Sigma_{q \in Q} P_I(q)=1$, es la distribución de probabilidad inicial sobre los estados de Q. Representa las probabilidades iniciales de los estados, de modo que $P_I(q)=P(q_1=q)$ para todo estado q de Q.

- P_T: $Q \times (Q \cup \{q^F\}) \rightarrow [0, 1]$, $\Sigma_{q' \in Q \cup \{qF\}} P_T(q'/q)=1$, $\forall q \in Q$, es la distribución de probabilidad sobre los estados de Q que representa las probabilidades de transición desde el estado q, de modo que $P_T(q'/q)=P(q_{t+1}=q' | q_t=q)$ es la probabilidad de transición desde el estado q al estado q´ para todo q´ de $Q \cup \{q^F\}$ y para todo $t \geq 1$.

- P_E: $Q \times A \rightarrow [0, 1]$, $\Sigma_{a \in A} P_E(a/q)=1$, $\forall q \in Q$, es la distribución de probabilidad sobre los elementos de A que representa las probabilidades de emisión (o densidades de probabilidad) para cada estado q de Q, de modo que $P_E(a/q)=P(a_t=a | q_t=q)$. Se define como la probabilidad de observar $a \in A$ en el estado q, en el instante de tiempo $t \geq 1$.

Un ejemplo de MOM con una topología ya fijada se puede ver en la figura 4.2. Los MOM se suelen clasificar en función del tipo de distribución de probabilidad que se utiliza para modelar las secuencias acústicas en cada estado. El caso más simple es el de los **modelos discretos**, donde las posibles observaciones $a \in A$ asociadas a cada estado son finitas (tras un proceso de Cuantificación Vectorial (CV)). Cuando las distribuciones de probabilidad tratan directamente con los vectores de parámetros se habla de **modelos continuos**. La representación más general de la probabilidad de observar el vector $a \in A$ en el estado q es una mezcla de M funciones densidad de probabilidad (fdp) normales gausianas (diferentes para cada uno de los estados q) (Huang et al, 1990) de la forma:

$$P_E(a/q) = \sum_{m=1}^{M} c_{qm} \eta(a, \mu_{qm}, \sigma_{qm}) \qquad (4.1)$$

donde η es una fdp gausiana con un vector media μ_{qm} y matriz de varianzas o covarianzas σ_{qm} para la componente de la distribución m en el estado q. c_{qm} es un factor de peso para la distribución m-ésima del vector *a* en el estado q.

Las distribuciones continuas permiten modelar mejor los fenómenos acústicos que no son de naturaleza discreta. Sin embargo, para lograr buenos resultados es necesario el uso de una gran cantidad de fdp. Esto implica la reestimación de un gran número de parámetros y de bases de datos acústicas muy grandes. Una solución intermedia son los **modelos semicontinuos**, que utilizan también fdp pero todos los estados de todos los modelos comparten el mismo conjunto de fdp. El número de fdp coincide normalmente con el número de codebooks derivados del proceso de CV para discretizar la señal, aunque ésta no es una condición necesaria (ver figura 4.1). En cada estado existen coeficientes que ligan el grado de afinidad con cada una de las fdp (Huang & Jack, 1989) (Huang, 1992). Por tanto, la probabilidad de emisión de una observación $a \in A$ se estima de acuerdo a:

$$P_E(a/q) = \sum_{m=1}^{M} P_E(s_m/q)\eta(a, \mu_m, \sigma_m) \qquad (4.2)$$

donde s_m es la etiqueta correspondiente a cada clase m∈(1..M) y la fdp de cada clase m se asume independiente del estado q.

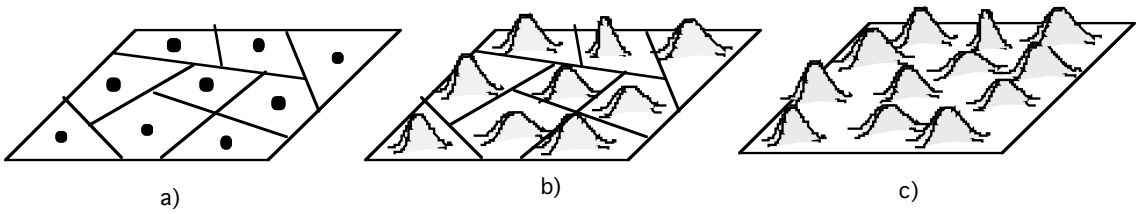


Figura 4.1.- (a) Espacio acústico cuantificado vectorialmente. (b). Espacio acústico convertido en conjuntos de mezclas de fdp, una por cada elemento que elemento que resultaría de la CV. (c) Espacio acústico con densidad de probabilidad gausianas independientes.

Aprendizaje: el aprendizaje de los parámetros se realiza en dos etapas: deductiva e inductiva. En la etapa deductiva, el diseñador del sistema de RAH fija ciertas características del modelo, basándose en su conocimiento a priori del problema. Tras algunos experimentos preliminares se fijo la topología a tres estados emisores por unidad subléxica, con probabilidad de transición al propio estado y al siguiente y sin posibilidad de omisión de ninguno de los estados al modelar las secuencias. Tienen un único estado inicial, el primero, y un único estado final, el último (figura 4.2).

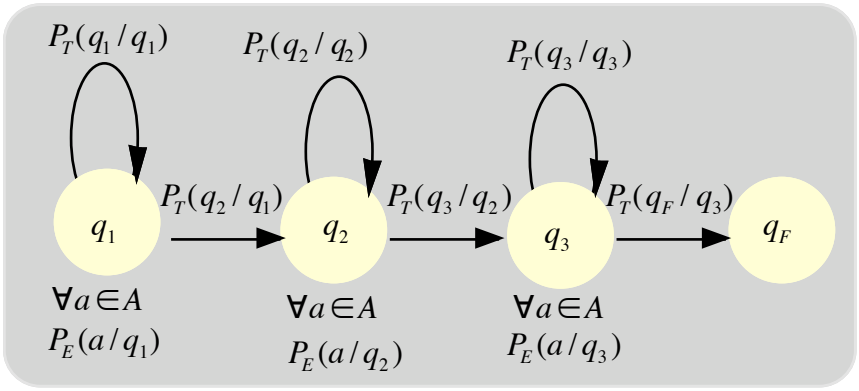


Figura 4.2.- Topología de los MOM utilizados en la experimentación.

En la etapa inductiva se deben estimar el resto de los parámetros de los MOM. Para ello se necesitan bases de datos con una adecuada distribución acústica y con un número suficiente de muestras. Se asume que una secuencia acústica A se ha generado por un autómata finito construido a partir de la concatenación de un conjunto de modelos de M: $M=m_1 m_2 ... m_{|M|} \in M^+$. Dada la muestra de aprendizaje S formada por pares compuestos de una secuencia acústica A y la secuencia de modelos M asociada a su interpretación correcta, se desea estimar el conjunto de parámetros de M que maximice una determinada "función objetivo". En este caso:

$$P(A^S/M^S) = \sum_{Q \in Q(M^S/A^S)} \prod_{t=1}^{|A|} P_E(a_t/q_t) P_T(q_t/q_{t-1}) \qquad (4.3)$$

siendo $Q(M^S/A^S)$ las secuencias de estados de M^S que pueden generar A^S.

Este problema de optimización constituye un problema clásico de estimación de probabilidades, según el criterio de máxima verosimilitud, para el que se conoce una solución eficiente: el algoritmo de Baum-Welch (Rabiner & Juang, 1989). Pero la evaluación directa de la ecuación 4.3 requiere la enumeración de todos los caminos de $Q(M^S/A^S)$ y presenta un coste prohibitivo. Otra solución viene dada por el algoritmo de Viterbi (Jelinek, 1976). Sólo se considera el camino de máxima probabilidad (mejor secuencia de estados), lo que simplifica enormemente los cálculos. En realidad se trata de una aproximación para el calculo de $P(A^S/M^S)$.

$$\hat{P}(A^S / M^S) = \max_{Q \in Q(M^S / A^S)} \prod_{t=1}^{|A|} P_E(a_t / q_t) P_T(q_t / q_{t-1}) \tag{4.4}$$

En este trabajo, para el entrenamiento de los modelos se cuenta con un pequeño conjunto de segmentos de cada unidad obtenidos a partir de un subconjunto de frases segmentadas manualmente en la UPV/EHU (ver apéndice A). Los modelos se entrenan individualmente a partir de estos segmentos utilizando el algoritmo de Baum-Welch. A partir de esa semilla inicial se utiliza 3 iteraciones del algoritmo subóptimo de Viterbi para el reentrenamiento de las unidades. Intentar la estimación de los parámetros de los modelos sin utilizar la segmentación manual inicial lleva a resultados no demasiado buenos (Varona, 1993).

Reconocimiento: en esta etapa, dado un conjunto de modelos entrenados M y una secuencia acústica $A=a_1...a_T$ de longitud T, se trata de encontrar la secuencia de modelos $\hat{M}$ que con más probabilidad pueda haber generado esa secuencia acústica A. Cada uno de los modelos de los fonemas o unidades contextuales se coloca en paralelo con una probabilidad entre unidades de transitar al resto de las unidades igual a 1/número de unidades, dando lugar al reconocimiento a nivel de DAF.

Se utiliza el algoritmo de Viterbi, de manera que la secuencia de estados más probable se recupera al final de la búsqueda (Rabiner, 1989). Dado un conjunto de Q(M) estados (todas las posibles secuencias de estados asociadas a M), la probabilidad asociada a cada estado $q \in Q(M)$ en el instante de tiempo t, $P_t(q)$, se calcula como:

$$P_t(q) = \max_{1 \le q' \le Q(M)} (P_{t-1}(q') P_T(q' / q)) P_E(a_t / q) \; 1 \le q \le Q(M) \quad 2 \le t \le T \tag{4.5}$$

Al llegar al instante final, se recorre el camino inverso lo que permite obtener la concatenación de estados que dan la máxima probabilidad. La información de vuelta atrás que se guarda para poder recuperar el mejor camino es:

$$\Psi_t(q) = \arg\max_{1 \le q' \in Q(M) \le N} (P_{t-1}(q') P_T(q / q')) \; 1 \le q \le Q(M) \; 1 \le t \le T \tag{4.6}$$

Desde el punto de vista del modelado estadístico, este criterio de búsqueda de la secuencia de estados de máxima probabilidad es únicamente una aproximación del criterio estadísticamente correcto y por tanto es subóptimo. También se pueden aplicar otros algoritmos como el de máxima discriminación o mínimo error (Juang & Katagiri, 1992).

4.1.3.3.- Unidades acústicas: fonemas y unidades contextuales obtenidas a partir de árboles de decisión

Una vez elegido el tipo de modelado acústico, habrá que decidir el tipo de unidades a modelar. El objetivo final es el reconocimiento de palabras. Sin embargo, ya hemos visto que el entrenamiento de palabras es inviable por lo que hay que optar por unidades de tipo fonético. Si A pesar de presentar inconvenientes propios debido a la variabilidad acústica y la dificultad de identificación de los límites, están relacionadas con las unidades lingüísticas y por tanto se puede caracterizar fácilmente el vocabulario de una aplicación concreta independientemente del número de palabras.

Llisterri propuso un conjunto de unidades alofónicas para la elaboración de una base de datos para el reconocimiento del Habla en Castellano dentro del marco del proyecto Albaycin elaborado por la Universidad Autónoma de Barcelona (Llisterri, 1991). Posteriormente partiendo del conjunto de alófonos del castellano (33 alófonos), siguiendo criterios cercanos al fonológico, y aplicando criterios de similitud de propiedades acústicas se restringió la lista inicial y se elaboró la propuesta del conjunto de unidades para el castellano. Los detalles del trabajo aparecen en (Torres, 1992) dando lugar al conjunto de 23 unidades que aparecen en la tabla 4.1.

En la mayor parte de los resultados presentados en esta memoria se han utilizado directamente estas unidades fonéticas. Se han elegido fundamentalmente porque su número es muy reducido (23 unidades + Silencio para el Castellano) y porque aparecen suficientemente bien representados en bases de datos acústicas no demasiado grandes.

Dentro del GRAH-MEAT de la UPV/EHU se han obtenido **unidades contextuales para el Castellano utilizando árboles de decisión** (López et al, 1999a). Estas unidades, como todas las unidades contextuales, tienen en cuenta los contextos no sólo dentro de la palabra sino también en los bordes.

Los árboles binarios se construyen a partir de muestras de entrenamiento teniendo como base las unidades independientes de contexto. Cada nodo del árbol de decisión representa una pregunta sobre la identidad fonética de uno o más contextos a derecha e izquierda, por tanto, cada una de las muestras de entrenamiento se traslada desde el nodo raíz a uno de los dos nodos hojas, que va a representar una categoría de contexto dependiente de la pregunta que se haya hecho (Bahl et al, 1994) (Odell, 1995). La división se realiza atendiendo principalmente a dos criterios: la evaluación de una medida de calidad que representa cuanto se incrementa la probabilidad de las muestras al producirse la división en los dos conjuntos y un número mínimo de muestras para cada unidad para garantizar el adecuado entrenamiento de las unidades. Los nodos hojas son alófonos contextuales de los fonemas incontextuales originales, por tanto, no queda ninguna unidad independiente de contexto dentro de una frase sin clasificar (existe cobertura total). Todas pueden ser clasificadas respondiendo a las preguntas de cada uno de los nodos del árbol hasta llegar a un nodo hoja (López et al, 1999b).

Tabla 4.1.- Conjunto de unidades para el Castellano (notación AFI), sus grafías y la notación utilizada por el transcriptor.

Clasificación por categoría fonéticas	Símbolo AFI asociado	Grafías	Transcripción del sistema
Oclusiva sordas	[p] [t] [k]	p, t, k, q, c (salvo ce, ci) en posición silábica prenuclear	' p ', ' t.', ' k '
Oclusivas sonoras	[b] [d] [g]	b, v, d, g (salvo ge, gi) todos los contextos p, t, k en posición silábica postnuclear	' b ', ' d ', ' g '
Nasales	[m] [n] [η]	m, n, ñ en todos los contextos	' m ', ' n ',' ñ '
Fricativas	[f] [θ] [s] [y] [x]	f todos los contextos, za, ce, ci, zo, zu. s todos los contextos y, hi inicial de palabra ll , j, ge, gi	' f ', ' z ', ' s ', ' y ', ' x '
Africada	[tf]	ch todos los contextos	' c'
Líquidas	[l] [r] [rr]	l todos los contextos r en posición intervocálica o grupo tautosilábico rr, r inicial de palabra.	' l ', ' r ', ' @ '
Vocales	[i] [e] [a] [o] [u]	i todos los contextos, conjunción y e, a, o, u, todos los contextos	' i ', ' e ',' i ', ' o ', ' u '

4.1.3.4.- Criterios de evaluación

La evaluación de un sistema de DAF presenta dificultades especificas que no están claramente resueltas en la literatura. Se presentan problemas derivados de las imprecisiones en la pronunciación de las frases de evaluación. Los porcentajes de errores se obtienen tras comparar la secuencia fonética obtenida como salida de un decodificador con la correspondiente transcripción fonética de la frase pronunciada. Sin embargo, se sabe que en las pronunciaciones se producen errores fonéticos muy a menudo, por lo que la comparación entre la salida del decodificador y la secuencia pronunciada (supuestamente sin errores) no es del todo correcta. La única forma de resolver este problema supone conocer siempre cual ha sido la secuencia que verdaderamente se ha pronunciado y eso supone un análisis supervisado de todas las frases, escuchándolas y observando los espectrogramas. Lo cual es inviable en bases de datos relativamente grandes (Castro, 1999).

Este problema no existe cuando la etapa de DAF esta incluida dentro del sistema completo de reconocimiento ya que se utiliza información de niveles superiores como léxico, sintáctico o semántico. En este caso la salida no será una secuencia fonética, sino una secuencia ortográfica de palabras.

Vamos a ver a continuación cómo medir el grado de acierto (o de error) de la etapa de DAF. Los resultados que se plantean en este trabajo se obtienen a partir del cálculo de la distancia (de Levenshtein) de edición entre la cadena fonética obtenida en la decodificación y la

transcripción fonética de referencia. El algoritmo de programación dinámica (programación fracional) que calcula esta distancia proporciona el número de caracteres correctos (f_acierto), número de sustituciones (f_sustituciones), número de borrados (f_borrados) y número de inserciones (f_inserción). Las distintas medidas que aparecen en la literatura suelen variar según los factores por los que se normalizan estas medidas (Prat et al, 1994). Tomaremos como medida de bondad del decodificador acústico fonético el porcentaje de unidades correctamente reconocidas también llamado porcentaje real de aciertos (PRA_{DAF}).

$$\%PRA_{DAF} = \frac{f_aciertos}{f_aciertos + f_sustituciones + f_borrados + f_inserciones} * 100 \qquad (4.7)$$

En esta medida se incluye el número de fonemas insertados (f_inserciones) lo que provoca unos resultados más pesimistas que los que se obtendrían si no añadiera este número a la suma del denominador.

Además, para algunos experimentos se muestra el intervalo de confianza $[PRA^{-},PRA^{+}]_{95\%}$ de los resultados obtenidos de acuerdo a la propuesta presentada en (Chollet, 1995). Si p es el valor promedio de los fonemas correctamente reconocidos (p=PRA/100) y N es el conjunto de fonemas que aparecen en el conjunto de evaluación (el tamaño de la muestra). El intervalo de confianza del valor real de la medida PRA se obtiene como:

$$PRA^{\pm} = \frac{p + \frac{z^2}{2N} \pm z\sqrt{\frac{p(1-p)}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} * 100 \qquad (4.8)$$

Para valores muy grandes de N se puede simplificar como:

$$PRA^{\pm} = p \pm z\sqrt{\frac{p(1-p)}{N}} * 100 \qquad (4.9)$$

donde z=1.96. Este valor indica que el "verdadero" porcentaje de reconocimiento tiene una probabilidad del 95% de encontrarse en el intervalo de confianza. Se dice que son valores estadísticamente distintos cuando los intervalos de confianza no se solapan, si se solapan aunque sean distintos son resultados estadísticamente equivalentes.

4.1.4.- Resultados de reconocimiento

En esta sección se van a presentar los porcentajes de reconocimiento obtenidos en la etapa de DAF. Ésto nos dará una idea de la bondad tanto de los modelos acústicos como de las unidades subléxicas bajo estudio. Como primera opción se va a evaluar el comportamiento de los MOM discretos utilizando como unidades fonéticas las unidades incontextuales. Para tratar de mejorar los resultados obtenidos con MOM discretos que representan unidades incontextuales existen fundamentalmente dos opciones:

a) utilizar modelos más complejos (por ejemplo modelos semicontinuos) y mantener el número de unidades subléxicas reducido (fonemas).

b) seguir usando los modelos acústicos simples (discretos) e incrementar el número de unidades (por ejemplo utilizando unidades contextuales obtenidas a partir de árboles de decisión).

Los resultados presentados en este punto, en primer lugar tratan de evaluar el comportamiento de los modelos semicontinuos frente a los modelos discretos (utilizando distinto número de codebooks) y utilizando como unidades básicas las unidades fonéticas incontextuales. En segundo lugar, se va a evaluar el comportamiento de las unidades contextuales obtenidos a partir de árboles de decisión.

4.1.4.1.- Unidades fonéticas incontextuales: fonemas

En esta sección se presentan los resultados de reconocimiento obtenidos de la etapa DAF cuando los modelos acústicos representan a los fonemas. Los modelos acústicos han sido entrenados con el conjunto de entrenamiento de SENGLAR (ver apéndice A) que consta de 1529 frases pronunciadas por 57 locutores. La evaluación de los modelos acústicos se lleva a cabo con el conjunto de prueba que es independiente del locutor y del vocabulario (test acústico). Esta formado por 700 frases y 33 locutores distintos (32034 fonemas). Se presenta además el reconocimiento a nivel acústico para las dos tareas de la aplicación (MLA (108576 fonemas) y BDGEO (25098 fonemas)) (ver apéndice A). Hay que destacar que tanto para la tarea MLA como BDGEO, los conjuntos de evaluación no corresponden a conjuntos de evaluación fonéticos y por tanto la distribución de los fonemas no es la más adecuada, como se puede ver también en el apéndice A, para obtener resultados de decodificación acústico fonética.

En la tabla 4.2 se presentan los resultados obtenidos utilizando modelos discretos y en la tabla 4.3 utilizando los modelos semicontinuos. Se presentan también resultados para diferentes números de codebooks:

* 2MC considera dos conjuntos de vectores:	Los cepstrales
	Energía y su derivada
* 3MC considera tres conjuntos de vectores:	Los cepstrales
	Primera derivada de los cepstrales
	Energía y su primera derivada
* 4MC considera cuatro conjuntos de vectores:	Los cepstrales
	Primera derivada de los cepstrales
	Segunda derivada de los cepstrales
	Energía y su primera derivada

En ambas tabla 4.2 y tabla 4.3, se muestran los resultados en la tercera iteración del algoritmo de Viterbi. Se ha introducido un peso entre cada fonema igual a 1/número_unidades

(1/24). También se muestran los intervalos de confianza $[PRA_{DAF}^{-}, PRA_{DAF}^{+}]_{95\%,}$. Se pueden extraer las siguientes conclusiones:

- Con respecto al reconocimiento, las mejores tasas se dan para el test_acústico. Para la tarea MLA los resultados se degradan bastante. A pesar de que se trata de una tarea muy sencilla de 30 palabras y que como consecuencia, con ML a nivel de palabra se obtendrán los mejores resultados (ver apéndice A y secciones 4.2 y 4.3). Cabe destacar que en el conjunto de frases de evaluación no están presentes algunos fonemas y que la distribución de los restantes está bastante desequilibrada con respecto al conjunto representado en el test acústico.

Tabla 4.2.- Resultados de reconocimiento $\%PRA_{DAF}$ en DAF utilizando modelos discretos.

	test acústico		MLA		BDGEO	
	%PRA	$[PRA_{DAF}^{-}, PRA_{DAF}^{+}]_{95\%}$	%PRA	$[PRA_{DAF}^{-}, PRA_{DAF}^{+}]_{95\%}$	%PRA	$[PRA_{DAF}^{-}, PRA_{DAF}^{+}]_{95\%}$
2MC	55.84	[55.30,56.38]	52.23	[51.94,52.52]	54.82	[54.20,55.43]
3MC	60.32	[59.78,60.85]	57.93	[57.64,58.22]	59.33	[58.72,59.94]
4MC	**63.82**	[63.32,64.37]	**61.56**	[61.27,61.85]	**62.60**	[62.00,63.20]

Tabla 4.3.- Resultados de reconocimiento $\%PRA_{DAF}$ en DAF utilizando modelos semicontinuos.

	test acústico		MLA		BDGEO	
	%PRA	$[PRA_{DAF}^{-}, PRA_{DAF}^{+}]_{95\%}$	%PRA	$[PRA_{DAF}^{-}, PRA_{DAF}^{+}]_{95\%}$	%PRA	$[PRA_{DAF}^{-}, PRA_{DAF}^{+}]_{95\%}$
2MC	58.82	[58.28,59.36]	51.33	[51.04,51.62]	57.56	[56.99,58.17]
3MC	62.40	[61.87,62.93]	56.97	[56.68,57.26]	62.06	[61.46,62.66]
4MC	**65.07**	[64.55,65.59]	**60.89**	[60.60,61.18]	**64.30**	[63.71,64.89]

- Cuando se trabaja con modelos discretos el uso de 4 codebooks incrementa notablemente las tasas de reconocimiento con respecto a 3 codebooks (más de tres puntos) y a 2 codebooks (aproximadamente 8 puntos).

- Cuando se trabaja con modelos semicontinuos este incremento es algo menor aunque también se aprecia una mejora significativa. Con respecto a 3 codebooks (aproximadamente 2 puntos) y a 2 codebooks (aproximadamente 7 puntos).

Vemos que cuando se trabaja con 4MC la diferencia entre los modelos discretos y semicontinuos es de aproximadamente un punto y medio y sin embargo los modelos semicontinuos son mucho más costosos computacionalmente (trabajan con funciones densidad de probabilidad, en vez de usar valores de emisión discretos, lo cual supone incrementar los cálculos en más operaciones por cada una de las que se realizan con los modelos discretos llegándose a un incremento del tiempo de decodificación requerido de más de 10 veces).

Por otra parte, durante los últimos años se han ido abandonando los modelos discretos para pasar a trabajar con los modelos continuos. En este punto y una vez vistos los resultados obtenidos por la mayoría de los grupos de investigación, parece resurgir la idea de llegar a un compromiso y usar modelos semicontinuos para limitar el número de gausianas necesarias. Es

más, se vuelve a pensar que quizá la potencia de los modelos discretos ha quedado oculta por el optimismo de los defensores de los modelos continuos (Digalakis & Neumeyer, 1999) sobre todo en los grupos de investigación en los que la cantidad de los datos destinados a entrenamiento es bastante limitada.

Todo ello nos lleva a proponer los MOM discretos de 4MC para realizar la experimentación que se llevará a cabo a lo largo de la memoria. El otro punto de interés a evaluar, consiste en mantener los modelos acústicos discretos e incrementar el número de unidades acústicas, los resultados obtenidos se muestran a continuación.

4.1.4.2.- Unidades contextuales

En este punto se van a presentar los resultados obtenidos en DAF para distintos conjuntos de unidades contextuales. El objetivo principal es la evaluación de los resultados obtenidos cuando se usaron unidades contextuales obtenidas a partir de árboles de decisión. También se evaluaron los trifonemas obtenidos tras contar el número de apariciones en la base de datos de entrenamiento. Con ambas aproximaciones se han evaluado conjuntos con distintos números de unidades.

En la tabla 4.4 se presentan los resultados que se obtuvieron a partir de las unidades contextuales (trifonemas obtenidos por conteo o unidades contextuales obtenidas a partir de árboles de decisión). Se utilizan los MOM discretos de 4MC.

Tabla 4.4.- Resultados de DAF con diferentes conjuntos de unidades subléxicas, para el test acústico.

	número de unidades	$\%PRA_{DAF}$	$[PRA^-, PRA^+]_{95\%}$
Incontextuales	23	63.82	[63.29,64.34]
Trifonemas obtenidos por conteo	103	65.90	[65.38,66.42]
	209	65.93	[65.41,66.45]
	347	65.18	[64.66,65.70]
Unidades de árboles de decisión	**101**	**66.44**	[65.92,66.95]
	178	66.30	[65.78,66.81]
	359	65.75	[65.23,66.27]

- Se observa que las unidades contextuales ayudan a un mejor reconocimiento, obteniéndose aproximadamente 2 puntos y medio de incremento del reconocimiento en DAF.

- Se ve que para este conjunto de evaluación, incrementar el número de unidades no significa un incremento de las tasas y se obtiene un buen resultado alrededor de 100 unidades.

- Las unidades obtenidas a partir de árboles de decisión proporcionan resultados en números absolutos algo mejores que los trifonemas obtenidos por conteo pero no son resultados estadísticamente independientes.

Para ver si el uso de estas unidades proporciona realmente beneficios, habrá que hacer la decodificación de palabras como se verá en la siguiente sección. A lo largo de la memoria

cuando se utilicen unidades distintas a los fonemas serán unidades contextuales obtenidas a partir de árboles de decisión (conjunto de 101 unidades).

En DAF el incremento del número de unidades no se ve recompensado por el incremento en las tasas de reconocimiento. Al colocarse todos los modelos en paralelo es necesario un incremento notable del tiempo de decodificación. Sin embargo este problema no se va a dar en los experimentos a nivel de palabras. Se parte de una única transcripción para cada palabra (en fonemas o unidades contextuales) por lo que el uso de unidades contextuales sólo supone un incremento de la memoria necesaria para guardar los MOM individuales. Cuando se trabaja a nivel de palabras, surge el problema derivado del desconocimiento de la unidad que se debe poner al principio y al final de cada palabra, puesto que los contextos no están definidos como se ve en la siguiente sección.

4.2.- Decodificación de palabras aisladas

En esta sección se van a presentar los resultados que proporciona el decodificador de palabras aisladas. Se trata de reconocer los modelos de palabras que se obtienen como concatenación de los modelos individuales de las unidades fonéticas, que componen la transcripción fonética de cada una de las palabras. El reconocimiento se lleva a cabo mediante el algoritmo de Viterbi. Los modelos de las palabras se colocan en paralelo con una probabilidad equiprobable de transitar de uno a otro. Los resultados que se obtienen sirven como base de la experimentación que se realizará en las siguientes secciones cuando el reconocimiento sea de palabras pero con la probabilidad de transición entre palabras dada por el ML.

4.2.1.- Criterios de evaluación

Cuando el decodificador da como resultado palabras, ya sea a nivel de palabras aisladas o por la inclusión de un ML, se utiliza de nuevo la distancia de Levenshtein para obtener en este caso la cuenta de palabras reconocida correctamente (w_acierto), sustituidas (w_sustituciones), borradas (w_borrados) e insertadas (w_inserción). A partir de esos valores se han obtenido los siguientes parámetros para evaluar el reconocimiento en palabras:

Porcentaje Real de Aciertos:

$$\%PRA = \frac{w_aciertos}{w_aciertos + w_sustituciones + w_borrados + w_inserciones} * 100 \qquad (4.10)$$

Porcentaje Real de Sustituciones:

$$\%PRS = \frac{w_sustituciones}{w_aciertos + w_sustituciones + w_borrados + w_inserciones} * 100 \qquad (4.11)$$

Porcentaje Real de Borrados:

$$\%\text{PRB} = \frac{\text{w_borrados}}{\text{w_aciertos} + \text{w_sustituciones} + \text{w_borrados} + \text{w_inserciones}} * 100 \quad (4.12)$$

Porcentaje Real de Inserciones:

$$\%\text{PRI} = \frac{\text{w_inserciones}}{\text{w_aciertos} + \text{w_sustituciones} + \text{w_borrados} + \text{w_inserciones}} * 100 \quad (4.13)$$

Y por tanto el Porcentaje Real de Errores será:

$$\%\text{PRE} = \frac{w_sustituciones + w_borrados + w_inserciones}{w_aciertos + w_sustituciones + w_borrados + w_inserciones} * 100 \quad (4.14)$$

Además se presenta para algunos experimentos el intervalo de confianza $[PRE^-,PRE^+]_{95\%}$ de los resultados obtenidos de la misma forma que se daba en DAF siendo p es el valor promedio de las palabras correctamente reconocidas (p=PRE/100) y N es el conjunto de palabras que aparecen en el conjunto de evaluación (el tamaño de la muestra).

$$PRE^{\pm} = \frac{p + \frac{z^2}{2N} \pm z\sqrt{\frac{p(1-p)}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} * 100 \quad (4.15)$$

Tanto el PRA como el PRE son realmente porcentajes, por lo que su valor puede variar (únicamente) entre el 0% y el 100%. Además no se ignoran las inserciones, de modo que el valor óptimo de ambas tasas (0% de errores/100% de aciertos) solo se alcanza en ausencia de errores. No obstante en las publicaciones de la comunidad científica internacional la medida que se suele dar es la del "*Word Error Rate*" (%WER) cuyo complementario no es el porcentaje de aciertos sino la precisión (Prat et al, 1994).

$$\%\text{WER} = \frac{w_sustituciones + w_borrados + w_inserciones}{w_aciertos + w_sustituciones + w_borrados} * 100 \quad (4.16)$$

4.2.2.- Resultados de reconocimiento

En esta sección se presentan los resultados de reconocimiento de palabras aisladas aplicando diferentes factores de peso entre cada palabra para al tarea de aplicación MLA (29 palabras + Silencio) y para la tarea BDGEO (1208 palabras + Silencio). El conjunto de evaluación de voz de la tarea MLA consta de 1600 frases y 26664 palabras y el de la tarea BDGEO consta de 600 frases y 5655 palabras. Los experimentos se van a realizar con los modelos discretos de 4 codebooks y los dos conjuntos de unidades fonéticas: fonemas y las unidades contextuales obtenidas a partir de árboles de decisión.

4.2.2.1.- Unidades incontextuales: fonemas

En la tabla 4.5 se presentan los resultados obtenidos para la tarea MLA cuando los modelos individuales con los que se construyen las palabras representan a los fonemas. Se aplican distintos factores de peso entre palabras: 1/1 (no hay factor de peso), 1/num_uni =1/29

(el factor de peso depende del número de unidades distintas). Con 1/5000, 1/8000, 1/10000 los factores de peso cada vez se va haciendo cada vez más pequeños. A medida que el factor de peso se hace más pequeño, las sentencias con un alto número de palabras (muchas transiciones) tendrán una probabilidad muy baja y no serán reconocidas. La consecuencia inmediata es la disminución de la inserción. En la tabla 4.6 se presentan los resultados para BDGEO. En este caso 1/um_uni= 1/1208.

Tabla 4.5.- Resultados de reconocimiento de palabras aisladas aplicando distintos factores de peso entre palabras (MLA).

Factor de peso	%PRE	%PRA	%PRS	%PRI	%PRB
1/1	42.88	57.12	19.15	22.12	1.61
1/num_uni	**36.66**	**63.34**	**14.56**	**17.69**	**4.45**
1/5000	33.92	66.08	11.80	13.08	9.04
1/8000	33.7	66.30	11.01	12.98	9.71
1/10000	33.58	66.42	10.80	10.50	12.28

A medida que se aumenta el factor de peso se observa un aumento significativo del reconocimiento (PRA) junto con una mayor afinidad entre los porcentajes de borrado e inserción. Si se introduce una penalización mayor, las tasas ya no aumentan demasiado y se observa un aumento del borrado. Al aplicar 1/10000 incluso el porcentaje de borrado es mayor que el de inserción.

Tabla 4.6.- Resultados de reconocimiento de palabras aisladas aplicando distintos factores de peso entre palabras (BDGEO).

Factor de peso	%PRE	%PRA	%PRS	%PRI	%PRB
1/1	76.37	23.63	49.02	25.17	1.15
1/num_uni	**62.32**	**37.68**	**44.15**	**12.65**	**5.33**
1/5000	61.66	38.34	43.10	10.88	7.68
1/8000	60.81	39.19	42.68	10.02	8.11
1/10000	59.05	40.95	41.87	8.28	8.90

En la figura 4.3 se presentan algunos resultados de reconocimiento observados para 4 frases elegidas arbitrariamente de la tarea BDGEO (1, 7, 14 y 15). Se observa una fuerte inserción de palabras cortas. A pesar de esa tendencia no es cierto que el sistema siempre prefiera las palabras cortas, aún no penalizando nada las inserciones resultan también algunas palabras largas, siempre y cuando estén de acuerdo con los resultados acústicos. Por ejemplo porte ->porque.

ORIGINAL	como se llama el mar que ba~na valencia
FONÉTICA (DAF)	S koma se ama ma e ke BaJa @alanTl an S
PALABRAS (1)	como se a mar mar que ba~na valencia han
PALABRAS (1/N)	como sea mar mar que ba~na valencia han
PAL(1/10000)	como se alhama que ba~na valencia han
ORIGINAL	donde desemboca el guadiana
FONÉTICA (DAF)	S y an ase mBa ka Galana D S
PALABRAS (1)	y han a san va cabo al han han
PALABRAS (1/N)	y han a san va cabo hallan han
PAL(1/10000)	y han desemboca guadiana
ORIGINAL	por que mar esta ba~nada asturias
FONÉTICA (DAF)	S por te mar asta BaJa stmriars S
PALABRAS (1)	porque mar hasta ba~na las turia ons
PALABRAS (1/N)	porque mar hasta ba~nan asturias
PAL(1/10000)	porque mar hasta ba~nan asturias
ORIGINAL	que comunidad ba~nada por el mediterraneo es la mas extensa
FONÉTICA (DAF)	S ke komo ni a DaJa tor e meitera ni Ge Des a mas escrnsa uB S
PALABRAS (1)	que como ni e a ba~na por a el me y tera ni de dos a mas estan sant
PALABRAS (1/N)	que como ni a ba~na trevenque la ni esa mas estan sant
PAL(1/10000)	que como ni a ba~na trevenque la ni esa mas estan sant

Figura 4.3.- Resultados de reconocimiento para cuatro frases elegidas arbitrariamente de la tarea BDGEO.

Para hacer una comparación real del número de fonemas y de palabras reconocidas sobre la tarea BDGEO se han transcrito (según la transcripción ortográfica) tanto los resultados como las frases de referencia y se han obtenido los resultados presentados en la tabla 4.7.

Tabla 4.7.- Representa el número de palabras y de fonemas que aparecen en las frases de referencia (Transcripción ortográfica) y en las reconocidas aplicando distinto factor de peso entre palabras.

	número_palabras	número_fonemas
referencia	5643	25335
bdgeo (1)	9415	25032
bdgeo (1/1208)	6729	24480
bdgeo (1/10000)	5993	24365

Analizando estos resultados se observa que:

1- Se reconocen más palabras que las que hay en realidad.

2- Se reconocen menos fonemas que los que hay en realidad.

3- El sistema tiende a reflejar los resultados de la etapa DAF y por eso se observan tantas palabras de pequeña longitud incluso penalizando fuertemente la inserción.

Es fundamental la dependencia acústica que se observa. Si no se aplica penalización entre las palabras o la misma es débil el sistema tiende a elegir una configuración de varias palabras cortas que se ajustan mejor a los resultados que se obtienen en la etapa de DAF. Si se aplica una penalización muy fuerte, tiende a ir por palabras algo más largas lo cual es lógico porque se penalizan las inserciones (se hace prohibitiva la transición entre palabras). No obstante, por los resultados observados se deduce que aún le queda mucho por hacer al módulo de ML para

lograr un sistema con bajo porcentaje de error. Cuando los locutores graban las frases puede que no pronuncien correctamente todas las palabras o incluso que no las pronuncien. Como el sistema tiende a recoger la configuración obtenida en la etapa DAF, será el ML el que tenga que imponer las restricciones para llegar a un compromiso entre la aportación de ambas fuentes de conocimiento y llegar a un bajo porcentaje de error.

En la tabla 4.8 se muestra la distribución de las palabras en función de su longitud. Se observa claramente la cantidad de palabras de 1, 2 y 3 fonemas que se insertan aunque se penalice fuertemente la inserción y creemos que es porque no encuentra palabras largas que se ajusten a los resultados obtenidos en la etapa de DAF.

Tabla 4.8.- Representa el número de palabras de longitud X en fonemas que aparecen tanto en las frases reconocidas como en las de referencia.

longitud	referencia	bdgeo (1)	bdgeo (1/208)	bdgeo (1/10000)
1	126	2941	1174	901
2	1551	3125	1436	992
3	1071	1717	1198	980
4	783	931	1073	1015
5	507	549	764	793
6	411	206	314	389
7	153	89	138	179
8	243	100	168	190
9	396	78	194	243
10	243	109	184	202
11	156	20	84	107
12	3	0	2	2

4.2.2.2.- Unidades contextuales: árboles de decisión.

A continuación se presenta la comparación con los resultados obtenidos utilizando las unidades contextuales obtenidas a partir de árboles de decisión para construir los modelos de las palabras a partir de su transcripción. Como se ha señalado anteriormente existe un problema derivado del desconocimiento de la unidad contextual que se debe utilizar en los bordes de las palabras, puesto que el contexto está sin determinar. Una posible solución consiste en utilizar todos los posibles contextos, lo cual implica una gran explosión combinatoria que lleva a un excesivo coste computacional.

Otra posible solución es utilizar las unidades incontextuales en los bordes lo que ha dado lugar al experimento AD_bor_I (Arboles de Decisión y en los bordes unidades Incontextuales) o entrenar unidades especificas para los bordes, que son unidades que tienen en cuenta el contexto derecho al comienzo de las palabras y el contexto izquierdo en los finales, lo que ha dado lugar al experimento AD_bor_CiCd (Arboles de Decisión y en los bordes unidades de Contexto izquierdo y Contexto derecho) (López de Ipiña et al, 1999b). En la tabla 4.9 se presentan los resultados de %PRE obtenidos para las tareas MLA y BDGEO.

Tabla 4.9.- Resultados de %PRE utilizando unidades contextuales.

		1/1	1/num_uni	1/5000	1/8000	1/10000
MLA	Incontextuales	42.88	36.66	33.92	33.70	33.58
	AD_bor_I	35.07	32.02	29.88	29.77	29.71
	AD_bor_CiCd	34.40	31.05	28.14	28.01	27.91
BDGEO	Incontextuales	76.37	62.32	61.66	60.81	59.05
	AD_bor_I	69.82	58.07	57.99	57.87	57.84
	AD_bor_CiCd	69.09	57.48	56.20	55.78	55.58

Parece que la inclusión de unidades contextuales ayuda bastante al reconocimiento frente a la utilización de unidades incontextuales. Un aumento de los resultados acústicos favorece notablemente los resultados a nivel de palabras aisladas. Los resultados obtenidos cuando se entrenan las unidades de los bordes son mejores que los obtenidos usando las unidades incontextuales. En el punto 4.4.3.2 se mostrarán de nuevo resultados haciendo uso de estas unidades contextuales.

4.3.- Decodificación integrando conocimiento acústico y sintáctico

En este punto se van a evaluar los resultados de decodificación obtenidos cuando se integran los modelos acústicos (MOM) junto con los modelos de lenguaje, que serán los modelos *k*-Explorables en Sentido Estricto (*k*-EE). En primer lugar, se analizan las diferentes posibilidades de integración entre ambos tipos de modelos para posteriormente, elegida una configuración, mostrar los resultados obtenidos.

4.3.1.- Integración de los modelos acústicos y de lenguaje

Para desarrollar un sistema de RAH continua es necesario hacer uso de conocimientos acústicos y sintácticos. Sin embargo el desarrollo de tal sistema no es único y depende de tres cuestiones fundamentalmente:

a) implementación secuencial o integrada de las distintas fuentes de conocimiento.

b) organización del léxico en memoria.

c) elección del algoritmo de búsqueda más adecuado y heurísticos para reducir el coste computacional.

A continuación se van a abordar estas tres cuestiones.

4.3.1.1.- Implementación de la integración entre DAF y ML

Fundamentalmente existen dos propuestas alternativas de construcción de un sistema completo de RAH para hacer cooperar las distintas fuentes del conocimiento (acústicas y de ML).

La primera de las propuestas (figura 4.3a)) consiste en integrar en un sólo módulo el conocimiento acústico y léxico (transcripción fonética de las palabras de la aplicación) en lo que se puede llamar decodificación léxica o decodificación de palabras aisladas. Este primer módulo debe proporcionar las N mejores secuencias de palabras que posteriormente pasan a un segundo módulo de análisis sintáctico, donde el ML debe decidir cual de la secuencias es la más probable según el conocimiento sintáctico o superior aportado.

El segundo esquema (figura 4.3b)) corresponde a un modelo completamente integrado en el que todas las fuentes de conocimiento se aportan conjuntamente, para dar finalmente la secuencia de palabras (única) más probable.

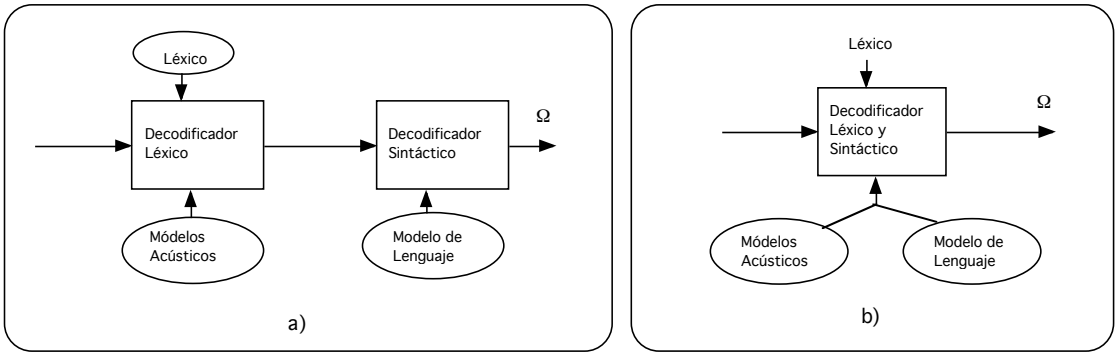


Figura 4.3.- a) Esquema de un sistema de RAH en el que los conocimientos acústicos y léxicos se integran en un primer módulo y el análisis sintáctico en un segundo módulo. b) Esquema de reconocimiento que integra en un único módulo todas las fuentes de conocimiento (acústica, léxica y sintáctica).

En un sistema completamente integrado (figura 4.3b)) todas las fuentes de conocimiento actúan conjuntamente para dar una única decodificación que esté de acuerdo con las restricciones acústicas, léxicas y sintácticas. El sistema trabaja desde el principio con mayor seguridad, ya que no está guiado por un único tipo de conocimiento (acústico). Sin embargo, se dice que los modelos de lenguaje a aplicar no deben ser muy complicados (Jelinek, 1991) ya que ello incrementará notablemente la estructura de búsqueda y por tanto los tiempos necesarios para las decodificaciones de las sentencias.

La ventaja del primer esquema de construcción del sistema (figura 4.3a)) es que la decodificación puramente acústica se separa de la aplicación de conocimientos sintácticos y superiores y por tanto se puede aplicar un ML todo lo complejo que se desee en un paso posterior. Sin embargo, en el esquema secuencial es necesario obtener del decodificador léxico las N mejores secuencias de palabras para que el siguiente módulo clasifique las más probables en función de los conocimientos léxicos, sintácticos, semánticos, etc (Schwartz & Austin, 1991). Puesto que la secuencia de palabras se obtiene sin tener en cuenta los conocimientos superiores, se necesita un valor de N bastante grande para encontrar la decodificación adecuada.

Una alternativa a la obtención de las N-mejores (N-*best*) secuencias de palabras son los grafos de palabras (Ney et al, 1994). La idea de los grafos de palabras consiste en que el decodificador léxico genere redes de palabras de manera que la decodificación correcta (elección tras aplicar el ML) sea uno de los caminos de la red. Es una aproximación mucho más eficiente puesto que sólo es necesario generar hipótesis de palabras localmente mientras que en los métodos de las N-mejores cada alternativa local requiere una sentencia completa que se suma a la lista de N-mejores.

En general se podría decir que este primer esquema de integración secuencial ha sido utilizado tradicionalmente para poder aplicar MLs bastante complejos. Sin embargo, resulta evidente, que si es posible introducir de forma eficiente ML complejos junto con los modelos acústicos se gana en eficiencia (no sería necesario la obtención de tantas decodificaciones que son sintácticamente incorrectas).

4.3.1.2.- Organización del léxico

Fundamentalmente existen dos alternativas de representación del léxico en memoria: lineal o en forma de árbol. En una representación lineal, cada palabra se representa como una secuencia lineal de unidades subléxicas (p. e. fonemas), independientemente de otras palabras. En una representación en forma de árbol cada arco de una estructura tipo árbol representa una unidad subléxica, de tal modo que una secuencia que pase desde el nodo raíz hasta una hoja del árbol representa una secuencia legal de fonemas que pertenece a una palabra del vocabulario. Esta representación puede resultar útil cuando hay muchas palabras que comparten los mismos fonemas iniciales.

En la etapa de decodificación, tanto si se trabaja con léxico representado en forma lineal o de árbol, es necesario hacer varias copias de los modelos de palabras (tantas como estados del ML que representen secuencias de palabras distintas, pero que coinciden en la última palabra de la secuencia).

Por tanto, cuando se trabaja con la representación del léxico en forma de árbol al hacer varias copias de las estructuras en forma de árbol el espacio de búsqueda se incrementa notablemente (Ney et al, 1993). Sin embargo, utilizando heurísticos para reducir el coste computacional y para la lengua alemana se han presentado resultados interesantes utilizando bigramas (Ney, 1993) y trigramas (Ormanns et al, 1998).

Con la representación del léxico en forma de árbol, se presenta una complicación adicional debido a que la identidad de la palabra sólo se conoce al final del árbol y solamente entonces se puede aplicar la probabilidad del ML. De esa manera la probabilidad del ML sólo se puede incorporar después de que se ha llegado a la segunda palabra que compone el bigrama, o a la tercera que compone el trigrama, etc.

Esta representación puede resultar de utilidad en aplicaciones sobre lenguas que tengan muchas palabras que se declinan como, por ejemplo, la lengua alemana (para la que ha sido introducida) o la lengua Vasca (Euskera). Para otras lenguas, puede implicar un coste espacial y temporal asociado excesivo sin que se obtengan beneficios, aún cuando la aplicación sea de vocabularios medianos y grandes. En concreto, para el Castellano y para aplicaciones como la estudiada en esta memoria no parece necesario el uso de esta aproximación.

4.3.1.3.- Algoritmos de búsqueda

La gran cantidad de recursos computacionales que necesitan los sistemas actuales, puede hacer en algunas ocasiones inviable su utilización. Para minimizar el esfuerzo, es necesario utilizar algoritmos que mejoren la eficiencia de la *búsqueda* de la mejor secuencia de palabras que corresponde a la frase pronunciada. Normalmente los algoritmos de búsqueda se suelen dividir en dos grandes grupos:

a) algoritmos cuya función objetivo es la búsqueda de la mejor secuencia de estados: Viterbi, A^*, Búsqueda hacia delante y hacia atrás.

b) algoritmos cuya función objetivo es la búsqueda de la mejor secuencia de modelos: Decodificación a Pila (*Stack decoding*), A^*.

a) Búsqueda de la mejor secuencia de estados

Algoritmo de Viterbi: este algoritmo ya ha sido tratado previamente como parte de la decodificación acústico-fonética. Su característica principal es que se trata de un algoritmo síncrono en el tiempo. Es decir, se procesa completamente el instante de tiempo t, antes de pasar a procesar el instante t+1. En el instante t la probabilidad asociada a cada estado se actualiza a partir del mejor resultado obtenido en el instante t-1 y la secuencia de estados más probable se recupera al final de la búsqueda (Rabiner & Juang, 1989). Desde el punto de vista del modelado estadístico, este criterio es subóptimo.

Algoritmo A^*: corresponde al conjunto de algoritmos cuya función objetivo de la búsqueda es la obtención de la mejor secuencia de estados. Sin embargo, también se puede utilizar cuando la función objetivo de la búsqueda es la mejor secuencia de modelos como se verá en el siguiente punto. El algoritmo A^* utiliza heurísticos. Al contrario que en el algoritmo de Viterbi se extienden únicamente los nodos que tienen mayor probabilidad de pertenecer al camino final, lo que implica que se trata de un algoritmo asíncrono en el tiempo (Paul, 1991). Para estimar la confianza depositada en un estado q se calcula la función:

$$f(q)=g(q)+h(q) \qquad (4.17)$$

La función g(q) representa la probabilidad del mejor camino desde el origen hasta el estado q (por ejemplo la probabilidad obtenida a través del algoritmo de Viterbi). La función heurística h(q) estima la probabilidad del mejor camino desde el estado q hasta el objetivo.

Este método de búsqueda heurística es admisible (es decir, se garantiza el encontrar el camino óptimo), si h(q) es una sobreestimación de la probabilidad exacta del mejor camino desde el estado q hasta el objetivo (condición necesaria). Pero también puede ser subóptimo (se pueden evaluar teorías incorrectas). El problema consiste en proponer un heurístico adecuado (Paul, 1992) y (Cremelie et al, 1995). La idea básica del algoritmo consiste en implementarlo mediante una *pila*, en la que se van guardando las probabilidades ordenadas de los caminos parciales y se van desarrollando los de mayor probabilidad.

Búsqueda hacia delante y hacia atrás (*fordward-backward search*): matemáticamente está basada en el algoritmo de entrenamiento de Baum-Welch conocido como "algoritmo *Fordward-Backward*". Utiliza una búsqueda simplificada hacia delante en tiempo real (*fordward*), seguida de una búsqueda detallada hacia atrás (*backward*), que se puede computar en muy poco tiempo después de que la primera haya acabado (Austin, 1991). La clave de este algoritmo es que a la vez que se lleva a cabo la búsqueda hacia delante, se guardan los valores de las probabilidades correspondientes al último estado de los modelos acústicos activos, que representan el final de una palabra. De este modo, en la búsqueda hacia atrás se eliminan rápidamente aquellos caminos que no tienen ninguna posibilidad de pertenecer al camino óptimo.

b) Búsqueda de la mejor secuencia de modelos

Decodificación a Pila (*stack-decoding*): el algoritmo de Viterbi es un algoritmo óptimo localmente lo cual no quiere decir que se obtenga la sentencia óptima global. Una posible solución a este problema la da el algoritmo de decodificación a pila que es óptimo globalmente y que originalmente ha sido usado en el campo de la transmisión de la información. Este algoritmo lleva a cabo una búsqueda no exhaustiva de izquierda a derecha a través de un árbol que se construye guiado por las probabilidades del ML.

La tarea del algoritmo de búsqueda es examinar de alguna manera ordenada caminos iniciales del árbol y dirigir la búsqueda a través de las extensiones de esos caminos que parezcan más prometedoras. El proceso de búsqueda utiliza una pila y depende de la evaluación de una función L que asigna valores a los caminos parciales (Jelinek, 1969).

1.- Inicialmente la primera entrada de la pila corresponde al nodo raíz del árbol.

2.- Extraer de la parte superior de la pila aquel camino parcial cuyo L-valor sea el más alto. Si el camino parcial es completo (es decir se ha llegado a un nodo terminal del árbol) se para. Si no se calcula la L-evaluación de todas las hojas que pueden ser extensión a partir del camino parcial bajo estudio y se insertan los nuevos caminos extendidos dentro de la pila en orden decreciente del L-valor.

3.- Repetir el paso 2.

Este algoritmo requiere especificar la función L. El problema derivado de calcular funciones heurísticas para podar todas las posibles hipótesis hacen que este algoritmo sea de difícil aplicación en los sistemas de RAH integrados (Bahl et al, 1993).

La clave de este algoritmo es decidir cual es la teoría que debe ser extraída de la cima de la pila para ser extendida. Si se usa como criterio desarrollar los caminos con probabilidades acumuladas máximas, el resultado será una búsqueda uniforme (Jelinek, 1976). La cantidad de cómputo requerida para esta búsqueda resulta prohibitiva y además se necesita una pila muy grande porque las probabilidades acumuladas decrecen muy rápidamente con la longitud del camino. Para evitar este problema lo mejor es el uso del criterio A* (Kenny, 1993).

Algoritmo A:* el algoritmo A* ya ha sido presentado en el punto anterior. En este caso se utiliza para obtener la mejor secuencia de modelos. Como en el caso anterior utiliza heurísticos y se debe evaluar la confianza que se tiene, en que el modelo pertenezca al conjunto de modelos que forman parte de la secuencia final. El criterio A* es la diferencia entre la probabilidad actual de llegar a un punto de un camino en un tiempo concreto t y un límite superior mínimo de la probabilidad de cualquier camino que llegue a ese mismo punto en ese instante. El algoritmo A^* cuando se utiliza para buscar la mejor secuencia de modelos también puede utilizar la maximización de Viterbi, para no realizar el cálculo de las probabilidades a través de todos los caminos.

c) Reducción de la carga computacional

De entre todos los algoritmos anteriores, el más simple y no por ello menos útil es el algoritmo de Viterbi. Se trata de obtener la ruta más probable a través de la red integrada, dada la secuencia de unidades acústicas y/o léxicas. Sin embargo, a pesar de su utilización, el coste del reconocimiento para tareas de vocabularios medios o grandes es prohibitivo, ya que van a estar activos todos los estados de la red integrada.

Para hacer más eficientes tanto el algoritmo de Viterbi como el resto de algoritmos se pueden utilizar heurísticos que reduzcan la carga computacional, como por ejemplo: la búsqueda en haz, la poda a partir del histograma o los algoritmos de búsqueda rápida hacia delante. Con este tipo de técnicas se intenta explorar eficientemente las restricciones impuestas por el ML, y evaluar las ecuaciones sólo en las regiones más relevantes del espacio de búsqueda.

Búsqueda en haz (*Beam_search*): esta técnica se puede considerar como clásica y se utiliza desde los primeros sistemas (Lowerre & Reddy, 1980). Consiste en seguir desarrollando únicamente las hipótesis más prometedoras. La estrategia de poda se basa en que aquellas hipótesis dentro de un camino que tengan una probabilidad significativamente menor que la máxima, tienen muy pocas posibilidades de resultar en el camino óptimo. Por eso, para cada instante de tiempo todos los caminos con probabilidades que no lleguen a un umbral, obtenido a partir del mejor camino observado hasta la fecha, no serán considerados para futuros cálculos. Puesto que es un algoritmo síncrono en el tiempo, es posible comparar las distintas hipótesis en un instante de tiempo t (reconoce la misma secuencia acústica parcial) y dejar únicamente aquellas con mayor probabilidad de pertenecer a la secuencia óptima. En concreto: sea p la probabilidad de la mejor hipótesis hasta el momento y *ah* una constante que regula el ancho de haz entre 0 y 1. Se eliminan aquellas hipótesis cuya probabilidad sea menor que p**ah*. Este procedimiento es muy común y se conoce como algoritmo de Búsqueda en Haz (*Beam-Search*).

ah es el parámetro que regula la anchura del haz de búsqueda, cuanto menor sea *ah* más conservador será el criterio. Si este factor es próximo a 1 el haz de probabilidades que se permite es relativamente pequeño al contrario que si el factor es próximo a cero. La ventaja principal es que el coste de la búsqueda pasa a ser proporcional al número de hipótesis e independiente de la red integrada. El inconveniente principal es la pérdida de optimalidad de la búsqueda (utilizando el algoritmo de Viterbi, la búsqueda de por sí, ya no es óptima). Sin embargo, la experimentación demuestra que eligiendo un ancho de haz suficientemente ancho, la tasa de reconocimiento es muy similar a la obtenida en la búsqueda completa (*full search*). Se puede concluir que existe, para cada tarea y conjunto de test, una anchura de haz mínima a partir de la cual los resultados se degradan bastante. Normalmente para el desarrollo de la búsqueda en haz se utiliza una estructura llamada lista de nodos activos que contiene los nodos que deben ser expandidos en el instante de tiempo t y la lista de nodos activos asociada al índice temporal t se crea a partir de la lista de nodos activos asociada al índice temporal t-1 (Steinbiss et al, 1994).

Una red de hipótesis parciales guarda información asociada a los estados de la gramática, que han resultado activados en el proceso de decodificación para cada frame. Al terminar la

construcción de la lista de estados activos para cada frame, se coloca en la cabeza de la lista el estado más probable, ya que seguramente él dará lugar al estado de mayor probabilidad del frame siguiente haciendo así que la poda en el frame siguiente sea más eficiente.

Por tanto, la probabilidad obtenida en la búsqueda se utiliza tanto para valorar las probabilidades de las distintas hipótesis en los caminos, como para podar aquellas que no superan un umbral. Esta estrategia de poda presenta dos características muy atractivas. La primera es que no es necesario normalizar las probabilidades acumuladas, puesto que el sistema es síncrono con el tiempo. La segunda es que hay un efecto de auto-focalización en la búsqueda de tal forma, que el número de hipótesis se ajusta automáticamente a la ambigüedad de la señal de voz y a la del ML (Ney et al, 1992b).

Poda a partir del histograma (*histogram pruning*). Se trata de un criterio de limitación de la búsqueda que normalmente complementa a la aplicación de la búsqueda en haz. Consiste en fijar el número máximo de nodos que se pueden expandir para cada lista de nodos activa. Para no tener que ordenar los nodos según su probabilidad, se construye un histograma de las probabilidades asociadas a los nodos de la lista de nodos activos (Steinbiss et al, 1994) (Bonafonte, 1995) (Llorents & Casacuberta, 1999).

Búsqueda rápida hacia delante (*Fast Look-Ahead*): dado un vocabulario con un gran número de palabras, el número de alternativas a explorar puede ser muy grande. La solución puede ser descartar aquellas palabras que no sean buenas candidatas mediante algún procedimiento de búsqueda rápida hacia delante (Fast Look-*Ah*ead) (B*ah*l et al, 1989). Simplemente se trata de mirar un poco más allá, a través de unos pocos elementos de la observación, para evitar la activación innecesaria de demasiados modelos.

La propuesta más conocida es el *Fast Match* (B*ah*l et al, 93). El método más usual de implementar el algoritmo *de Fast Match* en un sistema de RAH consiste en utilizar dos reconocedores en equipo. El primero de ellos realiza un examen burdo y actúa como un filtro para podar todas aquellas palabras que no tienen ninguna posibilidad de ser reconocidas por el segundo, que trabaja más detalladamente. Se reemplaza el modelo acústico detallado de cada unidad por un modelo más simple para el cual los cálculos sean mucho más rápidos. Al mismo tiempo debe mantener cierta precisión para no incluir demasiadas palabras en la lista de candidatas. El segundo sistema, síncrono en el tiempo, trabaja de la forma usual (por ejemplo utilizando la decodificación que proporciona el algoritmo de Viterbi).

La principal desventaja que sufre este algoritmo es que es necesaria una etapa previa de ajuste de parámetros (etapa de *tunning*), para llegar a un buen compromiso entre la selectividad (que palabras incluir en la lista) y evitar la introducción de errores irrecuperables (situaciones en las que se elimine de la lista, palabras que realmente son válidas). Naturalmente lo mejor es mirar hacia delante la mayor cantidad posible de señal, para determinar mejor la lista de palabras cuya continuación está permitida. Pero esto implica mayor demora de la búsqueda exhaustiva del segundo sistema. Para alcanzar gran velocidad la búsqueda hacia delante debe ser muy rápida. Esto implica que en algunas ocasiones sólo se elige una palabra por sus dos o tres primeros

fonemas lo cual no es muy efectivo puesto que palabras que comiencen igual, pueden tener significados muy diferentes. Por tanto, aunque este algoritmo reduce la cantidad de esfuerzo del segundo sistema, el segundo tiene que aplicar aún bastante esfuerzo.

4.3.2.- Condiciones experimentales y criterios de evaluación

En este punto se van a presentar las condiciones experimentales del sistema completo de RAH. Al igual que en las secciones precedentes los modelos acústicos son los MOM. En lo que se refiere a la modelización sintáctica se van a utilizar los modelos de orden K que integran K modelos *k*-Explorables en Sentido Estricto (*k*-EE) con k=1...K, presentados en el capítulo 3. Esto nos permite trabajar con autómatas de estados finitos. Los modelos de lenguaje se obtienen a partir de grandes bases de datos de texto (ver apéndice A). Se van a evaluar los modelos con diferentes valores de K. La técnica de suavizado aplicada a los ML es la de back-off sintáctico utilizando el descuento de Witten-Bell como se presento en el capítulo 3. Será en el capítulo 5 donde se estudiarán otras técnicas de suavizado.

Se va a implementar el sistema completamente integrado, en el que se aportan conjuntamente todas las fuentes de conocimiento. Con este tipo de implementación es prohibitivo reservar espacio "a priori" para todos los estados del modelo, lo cual implica reservar espacio para muchas repeticiones de una misma palabra. Es fundamental realizar el reconocimiento sobre una red de búsqueda conducida por los datos. Se parte de una situación inicial en la que hay un único modelo por el que presumiblemente comienza el reconocimiento (silencio inicial) y la red de búsqueda se va construyendo dinámicamente (Ney et al, 1992b).

Cuando se trabaja con el sistema integrado, lo que se trata de reconocer son palabras. Aunque ha habido alguna propuesta alternativa al uso de palabras (Peñagaricano et al, 1999) el uso de estas unidades alternativas no está muy extendido. Las palabras se transcriben automáticamente y los modelos se construyen como concatenación de los modelos de las unidades fonéticas. La representación de las palabras del léxico se ha realizado en forma lineal.

Los modelos de las palabras se integran en el autómata de orden K. Cada uno de los estados del autómata representa una secuencia de como máximo (K-1) palabras del vocabulario y en las transiciones entre los estados se integra el modelo de la unidad léxica correspondiente (ver figura 4.4).

Como algoritmo de decodificación se utiliza el clásico algoritmo de Viterbi (ver figura 4.5), junto con la búsqueda en haz para reducir el coste computacional. Aplicando la técnica de búsqueda en haz no es necesario mantener activos todos los nodos asociados a los modelos acústicos y de lenguaje en una red completa sino que se sólo se desarrollan los necesarios sobre una red conducida en todo momento por los datos. Los experimentos se llevaron a cabo en una estación de trabajo *Silicon Graphics* O_2 con un procesador R10000. Las bases de datos utilizadas son de nuevo MLA y BDGEO.

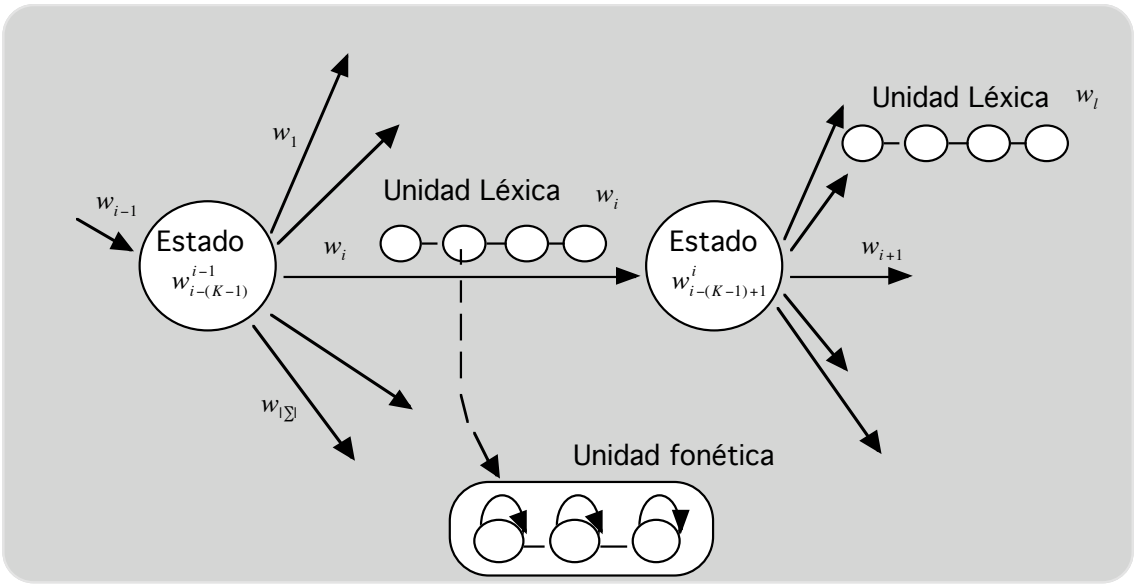


Figura 4.4.- Cada estado del ML representa a una determinada secuencia de palabras de longitud máxima K-1. En las transiciones entre los estados se integran los modelos de las unidades léxicas correspondientes que a su vez están formados por una concatenación de unidades fonéticas representadas mediante MOM.

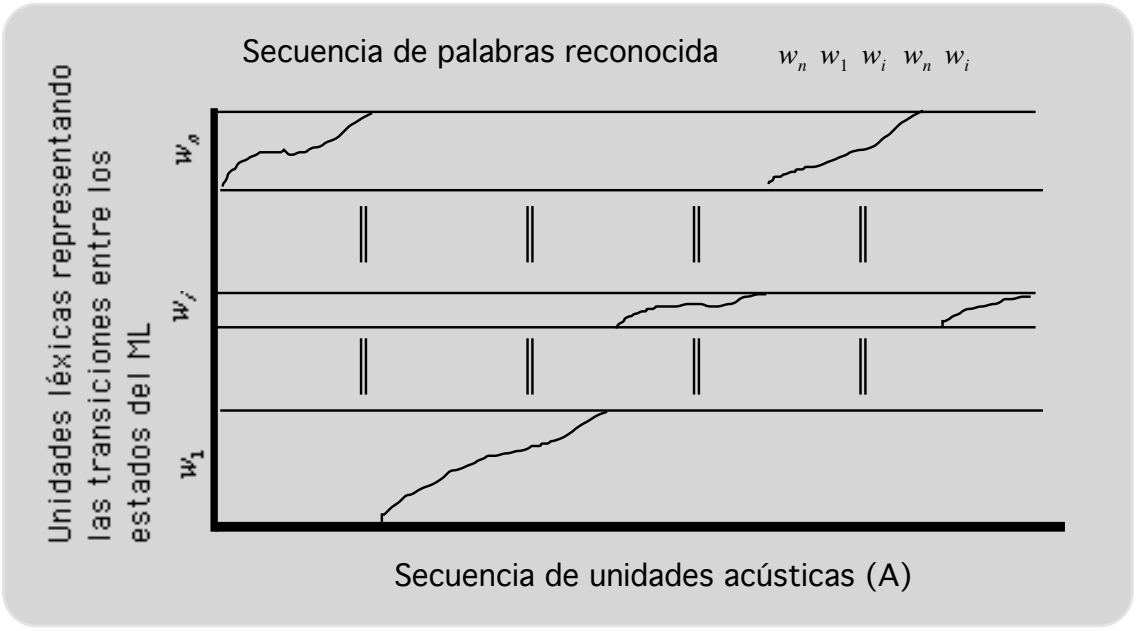


Figura 4.5.- El algoritmo de Viterbi obtiene el camino de máxima probabilidad a través de las unidades léxicas representadas en las transiciones entre estados del ML. Estas unidades son las que se sitúan en el eje de ordenadas de la figura.

La evaluación se realiza igual que para palabras aisladas calculando el Porcentaje Real de Errores:

$$\%\text{PRE}= \frac{w_sustituciones + w_borrados + w_inserciones}{w_aciertos + w_sustituciones + w_borrados + w_inserciones} * 100 \qquad (4.18)$$

Además se da para algunos experimentos el intervalo de confianza $[\text{PRE}^-,\text{PRE}^+]_{95\%}$ de los resultados obtenidos.

4.3.3.- Resultados experimentales con ajuste del parámetro que regula el ancho de haz

En el esquema de decodificación utilizando el algoritmo de Viterbi, cada vez que el sistema considera una posible transición a través de un estado q del ML, se deben evaluar las transiciones desde ese estado q por cada una de las palabras del vocabulario de la aplicación w. Al trabajar con modelos integrados de orden K, se debe utilizar la función de transición $\delta(q/w)$ que maneja la información a través del array propuesto en el capítulo 3. De esta manera se obtiene para cada estado q y para cada palabra w, el estado de destino q_d y la probabilidad asociada $P(w/q)$.

A continuación se presentan los resultados obtenidos considerando distintos factores de ancho de haz (*ah*) para obtener experimentalmente el valor de *ah* que permita abordar las tareas MLA y BDGEO (figura 4.6). Cuando el valor de *ah* es próximo a cero se podan muy pocos caminos, lo cual implica que se obtendrán tasas de error más bajas pero con un coste computacional alto. A medida que se hace próximo a uno se invierten los resultados (mayor tasa de error, menor coste computacional). Los ejes en los que se representan los resultados muestran por una parte, en el eje de ordenadas y en escala logarítmica, el número medio de nodos activos por frame (tanto estados de los modelos acústicos como del ML), que han sido necesarios durante el desarrollo de la decodificación. Este dato es proporcional al tiempo medio necesario para decodificar cada sentencia. En el eje de abcisas se presenta el Porcentaje Real de Error en palabras (%PRE).

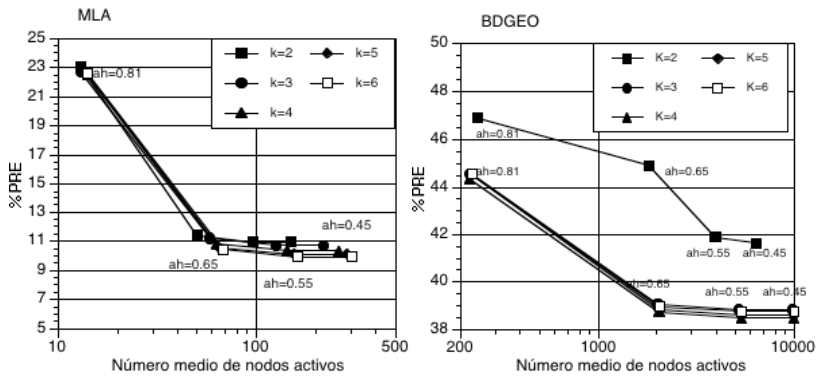


Figura 4.6.- Resultados obtenidos evaluando el porcentaje real de errores (%PRE) para las dos tareas (MLA y BDGEO), diferentes modelos integrados de orden K (K=2, 3, 4, 5 y 6) y factores de ancho de haz (*ah*).

Con la tarea MLA se observa un comportamiento muy similar para todos los valores de K. Con la tarea BDGEO se observa un comportamiento diferente para K=2, siendo el porcentaje de error significativamente más alto.

En todos los casos con un ancho de haz *ah*=0.81 (muy restrictivo) el reconocimiento se degrada bastante. Para valores del parámetro *ah* menos restrictivos (valores de *ah* más pequeños) el resultado que se observa prácticamente es similar en %PRE. Sin embargo, a medida que el parámetro *ah* se hace más pequeño es necesario mantener un mayor número

medio de nodos activos por frame para decodificar las sentencias. El número de nodos que deben estar activos por frame, está directamente relacionado con el tiempo medio necesario para decodificar cada sentencia. Como el tiempo representado en cada frame es de 10 milisegundos, si el tiempo medio por frame es menor a 10 milisegundos se consideran decodificaciones obtenidas en tiempo real. A continuación (tabla 4.10 y tabla 4.11) se muestran los resultados obtenidos en número medio de nodos activos, tiempo medio necesario para decodificar cada frame y %PRE.

Tabla 4.10.- Resultados obtenidos para la tarea MLA, número medio de nodos activos, tiempo medio por frame y %PRE para distintos valores de K.

K	Número medio de nodos activos				Tiempo medio por frame (ms)				%PRE			
	ah=0.81	*ah*=0.65	*ah*=0.55	*ah*=0.45	*ah*=0.81	*ah*=0.65	*ah*=0.55	*ah*=0.45	*ah*=0.8	*ah*=0.65	*ah*=0.55	*ah*=0.45
2	13	50	96	120	1	3	**6**	**9**	23.07	11.40	10.95	10.95
3	13	58	126	183	1	4	**8**	11	22.73	11.20	10.71	10.71
4	14	63	143	212	1	4	**9**	12	22.51	10.79	10.38	10.38
5	14	65	155	231	2	5	**10**	14	22.57	10.59	10.12	10.12
6	14	68	162	299	2	5	11	15	22.60	10.43	**9.93**	**9.93**

La tarea MLA se decodifica en tiempo real con porcentajes de error alrededor del 10%. Los resultados para la tarea BDGEO no son en tiempo real en la mayoría de las combinaciones como se puede ver en la tabla 4.11.

Tabla 4.11.- Resultados obtenidos para la tarea BDGEO: número medio de nodos activos, tiempo medio por frame y %PRE para distintos valores de K.

K	Número medio de nodos activos				Tiempo medio por frame (ms)				%PRE			
	ah=0.81	*ah*=0.65	*ah*=0.55	*ah*=0.45	*ah*=0.81	*ah*=0.65	*ah*=0.55	*ah*=0.45	*ah*=0.8	*ah*=0.65	*ah*=0.55	*ah*=0.45
2	243	1224	3964	6358	6.3	33	85	174	46.91	41.89	41.62	41.62
3	222	2027	5189	9737	6.5	35.3	114	288	44.58	39.09	38.85	38.85
4	221	2052	5374	10050	6.5	38.6	126	392	44.36	38.75	**38.5**	**38.5**
5	222	2057	5401	10065	6.8	39.4	129	298	44.54	38.87	38.63	38.63
6	222	2060	5410	10068	7.3	39.2	123	301	44.57	38.99	38.77	38.77

Se observa que para la tarea BDGEO sólo se obtienen decodificaciones en tiempo real para un valor de *ah*=0.81, obteniéndose un porcentaje de error bastante alto, alrededor del 45%. Para *ah*=0.65 se reconoce en 3.8 veces tiempo real con un porcentaje de error nunca inferior al 38% y para valores de *ah* más permisivos los tiempos son ya muy altos sin que se logre un decremento del porcentaje de error.

En la figura 4.7 se pueden ver las salidas que se obtienen del decodificador de DAF, palabras aisladas con un factor de peso 1/1000 y K=4 con *ah*=0.65 para las cuatro frases elegidas arbitrariamente de la tarea BDGEO ya presentadas anteriormente.

ORIGINAL	como se llama **el** mar que ba~na valencia
FONÉTICA (DAF)	S koma se ama ma e ke BaJa @alanTl an S
PAL (1/10000)	como se alhama que ba~na valencia han
DAF+ML	como se llama mar que ba~na valencia
ORIGINAL	donde desemboca el guadiana
FONÉTICA (DAF)	S y an ase mBa ka Galana DS
PAL (1/10000)	y han a san va cabo hallan han
DAF+ML	y han **desemboca guadiana**
ORIGINAL	por que mar esta ba~nada asturias
FONÉTICA (DAF)	S por te mar asta BaJa stmriars S
PAL (1/10000)	porque mar hasta ba~nan asturias
DAF+ML	por que mar **esta** ba~na asturias
ORIGINAL	que comunidad ba~nada por el mediterraneo es la mas extensa
FONÉTICA (DAF)	S ke komo ni a DaJa tor e meitera ni Ge Des a mas escrnsa uB S
PAL (1/10000)	que como ni a ba~na trevenque la ni esa mas estan sant
DAF+ML	**que como ni a** ba~na por el maitena ni de dos la mas extensa un

Figura 4.7.- Resultados de decodificación obtenidos para cuatro frases elegidas arbitrariamente de la tarea BDGEO. Los resultados muestran la decodificación fonética, los resultados de palabras aisladas obtenidos cuando se aplica un factor entre palabras 1/10000 y los obtenidos con DAF+ML (K=4, *ah*=0.5).

Con la aplicación del ML se logra una mejor salida del decodificador: en la primera frase el ML resuelve bastante bien la situación que acústicamente ya estaba próxima a la correcta. Le falta introducir, la palabra "el" que no aparece dentro de la secuencia acústica y que puede que el hablante ni siquiera haya pronunciado.

En la segunda frase ha sido capaz de introducir las palabras "desemboca" y "guadiana" pero no ha podido arreglar el comienzo de la frase y se sigue reconocimiento " y han".

En la tercera frase el ML ha sido capaz de sustituir "hasta", asociado a los resultados acústicos obtenidos, por "ésta".

Pero sigue habiendo situaciones muy persistentes asociadas a las componentes acústicas. Por ejemplo en la cuarta frase "que comunidad" se sigue reconociendo como "que como ni a". Por lo que parece que la ayuda que proporciona el ML no es todo lo buena que cabría esperar.

4.4.- Evaluación de las ponderaciones de las probabilidades acústicas y/o de ML

Una vez integrada la aproximación gramatical (integración de gramáticas *k*-Explorables en Sentido Estricto con los modelos acústicos en un sistema de RAH), se estudiarán los efectos derivados de las ponderaciones de las probabilidades tanto sobre los modelos acústicos como sobre los ML necesarias para el óptimo funcionamiento del sistema. El objetivo es buscar una óptima relación entre ambas distribuciones de probabilidad. Finalmente se presentarán las conclusiones obtenidas.

En RAH se trata de conseguir la secuencia de palabras Ω que con mayor probabilidad corresponde a la observación acústica A. La probabilidad final es el producto del conocimiento aportado por ambas fuentes (acústica (P(A/Ω)) y sintáctico-semántica (P(Ω))). Normalmente los modelos acústicos y de lenguaje se construyen por separado. Los modelos acústicos se obtienen a partir de bases de datos acústicas que no tienen en cuenta las grandes bases de datos de texto utilizadas para obtener los modelos de lenguaje, mientras que los ML se obtienen a partir de gran cantidad de texto sin tener en cuenta la posible configuración acústica. Como cada operando se obtiene a través de dos caminos muy distintos el producto final no es un probabilidad *real*. Por otra parte, ambas probabilidades no son reales (Jelinek, 1990), sino estimaciones.

En el *EUROSPEECH* del 95 (Bourlard, 1995) y después en el "*Speech Comunication*" (Bourlard et al, 1996) se hace un recorrido sobre las grandes lagunas que aún afectan a las diferentes etapas de un sistema de reconocimiento automático del habla. Como un punto destacado se señala este mismo problema. Se dice que es necesaria la modificación de la contribución acústica o la del ML, lo cual no es una solución muy ortodoxa desde el punto de vista de la teoría estadística. Los motivos que habitualmente se utilizan para justificar estas ponderaciones radican en la diferencia en el rango dinámico de las probabilidades asociadas a cada fuente, diferencia que suele deberse a dos razones. En primer lugar, porque las fuentes están definidas sobre alfabetos de tamaño distinto. De hecho, las probabilidades de observación de los elementos se la secuencia acústica se pueden estimar mediante funciones de distribución de probabilidades continuas, que equivale a utilizar un alfabeto infinito. En segundo lugar, porque el número de probabilidades que utiliza cada fuente para evaluar una frase es muy distinto. Por ejemplo, el número de palabras en una frase es mucho menor que el número de observaciones acústicas. Por tanto, el número de probabilidades que aplica la fuente del ML es mucho menor que el que aplica la fuente de información acústico-fonética. Este mismo fenómeno se manifiesta en el impacto insignificante que tienen las probabilidades de transición de los modelos acústicos (Ney, 1993) (Bourlard et al, 1996).

Los resultados experimentales muestran que los mejores resultados de reconocimiento se obtienen cuando se introduce una modificación en la contribución de una o ambas fuentes de conocimiento mediante factores γ y α, aplicados a las probabilidades de la ecuación 1.3 de la siguiente forma:

$$P(A/\Omega)^{\gamma}P(\Omega)^{\alpha} \qquad (4.19)$$

Lo más usual, representado en la literatura, es mantener el parámetro $\gamma=1$ e introducir un $\alpha>1$. La constante α normalmente se obtiene experimentalmente mediante un proceso de prueba y error, de tal manera que se minimiza el error introducido en el reconocimiento.

En la literatura aparece reflejado el uso del parámetro α en bastantes artículos (Bonafonte et al, 1995) (Rubio et al, 1997), (Ogawa et al, 1998) (Müller & Psutka, 1999), sin embargo no hay por el momento una explicación teórica suficientemente satisfactoria para este hecho puesto que la regla de Bayes predice que este valor de α debe ser también igual a 1 y que esa debiera ser la solución correcta.

En (Rubio et al, 1997) se dice que este factor de compensación es necesario debido a la mala estimación de los modelos acústicos de manera que las decisiones son debidas fundamentalmente a la gramática con poca influencia de las componentes acústicas. Valores de α superiores a 1 podrían disminuir la importancia de la gramática. A lo largo de la presente memoria se verá, que efectivamente el valor de α disminuye las probabilidades de la gramática, pero no para limitar su importancia, sino todo lo contrario.

En (Huang et al, 1993) se presentó una aproximación estocástica que permite ajustar el peso del ML frente a los modelos acústicos utilizando un método de descenso por gradiente. Dice que se podría no solo ajustar las probabilidades del ML para complementar ambos factores sino incluso los parámetros de los modelos acústicos. Se aplica a un modelo secuencial donde se obtienen las N-mejores sentencias sin aplicación del ML. No se obtienen unos resultados excesivamente buenos que se argumentan debido a la escasez de datos.

Otro punto importante a tratar es que en decodificación no se puede controlar la longitud de las frases (Rabiner & Juang, 1993). De esta manera se pueden generar sentencias que son arbitrariamente largas por inserción de una gran cantidad de palabras de longitud corta. Una forma de prevenir este efecto es incorporar una penalización a la inserción de palabras dentro de la decodificación por Viterbi, multiplicando por un valor fijo positivo inferior a la unidad, a la probabilidad obtenida al final de cada modelo de palabra $\beta P(\Omega)$. Ajustando esa penalización de las palabras se puede controlar el porcentaje de inserción y de borrado. Una penalización muy alta reduce la inserción de palabras e incrementa el porcentaje de borrados, y un valor pequeño de penalización tendrá el efecto contrario. Normalmente también este valor se determina experimentalmente.

Puesto que todos los valores de las probabilidades de $P(\Omega)$ están comprendidas entre 0 y 1, lo que hace la constante α es disminuir de forma exponencial todas las probabilidades de transición del ML. A medida que aumenta el valor de la constante se hace más notable la

diferencia entre las transiciones de probabilidad alta (que disminuyen menos) y las de probabilidad baja (que disminuyen más). Sin embargo, la constante β disminuye las probabilidades linealmente como puede verse en la figura 4.8.

Cuando se aplica un parámetro γ sobre las probabilidades $P(A/\Omega)^{\gamma}$, el valor de γ debe ser inferior a la unidad de manera que se incrementen exponencialmente las probabilidades acústicas. La distribución de probabilidad tenderá a ser menos distribuida (los valores de las probabilidades se concentrarán en un rango menor) a medida que γ decrezca, ya que las transiciones de probabilidad bajas verán incrementado cada vez más su valor, en contra de lo que les ocurre a las transiciones de probabilidad alta que incrementan sus probabilidades de forma poco significativa.

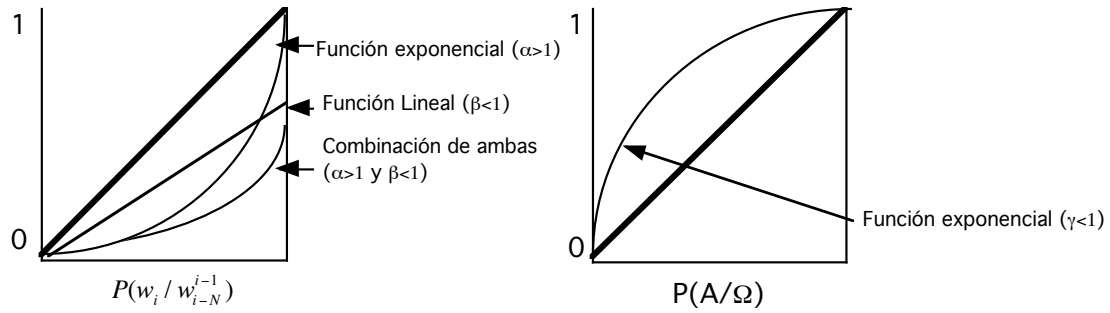


Figura 4.8.- a) Reducción de las probabilidades del ML debido a los factores α y β. b) Incremento de las probabilidades acústicas debido al parámetro γ.

A continuación se van a presentar los resultados que se obtienen cuando se realizan diversas alteraciones sobre el modelo de lenguaje (función exponencial, lineal, escalón y combinación de las mismas), y sobre los modelos acústicos (función exponencial y reducción del número de componentes acústicas).

4.4.1.- Efecto de las ponderaciones sobre las probabilidades del ML.

En los experimentos previos tanto las probabilidades acústicas como de ML se han mezclado directamente utilizando la regla de Bayes ($\alpha=\beta=\gamma=1$). Los resultados obtenidos ponen de manifiesto la gran influencia de las componentes acústicas frente a las del ML. Sabemos que las probabilidades acústicas son muy pequeñas frente a las probabilidades del ML y además se aplican un número mucho mayor de veces.

Un análisis realizado sobre las probabilidades finales (máximas) con las que las frases son decodificadas permite observar que la probabilidad del ML nunca aporta más del 1.5% al conjunto de la probabilidad final máxima, lo cuál pone de manifiesto que las probabilidades del ML no aportan demasiado a la probabilidad final.

Si se decrementan las probabilidades del ML, por ejemplo, elevándolas a una determinada potencia positiva, es posible que la aportación sea mayor, o al menos que esa aportación mínima tenga una mayor influencia en la decodificación obtenida.

Se van a estudiar las siguientes funciones de ponderación (disminución) de las probabilidades del ML:

a)- Función exponencial, $P(ML)^{\alpha}$ con $\alpha>1$.

b)- Función lineal, $\beta P(ML)$, $\beta>1$: esta función se utiliza para limitar la inserción.

c)- Función exponencial y lineal, $\beta P(ML)^{\alpha}$:

Con la función α se disminuyen todas las probabilidades de forma exponencial. En esta sección se va a probar también el efecto que se produce al eliminar de forma drástica las probabilidades más bajas mediante una función escalón.

d) Función escalón:

$$P(ML) = \begin{cases} P(ML) & P(ML) > e \\ 0 & P(ML) \le e \end{cases} \tag{4.20}$$

(Se eliminan todas las transiciones con probabilidades por debajo de un umbral e).

4.4.1.1.- Escalado exponencial: $P(\Omega)^{\alpha}$

En este punto se va a evaluar el efecto de la aplicación de un factor exponencial α a las probabilidades del ML junto con la aplicación de distintos factores de ancho de haz (*ah*) de búsqueda en haz. En la figura 4.9 se puede ver como varía el porcentaje real de error (%PRE) y el número medio de nodos activos implicados (escala logarítmicos) cuando se integran los modelos de orden K (K modelos *k*-EE con K=2, 4 y 6 (óptimo)) de la tarea MLA en el sistema de RAH. Se presenta un barrido del parámetro α, desde α=1 hasta α=7 para distintos valores del parámetro *ah*.

Los puntos representados el margen inferior izquierdo de cada figura son los que representan la mayor eficacia del sistema, es decir el %PRE más bajo junto con un menor número medio de nodos activos, lo cual es directamente proporcional al tiempo necesario para decodificar cada sentencia.

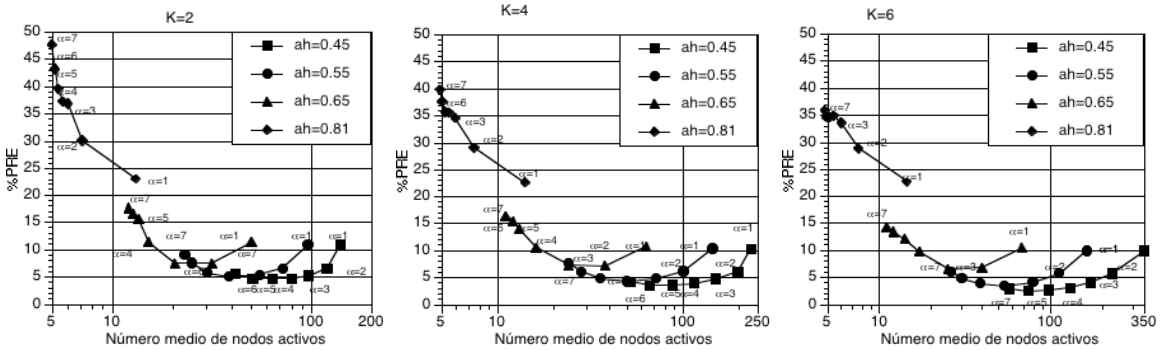


Figura 4.9.- Evolución del error con distintos valores de *ah* y de α para los modelos integrados de orden K con K=2, 4 y 6 (óptimo). Tarea de aplicación MLA.

A medida que aumenta el valor del parámetro α se observa un descenso significativo del número medio de nodos activos. Además, el %PRE disminuye, excepto cuando *ah*=0.81, hasta

alcanzar un valor óptimo a partir del cual vuelve a ascender. Estos mismos resultados junto con el tiempo medio para decodificar cada frame se pueden ver en la tabla 4.12.

Tabla 4.12.- Resultados del número medio de nodos activos, del tiempo medio por frame y del %PRE con distintos factores de ancho de haz (*ah*) para la tarea MLA.

K	Número medio de nodos activos			Tiempo medio por frame (ms)			%PRE		
	ah= 0.6 α= 3	*ah*= 0.5 α= 4	*ah*= 0.4 α= 5	*ah*= 0.6 α= 3	*ah*= 0.5 α= 4	*ah*= 0.4 α= 5	*ah*= 0.6 α= 3	*ah*= 0.5 α= 4	*ah*= 0.4 α= 5
2	21	39	64	1	3	4	7,53	5.01	4.60
3	26	47	80	2	3	5	7.48	4.83	4.36
4	28	50	87	2	4	6	7.17	4.33	3.62
5	29	52	92	2	4	7	6.76	3.64	3.26
6	30	54	98	3	5	7	6.82	3.51	2.68

Cuando el valor del parámetro *ah* es muy restrictivo *ah*=0.81, utilizar un valor de α>1 hace que el sistema produzca peores resultados en todos los casos (se desestiman un número excesivo de hipótesis). Sin embargo, para *ah*=0.65 se produce una mejora significativa hasta α=3. Con *ah*=0.55 y *ah*=0.45 se logran los mejores %PRE para α=4 y α=5 respectivamente. El mejor comportamiento del sistema (menor %PRE y NMNA) se logra para *ah*=0.55. Se llega a una tasa de error inferior al 3% para K=6 y ancho de haz suficientemente grande y por supuesto en tiempo real. En esta tarea a medida que se aumenta el valor de K, disminuye el error lo cual está de acuerdo con los resultados obtenidos en perplejidad (ver figura 3.12).

A continuación (figura 4.10 y tabla 4.13) se presentan los mismos resultados obtenidos para la tarea BDGEO en las mismas condiciones experimentales.

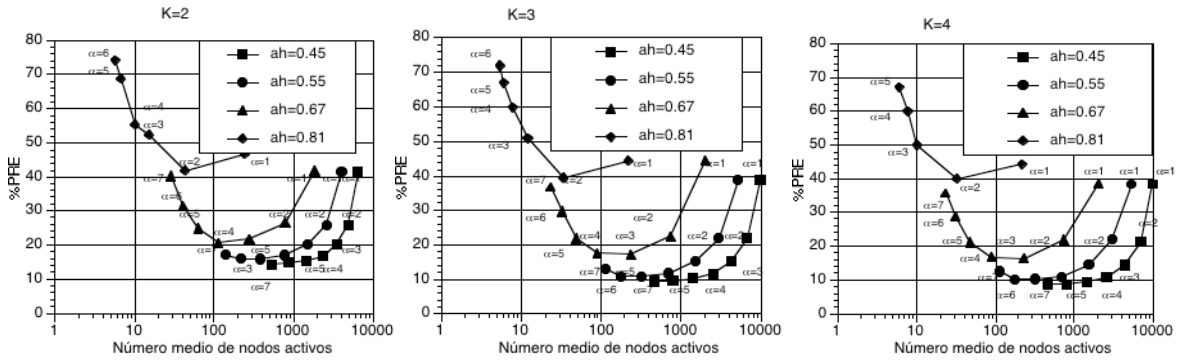


Figura 4.10.- Evolución del error con distintos valores de *ah* y de α para los modelos integrados de orden K con K=2, 3 y 4 (óptimo). Tarea de aplicación BDGEO.

En estos experimentos ocurre algo similar y se observan prácticamente los mismos efectos. Cuando *ah*=0.81, α=2 produce mejores resultados, pero con α>2 el sistema incrementa mucho el %PRE. Para valores de *ah* menos restrictivos se decrementan notablemente las tasas de error y se llega al óptimo con α=4, α=6 y α=7 para *ah*=0.65, *ah*=0.55 y *ah*=0.45 respectivamente. De nuevo el mejor comportamiento se observa para *ah*=0.55. En este caso las tasas de error más bajas se obtienen para K=4: %PRE igual a 8.58 en 2.6 tiempo real (ver tabla 4.13). En tiempo real se obtienen resultados de error cercanos al 10%.

Tabla 4.13.- Resultados del número medio de nodos activos, del tiempo medio por frame y del %PRE con distintos factores de ancho de haz (*ah*) para la tarea BDGEO.

K	Número medio de nodos activos			Tiempo medio por frame (ms)			%PRE		
	ah= 0.6 α= 4	*ah*= 0.5 α= 6	*ah*= 0.4 α= 7	*ah*= 0.6 α= 4	*ah*= 0.5 α= 6	*ah*= 0.4 α= 7	*ah*= 0.6 α= 4	*ah*= 0.5 α= 6	*ah*= 0.4 α= 7
2	114	218.21	526.28	6.6	11.7	25.1	20.54	15.95	14.29
3	90	179.01	467.66	6.0	10.9	25.5	17.51	10.85	9.45
4	89	177.99	469.60	6.2	11.2	26.3	16.71	10.12	**8.58**
5	90	180.57	478.33	6.5	11.6	27.0	16.73	10.25	8.72
6	91	182.32	490.50	6.8	12.0	27.1	16.51	10.66	9.07

En ambos conjuntos de experimentos a medida que se decrementa el valor del parámetro *ah* se necesita un valor de α mayor para obtener los resultados óptimos. En general, para un valor fijo de *ah*, cuanto mayor es el valor de α, la búsqueda se hace mucho más restrictiva (se activan un menor número de hipótesis, como se explica en detalle al final de esta subsección). Cuanto mayor es el valor del parámetro que regula el ancho de haz, el valor de α para el que se consigue el óptimo de %PRE se va incrementando.

Se aprecia una disminución muy notable del error tras aplicar un valor de α=2 y sigue aumentando con α hasta llegar a un valor óptimo a partir del cual comienza a aumentar el error. Con un factor α demasiado alto, las tasas de error aumentan porque se resta demasiado importancia a la aportación hecha por los modelos acústicos. A continuación, en la figura 4.11 se muestra la información más en detalle asociada a la aplicación de un factor *ah*=0.55 que es el que parece llegar a un buen compromiso entre las tasas de error, que no disminuyen más aunque se decremente el valor de *ah* y el número medio de nodos activos implicados (que sigue incrementándose cuando *ah* decrece).

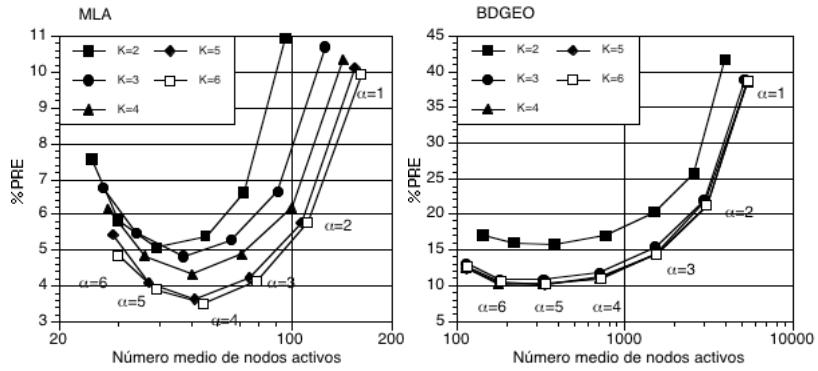


Figura 4.11.- Representa la evolución con α (1...7) del número medio de nodos activos (base logarítmica) para cada uno de los modelos integrados de orden K (K=2, 3, 4, 5 y 6) y el porcentaje de error observado (%PRE). Tareas MLA y BDGEO. Ambas figuras muestran los resultados para un factor *ah*=0.5, que será el elegido para los próximos experimentos de esta memoria.

Se observan notables diferencias entre ambas figuras, cuando la tarea es muy sencilla (MLA, 29 palabras) el incremento de K ayuda notablemente a disminuir el error. Es presumible que habrá cadenas de longitudes altas que se habrán visto una gran cantidad de veces y por tanto están bien entrenadas.

Para la tarea de talla media (BDGEO, 1208 palabras) hay un descenso notable del porcentaje de error al pasar de K=2 a valores de K superiores. Con K>2 no hay diferencias apreciables. El mejor resultado se da para K=4 y no se produce un incremento del error para valores de K superiores, como se podía apreciar con más detalle en la tabla 4.13.

Para la tarea MLA la disminución del error es muy significativa, ya que se partía de tasas de error muy bajas (se pasa de tasas cercanas al 10% a tasas del 3.5%). Para la tarea BDGEO el descenso del error es también importante. Se logra decrementar el error desde aproximadamente un 40% a un 10% independientemente del valor de K. En ambos casos los modelos acústicos son discretos de 4MC y representan unidades fonéticas incontextuales.

En resumen, el aumento de α significa un descenso muy notable en las tasas de error hasta alcanzar un valor mínimo de error o máximo de reconocimiento y también implica una notable disminución en el número medio de nodos activos que se tienen que activarse en el desarrollo del trellis de Viterbi.

En la figura 4.12 se pueden apreciar las salidas que se obtienen de decodificación, etapa DAF, palabras aisladas y K=4 con *ah*=0.55 y distintos valores del parámetro α. Se puede observar como el incremento de α ayuda notablemente al decremento del error hasta llegar a un óptimo a partir del cual se degrada otra vez. En la cuarta frase se puede apreciar que con α=6 se ha solucionado el error sintáctico "que como ni a" por "que comunidad". En la segunda frase se ha sustituido "y han" (acústicamente correcto) por "por donde" en la segunda frase. Con valores muy altos de α, se pueden llegar a conseguir frases sintácticamente correctas que poco tienen que ver con la secuencia acústica de entrada, como por ejemplo, lo que ocurre en la tercera frase que debiera ser "por que mar está bañada asturias" y se reconoce "en que mar están las canarias".

Tras observar realmente los números que se obtienen en los nodos de la red integrada utilizando el algoritmo de Viterbi se ha visto que el problema asociado a la integración de las probabilidades acústicas y de ML no es debido a que las probabilidades del trellis se vayan haciendo cada vez más pequeñas en comparación con las del ML. El problema surge debido a que la amplitud existente entre la probabilidad acumulada del camino de máxima probabilidad y la probabilidad de los siguientes caminos más probables es mucho mayor que el rango de dispersión de las probabilidades del ML, de forma que las probabilidades del ML no influyen la mayoría de los veces en las decisiones sobre el camino que se debe tomar.

ORIGINAL	como se llama el mar que ba~na valencia
FONÉTICA (DAF)	S koma se ama ma e ke BaJa @alanTl an S
DAF+ML (α=1)	como se llama mar que ba~na valencia
DAF+ML (α=2)	como se llama mar que ba~na valencia
DAF+ML (α=4)	como se llama el mar que ba~na a valencia
DAF+ML (α=6)	como se llama el mar que ba~na a valencia
DAF+ML (α=8)	como se llama el mar que ba~na a valencia
ORIGINAL	donde desemboca el guadiana
FONÉTICA (DAF)	S y an ase mBa ka Galana D S
DAF+ML (α=1)	y han desemboca guadiana
DAF+ML (α=2)	me han desemboca guadiana
DAF+ML (α=4)	donde desemboca el guadiana
DAF+ML (α=6)	donde desemboca el guadiana
DAF+ML (α=8)	donde desemboca el guadiana
ORIGINAL	por que mar esta ba~nada asturias
FONÉTICA (DAF)	S por te mar asta BaJa stmriars S
DAF+ML (α=1)	por que mar esta ba~na asturias
DAF+ML (α=2)	por que mar esta ba~na asturias
DAF+ML (α=4)	por que mar esta ba~nada asturias
DAF+ML (α=6)	por que mar esta ba~na las playas
DAF+ML (α=8)	**en que mar estan las canarias**
ORIGINAL	que comunidad ba~nada por el mediterraneo es la mas extensa
FONÉTICA (DAF)	S ke komo ni a DaJa tor e meitera ni Ge Des a mas escrnsa uB S
DAF+ML (α=1)	que como ni a ba~na por el maitena ni de dos la mas extensa un
DAF+ML (α=2)	que como ni a ba~na por el mediterraneo es la mas extensa
DAF+ML (α=4)	que comunidad a la por el mediterraneo es la mas extensa
DAF+ML (α=6)	que comunidad a la que el mediterraneo es la mas extensa
DAF+ML (α=8)	que comunidades autonomas atraviesa mas extensa

Figura 4.12.- Resultados de reconocimiento para cuatro frases elegidas arbitrariamente de la tarea BDGEO con un modelo K-EE con K=4 y *ah*=0.55.

En la figura 4.13 se muestra este hecho representando 3 palabras distintas en un instante de tiempo concreto del trellis de Viterbi. Al final de cada palabra hay una probabilidad acumulada. Las diferencias entre las probabilidades acumuladas son evidentemente más grandes que las diferencias existentes entre las probabilidades del ML debido fundamentalmente a las probabilidades acústicas (que se aplican un número mucho mayor de veces en el trellis de Viterbi y además sus valores de probabilidad son muy pequeños).

La consecuencia inmediata es que los valores de las probabilidades del ML son irrelevantes en la mayor parte de las situaciones para elegir el mejor camino a seguir. Por ejemplo si P_A=5220 , P_B=5210 y P_C=3160 en el instante i (se trabaja con una transformación de los logaritmos de las probabilidades a enteros positivos, por lo que los números más pequeños son los que indican una mayor probabilidad) lógicamente el camino más probable es el que termina en la palabra C, aunque $P(W_A/W_C)$=50 (transición menos probable), $P(W_A/W_A)$=10 y $P(W_A/W_B)$=2 (transición más probable) el sistema elegirá el camino que termina en W_C ya que $P_A+P(W_A/W_A)$= 5220+10, $P_B+P(W_A/W_B)$=5210+2 y $P_C+P(W_A/W_C)$=3160+50. Sin embargo,

si se aplica un α=2, las probabilidades del ML resultan ser más pequeñas (disminución exponencial) y se elegirá otro camino ya que P_A+P(W_A/W_A)= 5220+100, P_B+P(W_A/W_B)= 5210+4 y P_C+P(W_A/W_C)= 3160+2500. En el camino que termina en la palabra W_B se ha llegado a un compromiso entre la importancia de las componentes acústicas y del ML. Por supuesto, si α es muy grande las componentes acústicas dejarían de tener importancia como se ve en el ejemplo de la figura 4.13.

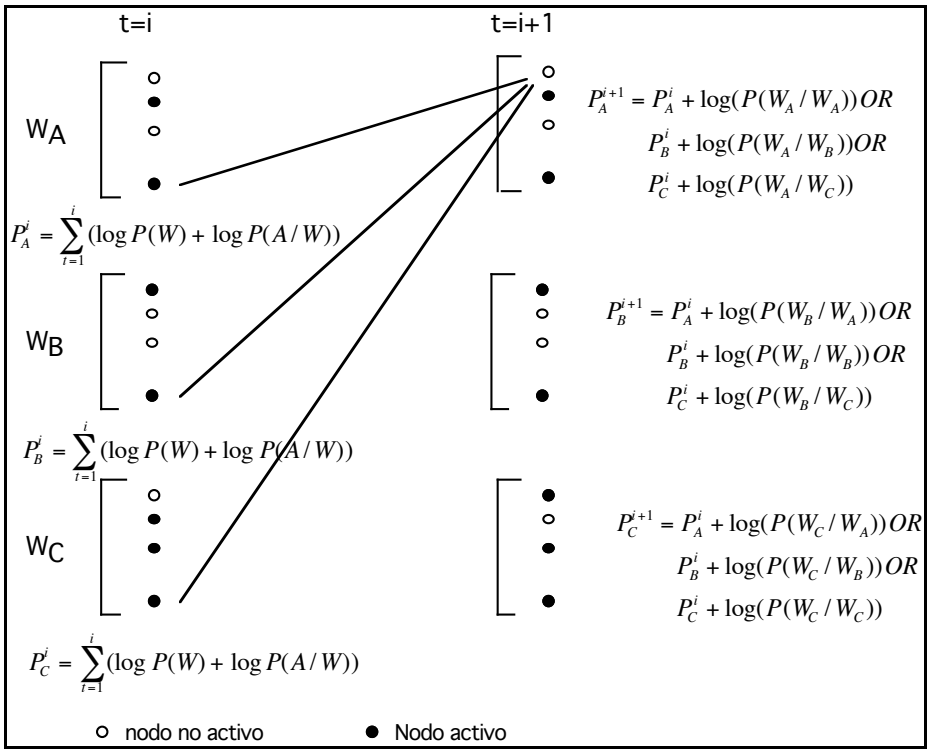


Figura 4.13.- Representa las probabilidades acumuladas en el desarrollo del trellis de Viterbi. Si la diferencia de probabilidad existente entre P_A, P_B y P_C es suficientemente grande y realmente lo es (debido a lo pequeñas que son las probabilidades acústicas) las probabilidades del ML son irrelevantes en la mayor parte de los casos para elegir el camino a seguir.

En resumen se podría decir que, el parámetro α debe dar solución al diferente rango de dispersión de las probabilidades acumuladas a lo largo de los nodos del trellis de Viterbi y las probabilidades del ML. Por tanto, este parámetro va a depender de las distribuciones de probabilidades acústicas y del lenguaje iniciales, así como del parámetro que regula el ancho de haz. No es posible encontrar su valor de forma teórica, ya que se trata de un valor heurístico que depende de la tarea y de las relaciones entre las condiciones experimentales de integración. Por ejemplo, si en vez de utilizar modelos acústicos con 4 codebooks (4 probabilidades distintas que se acumulan en cada frame) se utilizaran modelos acústicos menos precisos con 2 codebooks, es de suponer, que la probabilidad acumulada sería menor, por lo que se necesitaría un valor de α menor para lograr el óptimo (aunque este valor también dependería del ancho de haz utilizado, que quizá debiera ser más permisivo). Esta hipótesis se evaluará en el apartado 4.4.2.

4.4.1.2.- Escalado lineal: βP(Ω)

A continuación, en la figura 4.14 se va a probar el efecto de penalizar las inserciones aplicando un factor β a las probabilidades del ML sobre la tarea BDGEO. El valor del parámetro *ah* se ha mantenido fijo (*ah*=0.5). En este caso, el parámetro α se mantiene fijo igual a 1. De nuevo los valores del error se sitúan por encima del 30%.

La aplicación de un factor β>1 significa incrementar linealmente todas las probabilidades. Podría haber incluso probabilidades que fueran superiores a la unidad. Por supuesto la aplicación de valores de β superiores a uno empeora la situación, consiguiéndose peores porcentajes de %PRE además de implicar un mayor número medio de nodos activos por frame.

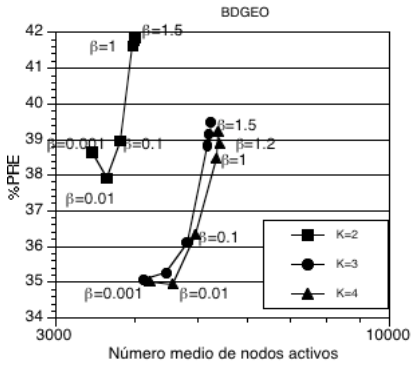


Figura 4.14.- Representa la evolución del error aplicando un barrido de los valores de β para los modelos integrados de orden K con K=2, 3 y 4 para la tarea BDGEO.

Si se aplican valores de β inferiores a la unidad (disminución lineal de todas las probabilidades del ML) para la tarea BDGEO se observa un descenso del error, más o menos en la misma proporción para cada uno de los modelos con distintos valores de K. Este decremento en el error es debido a la disminución de la inserción. De hecho en todos los casos para un valor de β=0.01 se puede observar el cambio. Por ejemplo, para K=4 se pasa de una situación original de %PRB=3.37% y %PRI=10.04% con β=1 a %PRB=4.65% y %PRI=6.70% para β=0.1 y a %PRB=6.26% y %PRI=4.33% para β=0.01. La tendencia es exactamente la misma para todos los valores de K. También, en todos los casos se produce una disminución del número medio de nodos activos en el trellis y por tanto del tiempo medio de decodificación de cada frase. Sin embargo, en todos los casos, el descenso del error es muy poco significativo, por lo que se va a probar del efecto que se produce al aplicar ambos escalados simultáneamente.

4.4.1.3.- Escalado lineal y exponencial: $\beta P(\Omega)^{\alpha}$

Vamos a ver a continuación, en la figura 4.15, cual es el resultado que se obtiene cuando lo que se hace es aplicar ambos factores de escalado α y β simultáneamente, para la tarea BDGEO. De nuevo el parámetro que regula el ancho de haz se ha mantenido fijo *ah*=0.5.

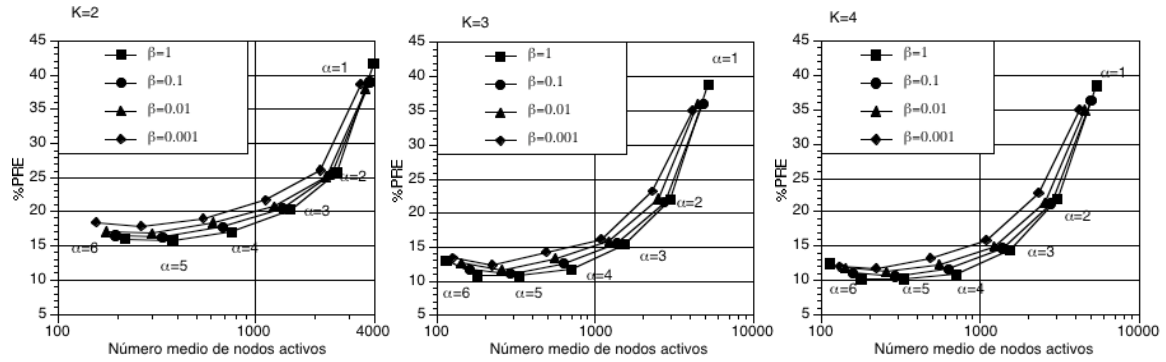


Figura 4.15.- Representa la evolución del error para los modelos integrados de orden K con K=2, 3 y 4 cuando se aplican ambos factores ($\alpha>1$ y $\beta<1$) sobre la tarea de BDGEO.

Se consiguen mejoras en los resultados sólo para $\alpha=1$, debido a la disminución de la inserción como ya habíamos visto en la sección anterior. Sin embargo, para $\alpha>1$ la aplicación del parámetro β empeora ligeramente los resultados (la disminución de la inserción ya está resuelta con la aplicación del parámetro α en sí mismo).

En la sección anterior se había visto que aplicar un parámetro $\beta>1$ con $\alpha=1$ incrementaba ligeramente las tasas de error. Sin embargo, para valores óptimos de α, aplicar un factor $\beta>1$ logra disminuir ligeramente los resultados de error (ver figura 4.16). La escala aplicada en el eje vertical es distinta en cada figura, simplemente para que se pueda apreciar la diferencia (realmente no significativa) alrededor de cada óptimo. Quizá el parámetro $\beta>0$, está compensando (incrementando de forma lineal), el descenso exponencial notable de las probabilidades del ML con valores de α altos. Ésto se podría explicar pensando que el parámetro α toma un valor no completamente adecuado y el parámetro β lo reescala ligeramente.

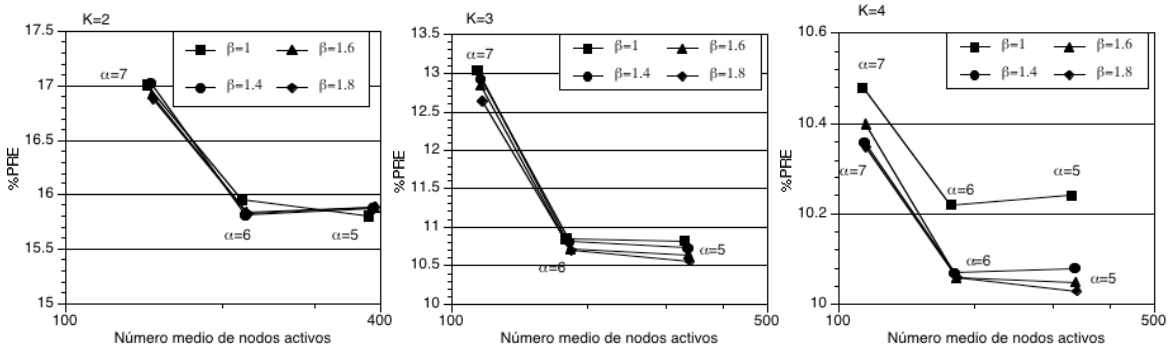


Figura 4.16.- Detalle de la evolución del error para los modelos integrados de orden K con K=2, 3 y 4 cuando se aplican ambos factores (α=5, 6 y 7 (alrededor del óptimo) y $\beta>1$) para la tarea de BDGEO.

Como conclusión se podría decir que un valor de β mayor que uno puede ayudar muy ligeramente al decremento del error cuando se trabaja con el valor de α adecuado.

4.4.1.4.- Aplicación de una función escalón

Cuando se aplica un valor del parámetro $\alpha>1$ se está utilizando una función exponencial que pondera las probabilidades de los MLs de forma que las probabilidades más bajas tienden hacia cero, a medida que el parámetro α aumenta de valor. Se va a probar el efecto que tiene

sobre las tasas de error la aplicación de una función escalón, de modo que sólo aquellas palabras con probabilidades por encima de un cierto umbral *e* sean evaluadas.

La función escalón que se va a aplicar es la siguiente:

$$P(w_i / w_{i-(K-1)}^{i-1}) = \begin{cases} P(w_i / w_{i-(K-1)}^{i-1}) & P(w_i / w_{i-(K-1)}^{i-1}) > e \\ 0 & P(w_i / w_{i-(K-1)}^{i-1}) \leq e \end{cases} \quad (4.21)$$

Los resultados obtenidos para dos valores del parámetro *e* se pueden ver en la figura 4.17. e=0.0, significa que no se ha aplicado ninguna función escalón y por lo tanto los resultados corresponden a los originales.

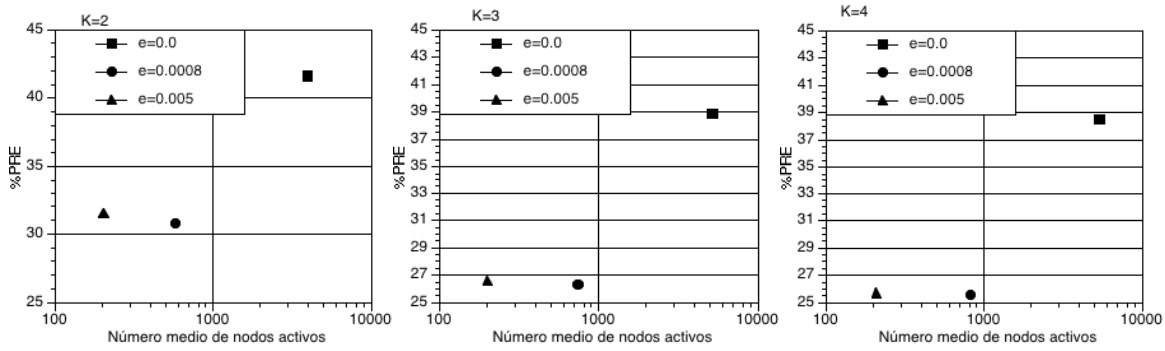


Figura 4.17.- Representa los resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 y distintos valores del parámetro e (e es el parámetro que sitúa el escalón, por tanto, e=0.0 significa no aplicar la función).

Se observa una disminución notable del error, así como del número medio de nodos activos por frame para decodificar las sentencias. Sin embargo, la disminución del error (se pasa desde valores próximos al 39% a próximos al 25%) es menor que la obtenida con la aplicación del valor óptimo del parámetro α (se pasa de valores próximos al 39% al 10%). A continuación, en la figura 4.18, se muestran los resultados de error obtenidos cuando se aplica la función escalón junto con un barrido sobre los valores del parámetro α, es decir se aplica una Función escalón+ Función exponencial de la siguiente forma:

$$P(w_i / w_{i-(K-1)}^{i-1}) = \begin{cases} P(w_i / w_{i-(K-1)}^{i-1})^{\alpha} & P(w_i / w_{i-(K-1)}^{i-1}) > e \\ 0 & P(w_i / w_{i-(K-1)}^{i-1}) \leq e \end{cases} \quad (4.22)$$

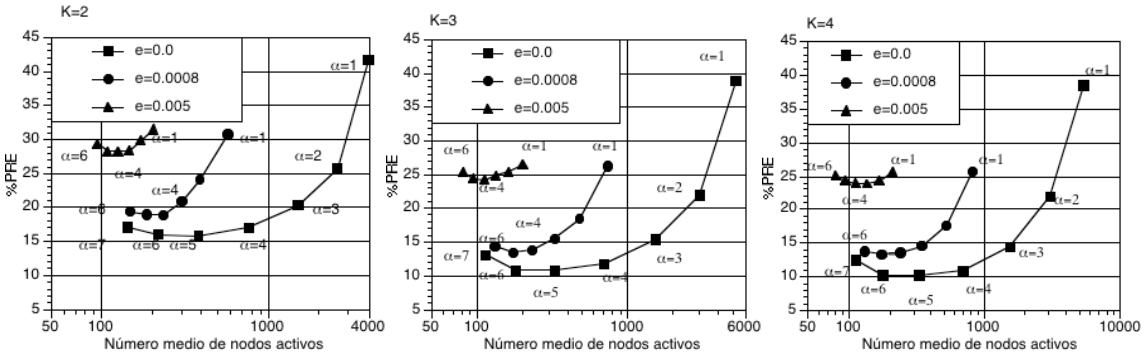


Figura 4.18.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 cuando se modifican los valores del parámetro α en la parte superior del escalón (BDGEO).

Los resultados de error obtenidos con e=0.005 no disminuyen significativamente con el incremento de α. Se han eliminado demasiadas transiciones del ML que son fundamentales en decodificación. Si nos fijamos en los resultados obtenidos con e=0.008 se ve que aplicar una función escalón proporciona buenos resultados para valores de α pequeños porque se eliminan a priori todos aquellos caminos con probabilidades demasiado pequeñas que con valores del parámetro α no muy altos no quedan suficientemente degradadas. Sin embargo, con α=óptimo y en general alto, no se llega a eliminar ningún camino, todos tienen una probabilidad aunque sea muy pequeña y parece que es una mejor solución para alcanzar el mínimo error.

Para eliminar el problema asociado a los caminos con probabilidad cero, cuando se aplica la función escalón directamente, vamos a ver a continuación los resultados que se obtienen cuando se utiliza una función escalón suavizada, de modo que aquellas probabilidades que queden por debajo del umbral dado por el parámetro *es2*, se elevan a una potencia 2*α (probabilidades aún más atenuadas).

$$P(w_i / w_{i-(K-1)}^{i-1}) = \begin{cases} P(w_i / w_{i-(K-1)}^{i-1})^{\alpha} & P(w_i / w_{i-(K-1)}^{i-1}) > es2 \\ P(w_i / w_{i-(K-1)}^{i-1})^{2\alpha} & P(w_i / w_{i-(K-1)}^{i-1}) \le es2 \end{cases} \tag{4.23}$$

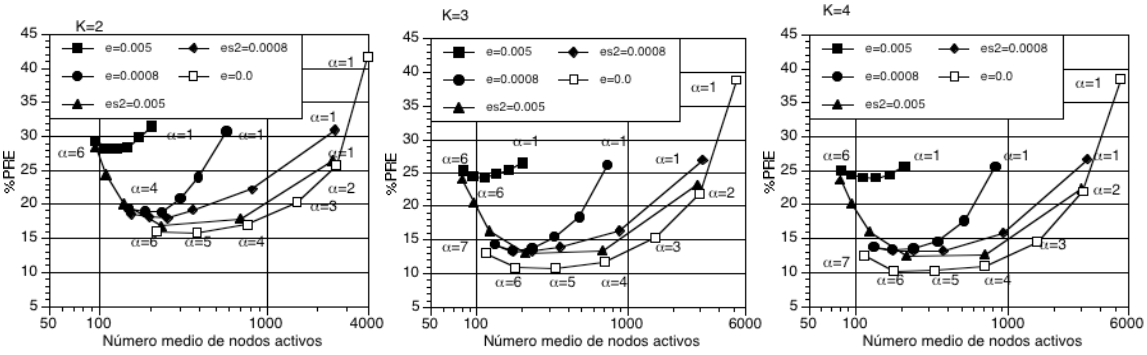


Figura 4.19.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 para una función escalón suavizada. Se aplican distintos valores de α en la parte superior del escalón y 2α en la parte inferior (BDGEO).

En general, se puede ver en la figura 4.19, que se obtienen mejores resultados que los obtenidos con la aplicación directa de la función escalón, pero nunca mejores que los obtenidos directamente con la aplicación de la función exponencial. Se ve que con e=0.005 se obtenían resultados pobres porque se desestimaban demasiados caminos. Sin embargo, al aplicar este método se obtiene en general buenos resultados, lógicamente para valores del parámetro a más pequeños. Hay que tener en cuenta que las probabilidades pequeñas se están disminuyendo mucho.

A pesar de los buenos resultados obtenidos tras aplicar la función exponencial 2α en la parte baja del escalón, parece que las probabilidades bajas se aproximan demasiado rápido a cero. Por ello se va a utilizar de nuevo la función escalón suavizada pero en la parte baja del escalón se utiliza una función exponencial $\alpha+1$ (ver ecuación 4.24).

$$P(w_i / w_{i-(K-1)}^{i-1}) = \begin{cases} P(w_i / w_{i-(K-1)}^{i-1})^{\alpha} & P(w_i / w_{i-(K-1)}^{i-1}) > es1 \\ P(w_i / w_{i-(K-1)}^{i-1})^{1+\alpha} & P(w_i / w_{i-(K-1)}^{i-1}) \le es1 \end{cases} \quad (4.24)$$

En la figura 4.20 se muestran los resultados obtenidos observándose muy buenos resultados pero nunca superando los resultados originales. La ventaja que se puede apreciar al utilizar la función escalón suavizada es que la dispersión del %PRE entre α=1 y α=óptimo es menor. Se pasa de aproximadamente el 27% de error al 10%, frente a la utilización directa de la atenuación de las probabilidades del ML mediante el parámetro α, en la que se pasa del 39% al 10%.

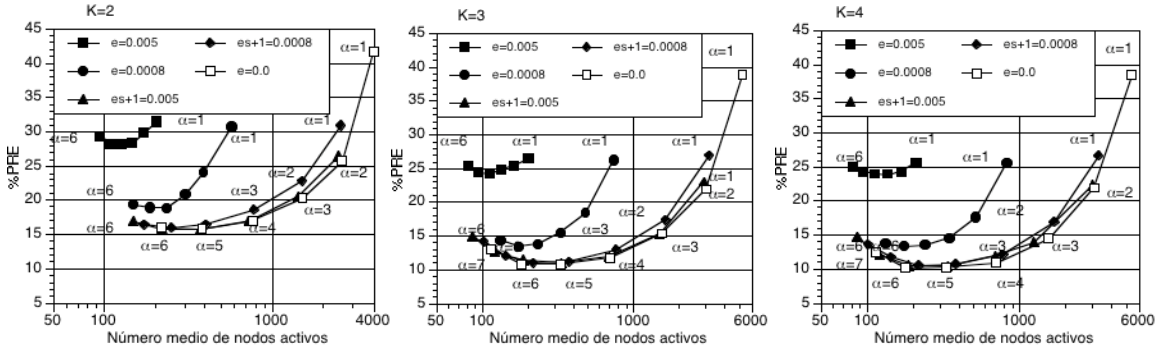


Figura 4.20.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 para una función escalón suavizada. Se aplican distintos valores de α en la parte superior del escalón y $1+\alpha$ en la parte inferior (BDGEO).

4.4.2.- Resultados ponderando las probabilidades acústicas

En todos los resultados presentados hasta el momento sólo se han modificado las probabilidades del ML. Vamos a ver ahora que ocurre cuando se alteran las componentes acústicas.

La componentes acústicas se pueden alterar (incrementar) directamente aplicando un factor γ a las probabilidades acústicas $P(A/\Omega)^{\gamma}$ con $\gamma<1$ o bien utilizando distintos conjuntos (menor número) de parámetros acústicos. Todos los resultados obtenidos hasta este punto se han realizado con modelos acústicos discretos con 4 codebooks. Es de suponer que si se utilizan un menor número de componentes acústicas, los resultados de decodificación serán peores, no obstante, es una alternativa para disminuir la aportación de la componente acústica y ver cual es su relación con las probabilidades del ML.

En ambas propuestas ($P(A/\Omega)^{\gamma}$ o utilización de menor número de componentes acústicas), la importancia de las probabilidades acústicas disminuye y se verá el efecto sobre los resultados de error finales. Se va a evaluar también el efecto que tiene la utilización de estas aproximaciones junto con la función exponencial de atenuación de las probabilidades del ML regulada por el valor del parámetro α vista en las secciones previas.

4.4.2.1.- Función exponencial: $P(A/\Omega)^{\gamma}$

Como ya se ha mencionado anteriormente, las diferencias entre las probabilidades acumuladas en el desarrollo de Viterbi es normalmente más grande que las diferencias existentes entre las probabilidades del ML. Esto es fundamentalmente debido a las probabilidades acústicas (que se aplican un número mucho mayor de veces en el trellis de Viterbi y además sus valores de probabilidad son muy pequeños). La consecuencia inmediata es que los valores de las probabilidades del ML son irrelevantes en la mayor parte de las situaciones para elegir el mejor camino a seguir. Para evitar este problema, en experimentos previos se ha visto el efecto de elevar las probabilidades del ML a una potencia $(P(\Omega))^{\alpha}$. Vamos a estudiar ahora cual es el efecto de elevar a una potencia los valores de las probabilidades acústicas $(P(A/\Omega))^{\gamma}$. No se pueden utilizar valores de $\gamma>1$ puesto que esto implicaría reducir todas las probabilidades acústicas y potenciar aun más la influencia de las mismas en decodificación. En este caso, para que las probabilidades del ML fueran competitivas y contrarrestar este efecto habría que aumentar aún más el valor de α. Se han realizado pruebas en este sentido observando realmente este efecto.

A continuación vamos a ver los resultados de error que se han obtenido para la tarea BDGEO (figura 4.21) cuando se aplica un valor fijo de ancho de haz *ah*=0.5 y distintos valores del parámetro γ.

La aplicación de un valor de γ entre $0<\gamma<1$ supone reescalar las probabilidades acústicas incrementando su valor y por tanto hacerlas menos importantes. Se observa que a medida que el valor de γ se hace mas pequeño se decrementa notablemente el error, pero se incrementa excesivamente el número medio de nodos activos por frame necesarios para decodificar cada sentencia (efecto contrario al observado cuando se aplica un valor de $\alpha>1$).

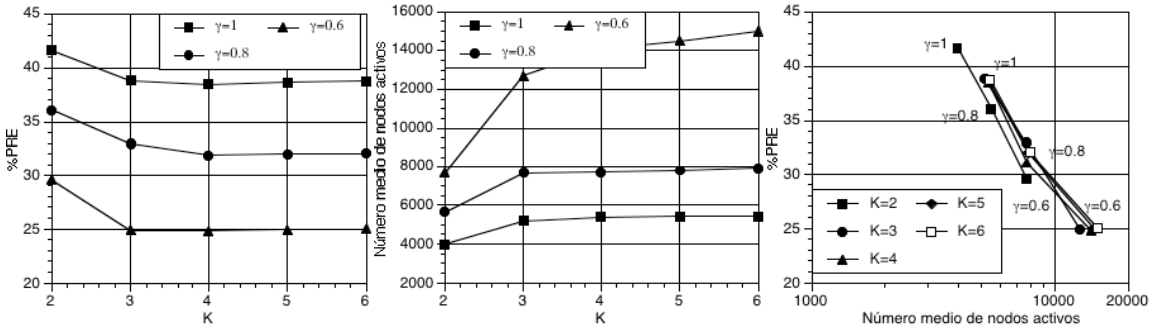


Figura 4.21.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ (*ah*=0.5).

El incremento del número medio de nodos activos es debido a que la aplicación de un factor γ es "en cierto modo equivalente" a aumentar el parámetro que regula el ancho de haz (ver figura 4.10), ya que al ser las probabilidades acumuladas menores (compuestas fundamentalmente de probabilidades acústicas que son más pequeñas y se aplican un mayor número de veces) se activan una mayor cantidad de hipótesis.

Sin embargo, cuando se muestran los resultados obtenidos con distintos factores de *ah* se podía ver (figura 4.10) que a partir de *ah*=0.5 para α=1 las tasas de error no disminuían, y lo único que aumenta es el porcentaje de nodos activos por frame. En este caso al alterar el valor de "γ" las tasas de error también disminuyen porque hay un doble efecto. Aunque α =1, el ML tiene una influencia mayor ya que las diferencias entre las probabilidades del ML se mantienen fijas mientras que son las probabilidades acústicas las que disminuyen y por tanto las probabilidades acumuladas totales. Por tanto, con α=1 no sólo es que se activen un mayor número de hipótesis (equivalente a un mayor ancho de haz) sino que las probabilidades del ML tienen una mayor influencia y por eso disminuyen las tasas de error.

En resumen, γ <1 supone una ampliación del número de hipótesis que se mantienen activas, y el descenso de las tasas de error es debido fundamentalmente a una mayor actuación de las probabilidades del ML. Para K=2 con α=1 es imposible lograr tasas de error inferiores al 40% ni inferiores al 35% para K=3 y K=4 por mucho que se incremente el ancho de haz (figura 4.10). Sin embargo a medida que el parámetro γ decrece las tasas aumentan llegándose a valores del 30% para K=2 y del 25% para K=3 y 4 siendo necesarios un mayor número medio de nodos activos por frame (excesivo para valores bajos del parámetro γ).

A continuación vamos a ver los resultados de error que se han obtenido para la tarea BDGEO con K=2, K=3 y K=4 (figura 4.22) cuando se realiza un barrido del parámetro α aplicado a las probabilidades del ML. En la figura 4.23 se pueden ver los resultados en detalle para los valores alrededor del óptimo.

La aplicación de un parámetro α>1 supone en todos los casos un decremento del porcentaje de error hasta llegar a un óptimo. Se ve que a medida que disminuye el valor del parámetro γ, el error disminuye alrededor del óptimo aunque también aumenta considerablemente el número medio de nodos activos por frame y por tanto el tiempo necesario para decodificar cada sentencia.

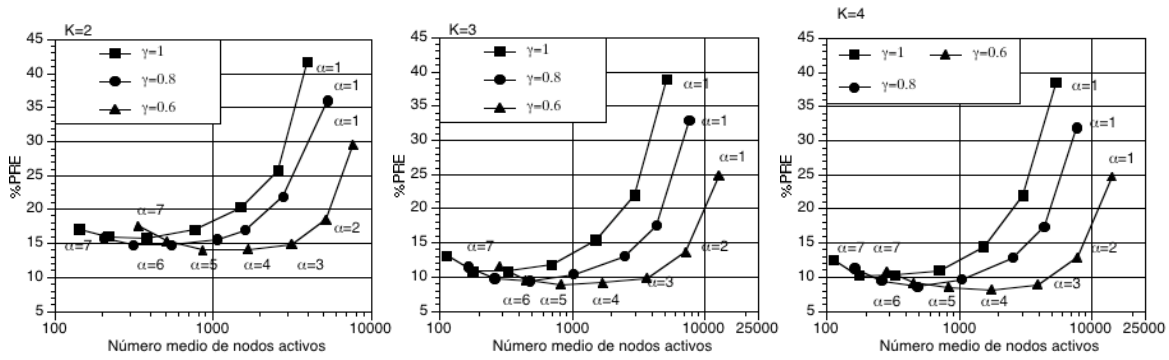


Figura 4.22.- Resultados de error para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ y del ML α (*ah*=0.5).

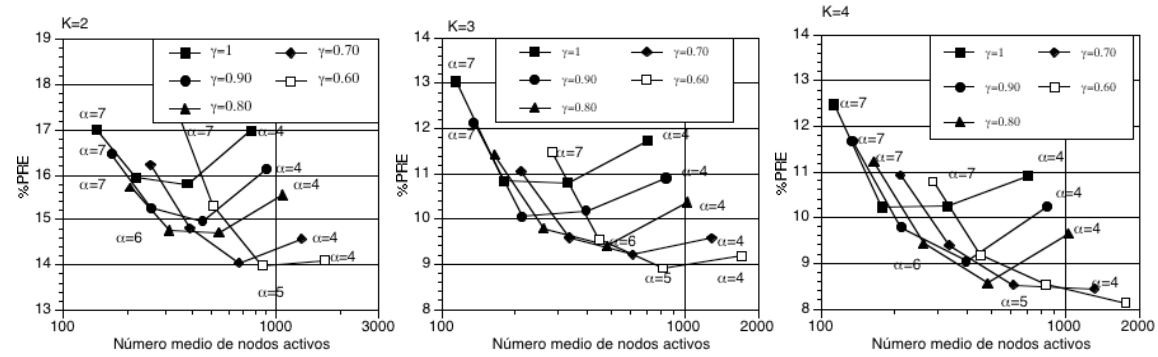


Figura 4.23.- Detalle de los resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ y del ML α=4, 5, 6 y 7 (*ah*=0.5).

A continuación se presentan los resultados obtenidos para valores de ancho de haz más restrictivos. Los resultados para *ah*=0.6 (figura 4.24) son muy similares a los obtenidos en el caso anterior (*ah*=0.5) y de nuevo se observa que a medida que disminuye γ el error es menor, aunque es mayor el número medio de nodos activos por frame. Sin embargo, para valores relativamente grandes del parámetro α el comportamiento es mucho peor, lo cual es lógico debido a la normal degradación que sufren las probabilidades de ML al otorgárseles una importancia excesivamente mayor que a las probabilidades acústicas.

Para *ah*=0.81 (ancho de haz bastante restrictivo) con (γ =1, α=1) se obtienen unas tasas de error muy altas, ya que hay una poda excesiva de los caminos a seguir. Un incremento de α=2 supone una mejora muy ligera y a continuación una degradación del modelo. En este caso los porcentajes de error disminuyen muy notablemente a medida que disminuye el parámetro γ, pero el hecho de utilizar valores de α>2 supone una degradación fortísima del modelo. Para un valor de γ =0.1 esta degradación comienza incluso con α=2. En este caso se llega a las tasas más bajas de error sin tener que atenuar las probabilidades del ML, sin embargo el número medio de estados es exageradamente alto.

Cuando se utiliza un valor de ancho de haz aún más restrictivo *ah*=0.9. Para γ =1 y α=1 las tasas de error son muy altas y es contraproducente la utilización del parámetro α en todos los casos, ya que hay un número excesivo de caminos podados. Al aumentar el valor de γ se

decrementan notablemente las tasas de error no siendo excesivo el número medio de nodos que hay que mantener activos.

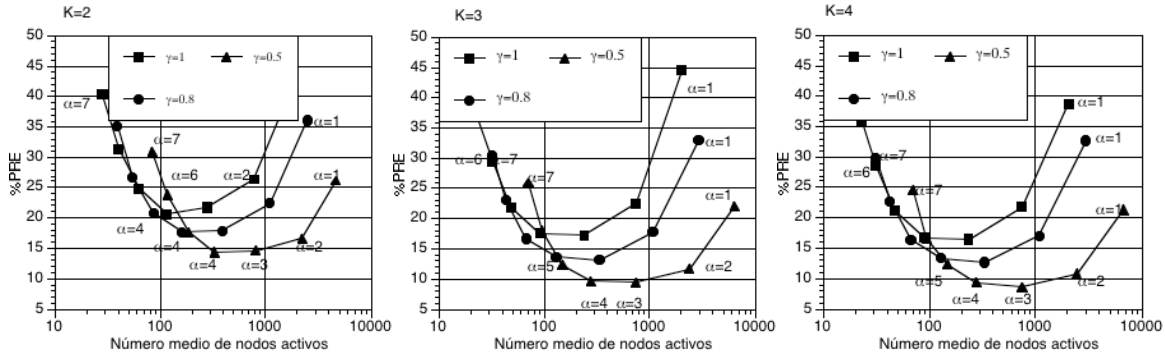


Figura 4.24.- Resultados de error para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ y del ML α (*ah*=0.6).

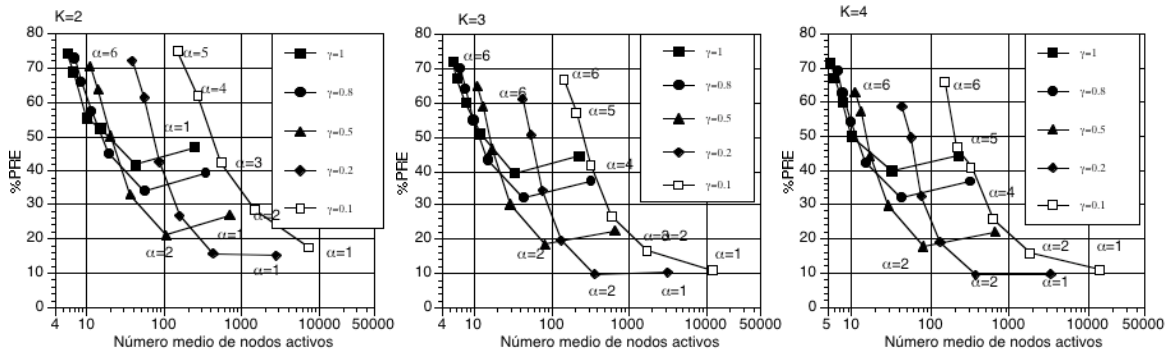


Figura 4.25.- Resultados de error para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ y del ML α (*ah*=0.8).

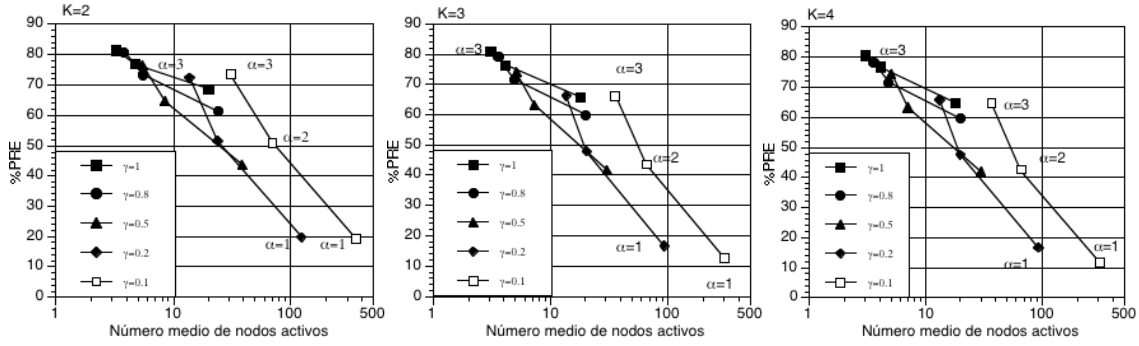


Figura 4.26.- Resultados de error para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores del parámetro de escalado acústico γ y del ML α (*ah*=0.9).

En resumen se puede decir que el parámetro α no es fundamental para obtener las mejores tasas de decodificación (tasas de error más bajas, aunque puede aumentar significativamente el número medio de nodos activos). Es posible conseguir prácticamente los mismos resultados pero para ello es necesario la redistribución de las probabilidades acústicas, utilizando el parámetro γ, de manera que las diferencias entre las probabilidades del ML sean significativas. En general seguiremos usando el parámetro α, porque hemos visto que la utilización del parámetro γ es otra manera de conseguir lo mismo y supone la multiplicación por

un factor un número mayor de veces puesto que las probabilidades acústicas se utilizan un mayor número de veces. Por tanto, aunque se mantenga el número medio de nodos activos por frame se incrementan los tiempos de decodificación y nos alejamos del tiempo real.

4.4.2.2.- Utilización de distintos conjuntos de parámetros acústicos

En base a los experimentos de DAF presentados en la subsección 4.1.4 se concluyó que los mejores resultados se obtenían con la utilización de modelos acústicos de 4 codebooks. Con este tipo de modelos se ha realizado toda la experimentación previa. A continuación vamos a ver los resultados que se obtienen cuando se reduce la importancia de la componente acústica. Es decir, aplicando solo 3 MC (Cepstrales, Primera derivada de los Cepstrales, Energía y derivada) o 2MC (Cepstrales, Energía y derivada). En la figura 4.27 se muestran los resultados obtenidos manteniendo fijo el valor del parámetro *ah*=0.5.

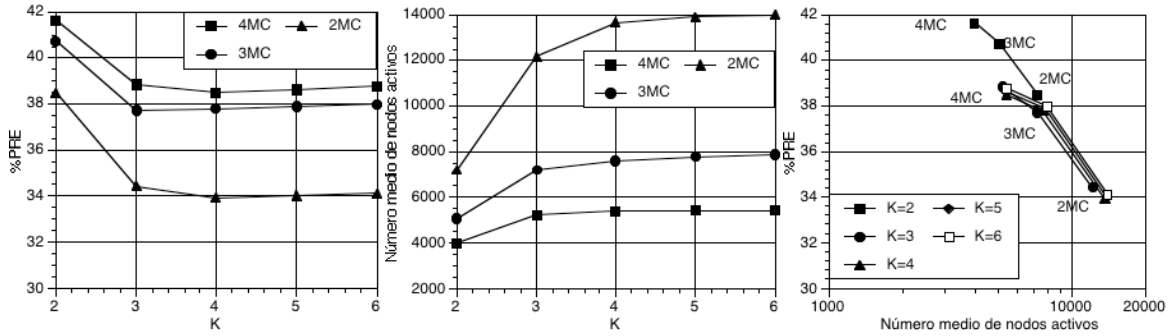


Figura 4.27.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 cuando se utilizan diferentes conjuntos de componentes acústicas (*ah*=0.5).

Se observa una disminución del porcentaje de error con 3MC, que es aún mayor con 2MC, aunque también aumenta el número medio de nodos activos. Al utilizar 2MC lógicamente las componentes acústicas tienen menor importancia (implica una menor probabilidad acumulada) y por tanto el ancho de haz utilizado, que se mantiene fijo, permite activarse un mayor número de hipótesis.

Que se obtengan mejores resultados de error utilizando un menor número de codebooks no quiere decir que los modelos acústicos de 4MC sean peores que los de 2MC. Lo que ocurre es que para valores de α<óptimo las componentes acústicas tienen una excesiva relevancia frente a las del ML y por ello los porcentajes de error son demasiado altos. Por tanto, el usar modelos acústicos con menor información hace que las probabilidades del ML sean más competitivas y favorece la decodificación de las sentencias.

En la figura 4.28 se pueden ver los resultados obtenidos cuando se realiza un barrido experimental sobre el parámetro α que regula la ponderación sobre las probabilidades del ML.

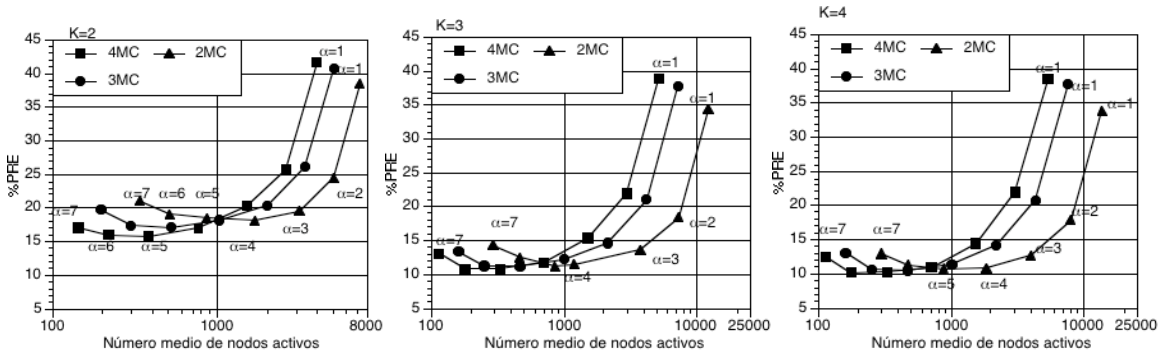


Figura 4.28.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3y 4 cuando se utilizan diferentes conjuntos de componentes acústicas y un barrido experimental sobre el parámetro α (*ah*=0.5).

A medida que aumenta el valor de α, se produce una notable disminución del error en todos los casos. Para K=2 y 2MC, el mínimo de error logrado con α es bastante menor que con 4MC. A medida que aumenta el valor de K se llega prácticamente a los mismos resultados (con K=4 se logra el óptimo) aunque se necesita un mayor número medio de nodos activos. En la práctica alrededor del óptimo se consiguen resultados similares a los que se obtendrían si se estuviera utilizando un ancho de haz menos restrictivo.

A continuación, en la figura 4.29 vamos a ver lo que ocurre cuando se utiliza una *ah* más restrictivo. Para α=1 los resultados en %PRE son notablemente mejores con 2MC. En todos los casos se llega al optimo de error para α=4, siendo mejores los resultados para 2MC aunque con la curva que muestra el comportamiento del sistema para distintos valores del parámetro α, desplazada a la derecha, es decir, se necesita un mayor número medio de nodos activos.

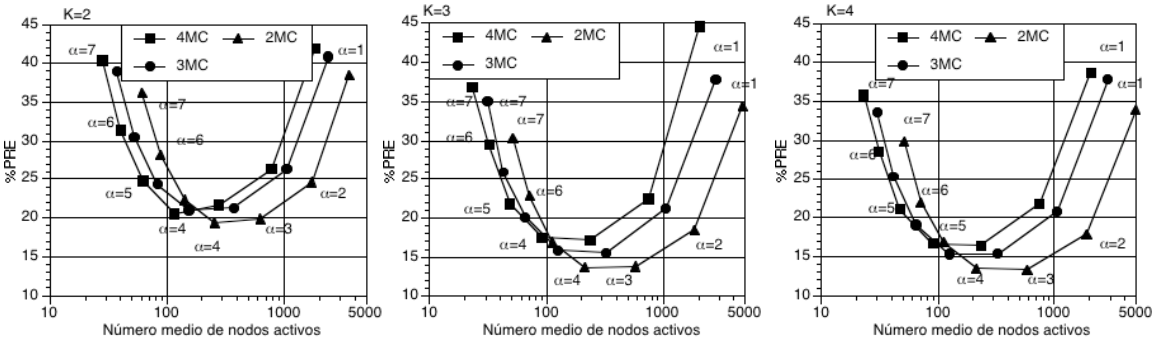


Figura 4.29.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 cuando se utilizan diferentes conjuntos de componentes acústicas (*ah*=0.6).

La conclusión que se puede extraer es que si se utiliza un *ah* más estrecho *ah*=0.6 se obtienen mejores resultados en %PRE con el α=óptimo si se utiliza menos información acústica. Esto se produce porque hay un mayor número de caminos que no se desactivan, por lo que también aumenta el número medio de nodos activos. Cuanta menos información acústica (menor número de MC) se utilice con el mismo *ah*, los resultados que se obtienen son comparables a los obtenidos al utilizar un *ah* menor (más permisivo).

4.4.3.- Utilización de modelos acústicos más precisos

Como ya se ha dicho anteriormente para mejorar los resultados obtenidos con MOM discretos que representan unidades incontextuales existen fundamentalmente dos opciones:

a) utilizar modelos más complejos (por ejemplo modelos semicontinuos) y mantener el número de unidades subléxicas incontextuales (reducido).

b) seguir usando los modelos acústicos simples (discretos) e incrementar el número de unidades (por ejemplo utilizando unidades contextuales obtenidas a partir de árboles de decisión).

Vamos a ver a continuación los resultados que se obtienen cuando se utilizan modelos acústicos con mayor información, o bien porque las distribuciones de probabilidad son continuas (usando modelos semicontinuos) o bien porque los modelos representan unidades contextuales obtenidas a partir de árboles de decisión.

4.4.3.1.- Resultados con MOM semicontinuos

En este caso, los resultados sólo se presentan para α=4, 5, 6 y 7 ya que el tiempo necesario para la decodificación de las sentencias que hasta ahora era prácticamente el mismo, se incrementa en mas de 10 veces al utilizar modelos Semicontinuos (SC) frente a la utilización de modelos Discretos (D).

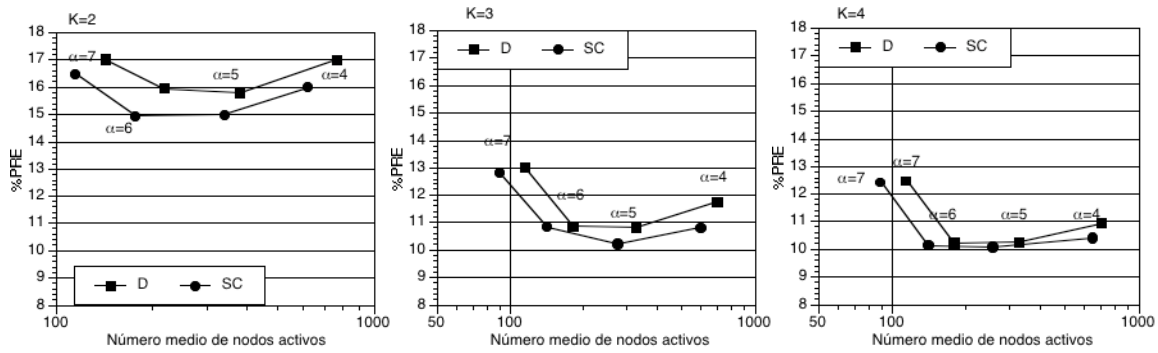


Figura 4.30.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 para α=4, 5, 6 y 7 utilizando modelos semicontinuos (BDGEO).

Se observa una ligera mejora del porcentaje de error y disminución del número medio de nodos activos por frame (sobre todo para valores bajos de K), que no se ve compensado por el tiempo de decodificación necesario para lograr los resultados.

4.4.3.2.- Resultados con unidades contextuales

A continuación se van a presentar los resultados obtenidos cuando se trabaja con unidades contextuales. Al igual que ocurría en el reconocimiento de palabras aisladas (ver subsección 4.2.2.2) surge el problema en la transcripción del léxico para la primera y la última unidad de cada palabra ya que no están definidos los contextos. O bien se utilizan las unidades incontextuales en los bordes dando lugar el experimento Arboles de Decisión en los bordes

Incontextuales (AD_bor_I) o se utilizan unidades contextuales de contexto izquierda o derecha en los bordes dando lugar el experimento Arboles de Decisión en los bordes Contexto izquierda- Contexto derecha (AD_bor_CiCd) (López de Ipiña et al, 1999b). En la figura 4.31 se presentan los resultados obtenidos.

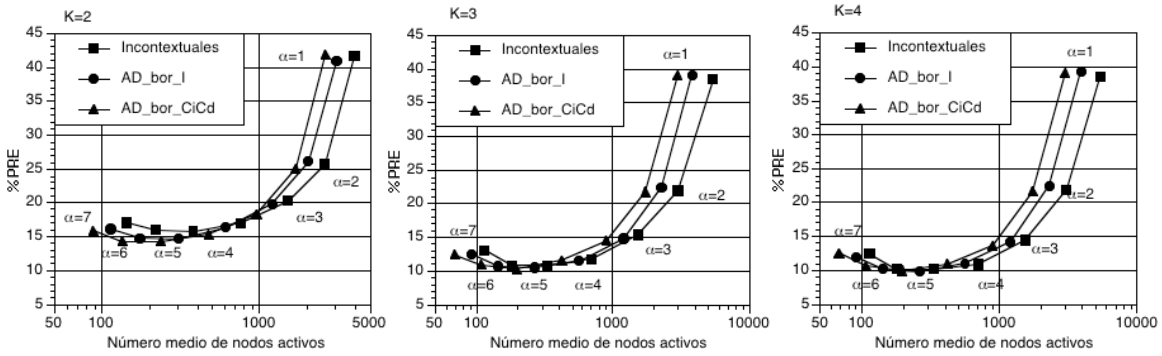


Figura 4.31.- Resultados de error obtenidos para los modelos integrados de orden K con K=2, 3 y 4 con distintos valores de α, cuando se utilizan las unidades contextuales para la tarea BDGEO.

Se observa una ligera mejora en los resultados para K=2. Para K=3 y K=4 se obtienen más o menos las mismas curvas de %PRE en función del parámetro α, un poco desplazadas hacia la izquierda (un menor número de nodos activos por frame). En la siguiente figura 4.32 se pueden observar los resultados en detalle para α=4, 5 y 6.

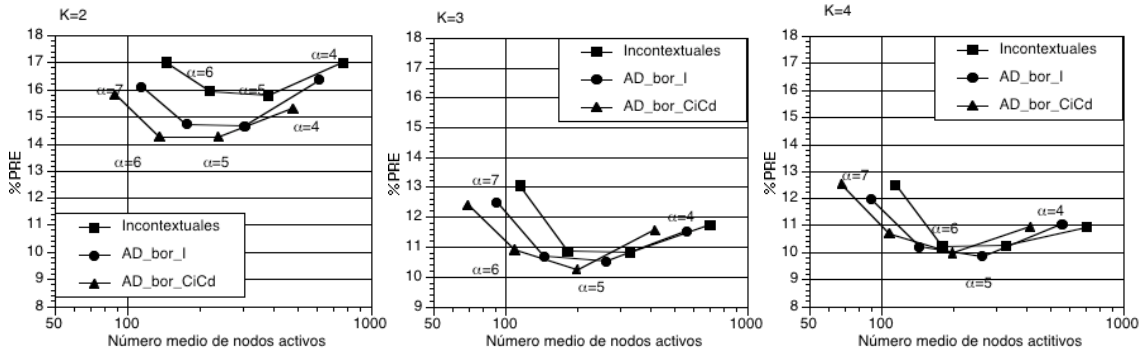


Figura 4.32.- Resultados de error obtenidos para los modelos integrados de orden K con K=2 , 3 y 4 a partir de las unidades contextuales (detalle para α=4, 5, 6, 7) BDGEO.

El uso de unidades contextuales parece aportar mejoras poco significativas a los resultados de decodificación sobre todo para valores relativamente altos de K. Este hecho podría estar unido al problema asociado a la transcripción de los bordes de las palabras, por lo que se podrían evaluar otras alternativas, aunque ese estudio no es objetivo del presente trabajo.

4.4.4.- Conclusiones

Tras estudiar los efectos de la utilización de distintos factores de ancho de haz, las modificaciones de las probabilidades acústicas y del ML, así como la utilización de modelos más precisos (modelos acústicos semicontinuos o modelos de unidades contextuales) se han observado las siguientes conclusiones:

Ancho de Haz (*ah*):

- Con valores del ancho de haz muy restrictivos el error es alto.

- A medida que se incrementa el ancho de haz, se decrementa el error hasta un umbral a partir del cual el error permanece estacionario.

- A medida que se incrementa el ancho de haz, se aumenta considerablemente el número medio de nodos activos (lo cual implica mayor tiempo de decodificación).

Ponderación de las componentes sintácticas:

* *Función exponencial* $P(w_i / w_{i-(K-1)}^{i-1})^{\alpha}$

- Redistribución de la distribución de probabilidades del ML (disminución desigual de las probabilidades sintácticas $P(w_i / w_{i-(K-1)}^{i-1})$).

- Disminución del número medio de nodos activos a medida que aumenta el valor de α.

- Descenso muy notable del error hasta un valor del parámetro α óptimo a partir del cual el error comienza a ascender de forma suave.

- Con α=1 la componente sintáctica no resulta relevante en el proceso de decodificación.

- Con α excesivo (>>óptimo) la componente acústica pierde su importancia. El resultado son frases muy cortas sintácticamente correctas.

* *Función lineal* $\beta P(w_i / w_{i-(K-1)}^{i-1})$

- Disminución lineal de las probabilidades sintácticas.

- Disminución ligera del número medio de nodos activos.

- Disminución poco significativa de las tasas de error.

* *Función lineal y Función exponencial* $\beta\ P(w_i / w_{i-(K-1)}^{i-1})^{\alpha}$

- El error es siempre algo mayor que el observado con la aplicación aislada de la función exponencial.

- No hay diferencia con respecto a la aplicación de la función exponencial en el número medio de nodos activos.

- Sólo es posible decrementar de forma muy poco significativa el error con la introducción de un valor de β>1 para el valor óptimo de α.

* *Función escalón*

- No se alteran las probabilidades que superan un cierto umbral, las demás son cero.

- Descenso notable del error y del número medio de nodos activos.

- Mejores resultados que los observados con la aplicación de la función lineal, pero peores que los obtenidos con el valor de α óptimo.

* *Función escalón y Función exponencial*

- Se incrementan los resultados con respecto a la aplicación de la función escalón si el umbral *e* no es demasiado restrictivo (alto).

- No se llega en absoluto a la obtención de los resultados óptimos (logrados con α).

* *Función escalón suavizada y Función exponencial*

- Se aplica una función exponencial más restrictiva (mayor decremento de la probabilidad) para las probabilidades por debajo del umbral *e*.

- Se llegan a obtener los mismos valores óptimos que los obtenidos con α, con un número medio de nodos activos similar.

- Ventaja: la diferencia en %PRE entre α=1 y α=óptimo es menor.

Ponderación de las componentes acústicas

* *Función exponencial* $P(A/\Omega)^{\gamma}$

- Decremento notable del error a medida que disminuye γ.

- Al contrario que lo que ocurría con el parámetro α, incremento notable del número medio de nodos activos.

- Para hacer competitivo el uso del parámetro γ y llegar a los resultados obtenidos con α (en %PRE y en número medio de nodos activos) hay que reducir enormemente el ancho de haz.

* *Función exponencial sobre los modelos acústicos y sobre los ML:* $P(A/\Omega)^{\gamma} \; P(w_i / w_{i-(K-1)}^{i-1})^{\alpha}$

- Cuando se trabaja con valores del ancho del haz intermedios, se necesita aplicar junto con γ un valor del parámetro α>1 para lograr los resultados óptimos.

* *Aplicación de distintos número (inferior) de parámetros acústicos*

- Se obtienen peores resultados óptimos (obtenidos con la aplicación del parámetro α) ya que hay menor información acústica a medida que se decrementa el número de conjuntos.

- Cuando α=1 o inferior al óptimo, se consiguen mejores resultados en %PRE cuanto menor información acústica se aplique.

- El número medio de nodos activos se incrementa ya que el ancho de haz permite la activación de un mayor número de hipótesis.

- Cuando se decrementa el ancho de haz se pueden conseguir mejores resultados aplicando un menor número de parámetros acústicos.

Aplicación de modelos acústicos más precisos

* *Modelos semicontinuos*

- Descenso ligero del %PRE y del número medio de nodos activos.

- El tiempo medio necesario para la decodificación se multiplica por más de 10.

* *Unidades contextuales*

- Problema asociado a la transcripción del borde de las palabras.
- Descenso ligero del número medio de nodos activos.
- El %PRE se mantiene.

4.5.- Arquitectura

Cuando se trabaja con el prototipo *on-line* se usa la placa de adquisición de una estación de trabajo Silicom Graphic O2 y un micrófono "hypercardioid" dinámico (transductor por presión de gradiente). Para los experimentos realizados en este trabajo (off-line) la adquisición de la base de datos BDGEO se realizó en Barcelona (ver apéndice A) y la de MLA en el departamento de Electricidad y Electrónica (ver de nuevo apéndice A). Una vez realizada la adquisición, la etapa de preproceso que sufren las señales (prototipo *on-line* o bases de datos pregrabadas) es exactamente la misma.

El sistema, tal y como fue pensado en su concepción, es distribuido y está basado en una estructura cliente-servidor. El servidor es el programa principal e incluye el módulo de búsqueda. El cliente, junto con el interface gráfico se encarga de la adquisición, procesamiento y posterior extracción de características de la señal de audio. A través del interface gráfico, basado en el Tcl/Tk (Ousterhat, 1994), el usuario puede controlar el sistema completo. La respuesta del sistema se da a través de mensajes en la pantalla. Los programas cliente y servidor pueden ejecutarse en máquinas diferentes y pueden estar trabajando en paralelo a través de hebras POSIX: por ejemplo el interface gráfico, adquisición de audio y extracción de características son tres hebras independientes del programa cliente. La figura 4.33 muestra las principales características de la arquitectura (Rodríguez et al. 1999b).

El uso de este tipo de arquitectura en paralelo y distribuida permite la máxima eficiencia del sistema. Es evidente que las aplicaciones del cliente deben funcionar en una maquina que posea utilidades gráficas y hardware de audio, sin embargo la potencia de CPU no es un punto fundamental. En contraste, las aplicaciones del servidor deben funcionar en una maquina con CPU muy rápida o bien con más de una CPU. Como el lógico, ambos cliente y servidor pueden estar ejecutándose en la misma maquina aunque el funcionamiento del sistema será menos eficaz.

Una vez que el prototipo ha sido puesto en funcionamiento la señal de audio se adquiere continuamente. Sin embargo, esta señal de audio no se procesa mientras la energía no supera un cierto umbral (si está por debajo del umbral se supone silencio) durante un cierto tiempo. Se trabaja con una cola circular de 30 segundos. En el momento en el que se detecta una señal de voz, se pasan los datos de audio al modulo de extracción de características. El segmento de señal que se envía consta también de los 10 ms anteriores y posteriores (para no perder fricativas iniciales, etc). A medida que se van extrayendo los vectores acústicos, se envían al servidor para

ser sumados a una estructura de tipo cola. Si la señal presenta saturaciones o es demasiado baja aparecen mensajes en la pantalla y no se sigue adelante con el reconocimiento. El servidor dispone de un modulo opcional de Cuantificación Vectorial que funciona cuando en el fichero de configuración se especifique que se va a trabajar con modelos discretos. El control y configuración de los mensajes entre cliente y servidor se puede activar en cualquier momento usando un canal de comunicación bidireccional e independiente.

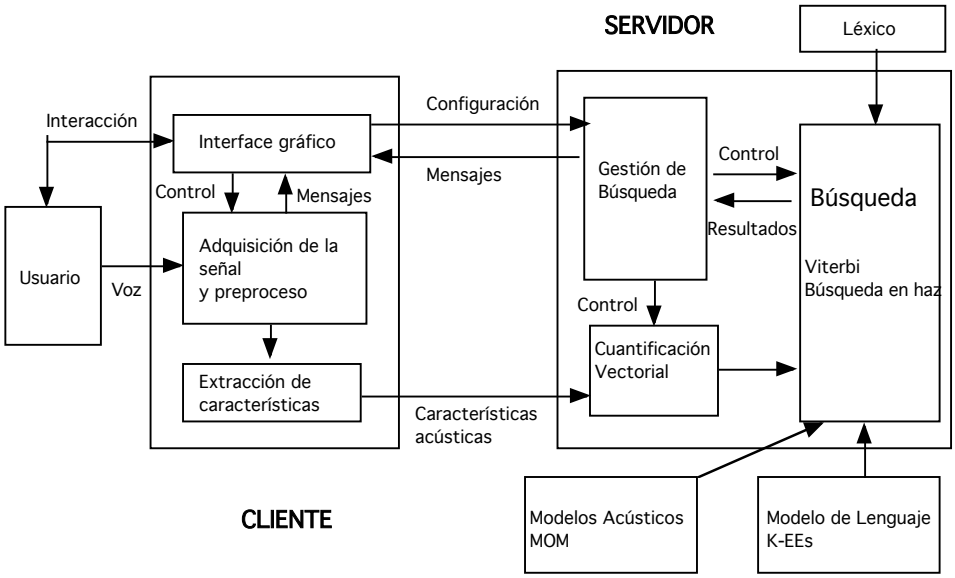


Figura 4.33.- Arquitectura del prototipo TORNASOL. Se muestra la comunicación entre el cliente y el servidor a través de una red local.

Capítulo 5

Evaluación del suavizado del ML en el sistema de RAH

Como se ha visto en el capítulo anterior, los mejores resultados de decodificación se obtienen cuando se introduce un reescalado sobre una o ambas distribuciones de probabilidad (acústica y/o del ML). La necesidad real de esta redistribución de las probabilidades pone de manifiesto que la distribución de probabilidad obtenida tras el suavizado de los modelos no es la más adecuada (sobrestima las probabilidades de los eventos no vistos en entrenamiento). La estimación de esta distribución se obtiene (ver capítulo 3) tras la aplicación de una técnica de suavizado que reescala las probabilidades del ML, sustrayendo una cierta cantidad de probabilidad de los eventos vistos en entrenamiento y distribuyéndola entre los eventos no vistos.

Como se ha presentado anteriormente en este trabajo, se utiliza la técnica de back-off sintáctico ya que se puede integrar de forma eficiente en el esquema recursivo de construcción del modelo que integra K modelos *k*-EE y que se utiliza como ML en el sistema de RAH. En los capítulos precedentes se ha evaluado el descuento de Witten-Bell en el que el factor de descuento depende del número de eventos distintos vistos en cada estado (ver capítulo 3). Pero este factor de descuento no es la única posibilidad, por tanto una vez visto que las distribuciones de probabilidad obtenidas deben ser alteradas para obtener el óptimo en decodificación, en este capítulo se van a evaluar diferentes propuestas de descuento, todas ellas desarrolladas bajo el formalismo dado por el back-off sintáctico.

Las alternativas más conocidas son: el descuento Absoluto y Lineal. En ambos casos el descuento depende del valor de un parámetro. También se probará un descuento muy simple basado en sumar un uno al número de eventos vistos en cada estado, llamado descuento Sencillo. En estos casos, igual que con el descuento de Witten-Bell, el descuento se realiza sobre todos los eventos vistos en cada estado. Como alternativa se va a probar un nuevo método llamado descuento Acotado y que se basa en el descuento de Turing. Este descuento depende del valor de un parámetro y sólo se alteran las probabilidades de los eventos vistos menos de un determinado número de veces (mal estimados). Todos estos descuentos se van a comparar con el descuento de Wittten-Bell que es la base de la experimentación desarrollada en este trabajo.

En este capítulo se van a presentar los distintos resultados experimentales que se han obtenido tanto a nivel de perplejidad como de error del sistema cuando los ML suavizados se integran dentro del sistema completo de RAH (prototipo TORNASOL). Como se ha visto en el capítulo anterior, la aplicación de una función exponencial sobre las probabilidades del ML es fundamental para obtener los mejores resultados de decodificación. En este capítulo se va a probar experimentalmente la fuerte relación existente entre las distintas distribuciones de probabilidad suavizadas, (dependientes de las técnicas de descuento) y el factor de escalado de las probabilidades del ML, necesario para alcanzar la máxima eficacia del sistema. Además se va a demostrar con resultados experimentales que no siempre el mejor suavizado de los ML evaluado en perplejidad resulta ser el mejor suavizado para conseguir el rendimiento óptimo del sistema lo que permitiría cuestionar la perplejidad como medida que permita predecir el comportamiento del ML en el sistema integrado.

La distribución del capítulo es la siguiente: en primer lugar se evalúa la incidencia directa e importante que tiene sobre la perplejidad y sobre el error la aplicación de un heurístico que regule el reescalado (disminución) de las probabilidades de transición a los estados de back-off (probabilidades relacionadas con las transiciones por los eventos no vistos en entrenamiento). Se verá como esta atenuación de las probabilidades implica en algunos casos tasas de error inferiores, lo que confirma la necesidad de una formulación alternativa a la obtención de las probabilidades suavizadas. Como consecuencia el resto del capítulo se dedica al estudio de distintas alternativas de descuento. Para ello, se presentan en primer lugar el conjunto de técnicas de descuento (dentro del back-off sintáctico) que serán posteriormente evaluadas tanto en perplejidad como en porcentaje de error del sistema de RAH.

La evaluación de las técnicas de descuento se lleva a cabo sobre la tarea BDGEO (ver Apéndice A). Una vez obtenidas las conclusiones, se evalúan las técnicas de descuento más relevantes sobre la tarea MLA para validar si las conclusiones siguen siendo válidas sobre bases de datos de vocabulario menor.

5.1.- Estudio de las distribuciones de probabilidad del ML

Como se ha visto en el capítulo anterior para la tarea BDGEO, a pesar de la inclusión del modelo de lenguaje, se puede observar a la salida del decodificador que en el sistema sigue predominando la etapa acústica y el modelo se muestra incapaz de resolver situaciones como:

frase original	**que comunidad**
salida DAF	**kekomonia**
salida del sistema integrado (α=1)	**que como ni a**

A pesar de que las probabilidades entre las combinaciones de palabras “que como”, “como ni” y “ni a” deben ser pequeñas (probabilidades probablemente obtenidas a partir de la técnica de suavizado), frente a “que comunidad”, el sistema prefiere la primera disposición de

palabras. Esto significa que, por supuesto, la componente acústica cuenta con una mayor aportación a la probabilidad final acumulada y que también probablemente se les esté aplicando a los ML un suavizado excesivo que hace que sean equiparables o incluso más probables las situaciones no vistas en el entrenamiento del ML frente a las vistas. Como ya se ha presentado en el capítulo anterior esta situación se soluciona (descenso notable del error) decrementando las probabilidades del ML mediante la aplicación de una función exponencial con $\alpha>1$ que logre una distribución total de probabilidad más dispersa.

Una posible alternativa a la aplicación de esta función exponencial sobre las probabilidades del ML consiste en disminuir la probabilidad que los modelos *k*-EE suavizados asignan a los eventos no vistos en entrenamiento. En este caso, el descuento sólo se aplica sobre las probabilidades asociadas a las transiciones de los eventos no vistos en entrenamiento, lo que provoca una distribución de probabilidades del ML con mayor rango de dispersión.

Posteriormente se va a presentar un estudio sobre la importancia real del suavizado de los ML presentado resultados de decodificación con los modelos *k*-EE no suavizados y por tanto no integrados.

5.1.1.- Atenuación de las probabilidades de las transiciones al estado de back-off en el ML suavizado

En este punto se va a probar el efecto que produce sobre los resultados de perplejidad y las tasas de error, la aplicación de un parámetro heurístico que regule la atenuación de las probabilidades de las transiciones al estado de back-off en el ML suavizado. Esta disminución se ha realizado multiplicando la probabilidad de transición al estado de back-off (asociado con las situaciones no vistas en entrenamiento) por un factor de atenuación del back-off (*ab*). De esta manera se castigan las transiciones asociadas a las situaciones no vistas en el corpus de entrenamiento de los modelos *k*-EE. Para valores altos de *k* existen cada vez más situaciones no vistas, frente a las vistas, y para muchas palabras será necesario transitar varias veces por el estado de back-off hasta encontrar una situación vista en entrenamiento en un modelo de *k* inferior. Por tanto esta atenuación será mayor sobre todo para los valores más altos de *k*.

En la figura 5.1 se puede ver el efecto que el parámetro de atenuación de las probabilidades de las transiciones por back-off (*ab*) provoca en los resultados de perplejidad obtenidos con los distintos modelos integrados de orden K (K modelos *k*-EE con K=2, 3, 4, 5 y 6). Se muestran los resultados de perplejidad sobre el conjunto de evaluación de texto del ML (Prueba_Texto) y sobre las frases que realmente se van a reconocer (Prueba_Voz) de la tarea BDGEO (ver apéndice A).

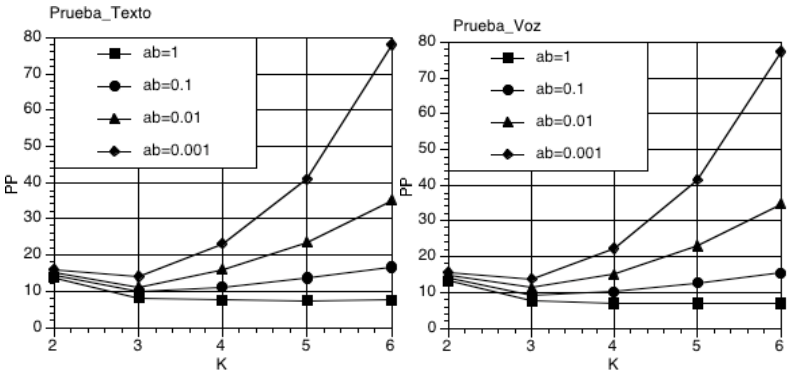


Figura 5.1.- Resultados de perplejidad obtenidos con los modelos integrados de orden K, K-EE (K=2...6) y distintos factores de atenuación de las probabilidades de transición al estado de back-off (*ab*) obtenidas con el descuento de Witten-Bell. Se utilizan los conjuntos de evaluación Prueba_Voz y Prueba_Texto.

Como se puede ver en la figura 5.1, ponderar la probabilidad asociada a las transiciones al estado de back-off con un factor *ab* incrementa notablemente los valores de la perplejidad sobre todo para valores altos de K. Este efecto es lógicamente mayor a medida que disminuye el valor del parámetro *ab*.

A continuación se muestran los resultados que se obtienen cuando los modelos integrados de orden K se integran en el sistema de RAH con distintos valores del parámetro de atenuación *ab* sobre las transiciones al estado de back-off. En la figura 5.2a se muestra el %PRE obtenido frente a K. En la figura 5.2b se puede ver también el Número Medio de Nodos Activos (NMNA) por frame para los distintos modelos integrados de orden K K=2...6, y en la figura 5.2c se muestra el %PRE frente al NMNA (base logarítmica).

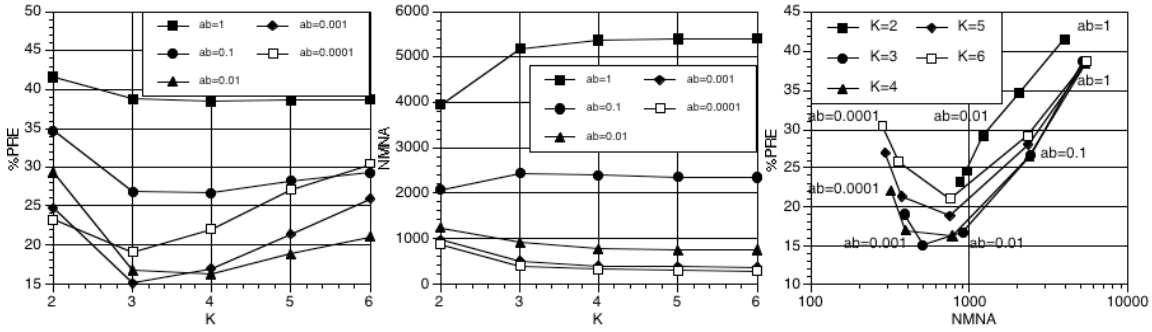


Figura 5.2.- a) Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) cuando se utilizan distintos factores de atenuación de las probabilidades asociadas al estado de back-off (ab). (Tarea BDGEO)). b) Número medio de nodos activos (NMNA) necesarios para la decodificación de las frases en función de K. c) %PRE obtenido en función del NMNA.

Para todos los valores de K (2...6) analizados la aplicación del parámetro de atenuación *ab* de las probabilidades asociadas a las transiciones al estado de back-off supone un decremento significativo del error acompañado de una disminución del número medio de nodos activos. La disminución del NMNA es debido a que la técnica de búsqueda en haz que se aplica en la decodificación utilizando el algoritmo de Viterbi (ver capítulo 4), necesita activar un menor número de posibles caminos a seguir, al haber cada vez más probabilidades (que han sido decrementadas) que no superan el umbral establecido. Con K=2 a medida que disminuye el

factor de atenuación *ab*, se obtienen cada vez mejores resultados. Sin embargo, para el resto de modelos (K >2) hay un mínimo alrededor de *ab*=0.01 y después los resultados se degradan, aunque siguen siendo notablemente mejores que los obtenidos sin la aplicación del parámetro *ab*, sobre todo para K=5 y K=6. Esto es lógico ya que esta alteración de las probabilidades de transición al estado de back-off en la forma en la que se han realizado, degrada mucho la distribución de probabilidad de los modelos con valores altos de K en los que cada vez habrá un mayor número de situaciones no vistas. Cada vez que se transita por el estado de back-off la probabilidad se multiplica por un factor de atenuación *ab*<1. Ésto implica que los caminos que tienen que transitar a través de varios estados de back-off tienen probabilidades excesivamente bajas. Por ejemplo para K=6 habrá bastantes situaciones en las que se multiplica hasta 5 veces, por un factor *ab* inferior a la unidad.

Parece que la perplejidad no mide bien la calidad del suavizado para el reconocimiento ya que las ponderaciones sobre el suavizado llevan a degradar notablemente los resultados de perplejidad obtenidos pero producen porcentajes de error inferiores.

El descuento utilizado para suavizar los modelos integrados de orden K (Witten-Bell) no parece ser el más adecuado desde el punto de vista del reconocimiento puesto que cualquier alteración sobre él (derivada de las alteraciones sobre las probabilidades de back-off o por aplicación de una función exponencial) nos lleva a una drástica reducción del número medio de nodos activos y de los porcentajes de error. No obstante, no hay que olvidar que no es lo mismo aplicar un descuento sobre las probabilidades asociadas a las transiciones de back-off, que aplicar una función exponencial sobre las probabilidades del ML (α>1). En el primer caso, sólo los eventos no vistos en entrenamiento ven alteradas sus probabilidades mientras que al aplicar la ponderación sobre las probabilidades del ML mediante la función exponencial (α>1), las probabilidades de todos los eventos vistos y no vistos se ven decrementadas. En este último caso puede que situaciones no vistas en entrenamiento, se mantengan con mayor probabilidad que algunas vistas, al obtener su probabilidad de un modelo inferior en el que esa transición fuera muy probable (ver capítulo 3).

A continuación vamos a evaluar (ver figura 5.3) la incidencia del reescalado de las probabilidades del ML cuando se aplica el factor de atenuación *ab* de las probabilidades de transición por el estado de back-off *ab*=0.01. En primer lugar se observa lo que ocurre cuando la integración con los modelos acústicos se realiza directamente utilizando la regla de Bayes, sin ponderación sobre las probabilidades del ML (α=1) (figura 5.3a). Como se había visto en la figura 5.2 hay un descenso muy notable del error y del número medio de nodos activos para todos los modelos integrados de orden K con K=2...6. En la figura 5.3b se muestran los resultados (%PRE y NMNA) obtenidos cuando se aplica un factor exponencial α sobre las probabilidades del ML. Se representan los valores obtenidos con *ab*=1 y *ab*=0.01 y los valores del parámetro α necesarios para conseguir el óptimo en cada caso.

Se observa que con *ab*=1 se alcanza un optimo de error para un valor de α mayor (α=6) frente al valor α =2 para *ab*=0.01, salvo para K=2 que se alcanza también para α=6. Para

cualquiera de los modelos integrados de orden K con K=2...6 las tasas de error son mejores (más bajas) cuando *ab*=1 y α=óptimo y se alcanzan con un menor número medio de nodos activos por frame lo cual implica un tiempo de decodificación menor.

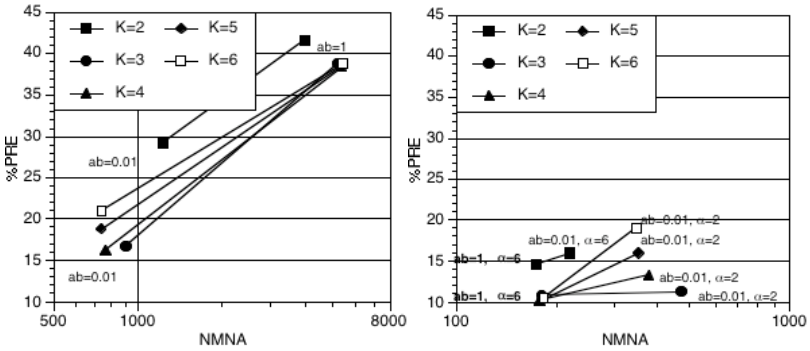


Figura 5.3.- Resultados de error obtenidos para los modelos integrados de orden K (K=2...6) con el parámetro de atenuación de las probabilidades de transición al estado de back-off *ab*=1 y *ab*=0.01. a) Resultados con α =1. b) Resultados con α = óptimo.

En resumen, para la tarea BDGEO cuando las probabilidades de transición por cada estado de back-off han sido alteradas por el valor del factor de atenuación de las probabilidades a través del estado de back-off *ab* se necesita un valor de α menor, α=2 para alcanzar el óptimo que es peor (sobre todo para valores altos de K) al obtenido sin la aplicación de ab y α=6. Para valores altos de K el modificar el factor *ab* sólo es útil, cuando el valor del parámetro α es pequeño y por tanto no se ha llegado a alcanzar el optimo, logrado para α=6.

Por supuesto, estos resultados indican que el reescalado directo de las probabilidades de transición asociadas al estado de back-off mediante la utilización de un parámetro de atenuación *ab* no es una buena estrategia de redistribución de la probabilidad (es muy drástica y toda la atenuación de probabilidades se realiza sobre los eventos no vistos) pero claramente pone de manifiesto que la distribución de probabilidad obtenida con el suavizado de Witten-Bell sobrestima la probabilidad asociada a los eventos vistos y por ello se necesita un valor del factor de ponderación de las probabilidades del ML α bastante alto (α=6) para alcanzar los resultados óptimos.

5.1.2 .- Importancia del suavizado

Antes de entrar en el estudio de diferentes métodos de descuento dentro de la técnica de suavizado del back-off sintáctico vamos a ver qué están aportando realmente las técnicas de suavizado. A continuación vamos a comparar los resultados que se obtienen cuando se aplica el suavizado de Witten-Bell (S) y cuando no se aplica suavizado (NS) sobre los modelos *k*-EE con *k*=2...6 en la tarea BDGEO (figura 5.4). Cuando no se aplica suavizado, los modelos no integran a los modelos inferiores. Se muestran los resultados que se obtienen cuando no se aplica reescalado sobre las probabilidades del ML (α=1) (figura 5.4a) y cuando se aplica el valor de reescalado necesario para conseguir el óptimo (figura 5.4b).

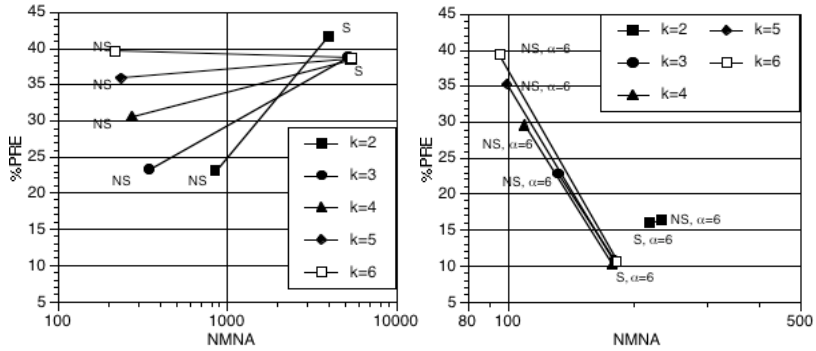


Figura 5.4.- Resultados de error obtenidos para la tarea BDGEO cuando se utilizan modelos integrados de orden K (K=2...6) suavizados con el descuento de Witten-Bell (S) y sin suavizar (NS). a)α=1, b) α=óptimo.

Cuando no se aplica el reescalado (α=1), se observan mejores resultados con los modelos sin suavizar (NS): disminución del %PRE salvo para *k*=6 y disminución del Número Medio de Nodos Activos (NMNA) por frame. Sin embargo cuando se trabaja con el valor del parámetro α que proporciona el óptimo (alrededor de 6) se ve que sólo para el modelo *k*=2 sin suavizar (no integrado) es posible llegar a resultados de error similares. A medida que se incrementa el valor de *k* la degradación en los modelos empieza a ser considerable. Lo cual pone de manifiesto algo que ya conocemos y es la necesidad de las técnicas de suavizado. No obstante, parece necesario lograr una distribución final de probabilidad más adecuada.

Quizá con otra distribución de probabilidades, utilizando otra estrategia de descuento, se necesite un valor de α distinto o se pueda llegar a tener un espectro de valores para los que el cambio en el error entre trabajar con α=1 y α=óptimo no fuera tan drástico. Esto nos lleva al estudio de otras técnicas de descuento dentro del suavizado sintáctico por back-off que se va a estudiar a lo largo del capítulo.

5.2.- Factores de descuento aplicables bajo el back-off sintáctico

Las técnicas de suavizado son necesarias dado que el conjunto de entrenamiento de los ML es limitado y no aparecen representadas todas las combinaciones de palabras. Como se ha presentado en el capítulo 2 básicamente existen dos posibilidades, el back-off o la interpolación. En el estudio desarrollado en (Ney et al, 1996) utilizando diferentes métodos de suavizado, se concluye que la diferencia entre ambas técnicas reside fundamentalmente en los detalles matemáticos proporcionando resultados muy similares en términos de perplejidad. En lo que se refiere a las técnicas de back-off se concluye que la técnica del descuento absoluto proporciona los mejores resultados y que los métodos de Katz que dejan las cuentas altas inalteradas no mejoran los resultados de perplejidad, si bien pueden simplificar el coste computacional del modelo. Todas las conclusiones se obtienen en función de los resultados de perplejidad que proporciona su sistema junto con los que proporciona el Toolkit de la CMU (Clarkson &

Robinson, 1997). En (Chen & Goodman, 1999) se puede encontrar también una revisión de las principales técnicas de suavizado.

Como ya se ha indicado, en ese trabajo se usa la técnica de descuento por back-off (ver capítulo 3). En este caso, los estados q del modelo k-EE a suavizar pertenecen al nivel k de modo que la etiqueta correspondiente al estado q representa una combinación de k-1 palabras[5] $w_{i-(k-1)}^{i-1}$ y el modelo suavizador será de nivel (k-1), de modo que los estados del modelo suavizador $w_{i-(k-1)-1}^{i-1}$ representan una combinación de k-2 palabras y serán estados de back-off (b_q) del modelo a suavizar (ver capítulo 3). Por tanto, si $q \equiv w_{i-(k-1)}^{i-1}$ es el estado correspondiente a un evento en el modelo k-EE y $b_q \equiv w_{i-(k-1)-1}^{i-1}$ es el estado correspondiente al mismo evento en el modelo (k-1)-EE. Entonces la probabilidad P(w/q) que se asocia a la transición $\delta^k(q,w)=(q_d$, P(w/q)) se obtiene a partir de (ver capítulo 2):

$$P(w/q) = \begin{cases} [1-\lambda]\dfrac{N(w/q)}{N(q)} & w \in \sum_q \\ \left(\displaystyle\sum_{\forall w_i \in \sum_q}\left[\lambda\frac{N(w_i/q)}{N(q)}\right]\right)\dfrac{P(w/b_q)}{1-\displaystyle\sum_{\forall w_i \in \sum_q} P(w_i/b_q)} & w \in \sum - \sum_q \end{cases} \quad (5.1)$$

donde Σ es el vocabulario de la tarea y Σ_q es el vocabulario asociado al estado q, es decir, el conjunto de palabras que aparecen en el corpus de entrenamiento detrás de la secuencia de palabras que etiqueta al estado q, $N(w/q)$ es el número de veces que la palabra w aparece después de la cadena que etiqueta al estado q, $N(q) = \sum_{\forall q \in \sum_q} N(w/q)$.

$P(w/ b_q)$ es la probabilidad estimada asociada al mismo evento en el modelo (k-1)-EE y (1-λ) es el factor de descuento que se debe aplicar sobre las probabilidades de los eventos vistos $w \in \sum_q$ para posteriormente ser repartido entre los eventos no vistos $w \in \sum - \sum_q$. El modelo (k-1)-EE, del que se tiene en cuenta el estado de back-off b_q, debe asignar probabilidad a todas las unidades, bien por ser un modelo base con cobertura del monoide libre o por haber sido suavizado previamente. Como se debe cumplir la condición estocástica:

$$\sum_{\forall w \in \sum} P(w/q) = 1 \quad \forall q \in Q^k \quad (5.2)$$

la probabilidad que se a asocia a cada evento no visto en entrenamiento $w \in \sum - \sum_q$ se estima a partir de:

$$P(w/q) = P(b_q/q)P(w/b_q) \quad \forall w \in \left(\sum - \sum_q\right) \quad (5.3)$$

lo que significa que la probabilidad que se asigna desde cada estado q a la transición a su estado de back-off b_q, $P(b_q/q)$, se estima de la siguiente forma:

---

[5] Una combinación de k-1 palabras $w_{i-(k-1)}...w_{i-1}$ viene dada como referencia a una posición concreta "i" dentro de la secuencia completa $w_1..w_i..w_L$.

$$P(b_q / q) = \left(\sum_{\forall w_i \in \sum_q} \left[\lambda \frac{N(w_i / q)}{N(q)} \right] \right) \frac{1}{1 - \sum_{\forall w_i \in \sum_q} P(w_i / b_q)} \qquad (5.4)$$

El factor $P(b_q / q)$ es la relación entre la probabilidad total no vista en q (el descuento) y la probabilidad total que el modelo inferior asigna a los eventos no vistos en q. La probabilidad que el modelo inferior asigna a cada uno de estos eventos participa también junto con $P(b_q / q)$ en el cálculo de la probabilidad suavizada. Esto presupone que si en el autómata de grado k no se encuentra el estado q correspondiente al evento analizado w_i tras $w_{i-(k-1)}^{i-1}$, lo que sucederá en aquellos casos en los que tal cadena no se haya dado en el conjunto de entrenamiento se le asignará al evento la probabilidad dada por el modelo inferior, el de k-1.

Dentro de las técnicas de descuento por back-off el factor de descuento (1-λ) tiene diferentes formas de expresarse lo que da lugar a descuentos muy distintos (Ney et al, 1997). Principalmente se puede distinguir entre los métodos en los que se aplica el descuento, en cada estado, sobre todos los eventos vistos y los que aplican el descuento sobre el conjunto de eventos vistos un número reducido de veces porque se supone que sus distribuciones de probabilidad estarán peor estimadas.

A continuación se presenta una clasificación de los métodos de descuento que se va a estudiar a lo largo del capítulo. Existen métodos clásicos como el descuento de Absoluto, Lineal y nuevas propuestas como el descuento Sencillo y Acotado.

i) Métodos basados en que el descuento se realiza sobre todos los eventos vistos.

a) Independientes de la evaluación de parámetros:

* *Descuento de Witten-Bell* (Bordel, 1996) (Clarkson & Rosenfeld, 1997):

- El descuento depende del número de eventos distintos vistos en cada estado.

- Ha sido el método utilizado como base de la experimentación previa.

* *Descuento Simple:*

- Propuesta aplicada en el toolkit de la CMU (Clarkson & Rosenfeld, 1997) como solución a aplicar junto con el descuento de Katz (ver subsección 5.2.2.1) cuando todos los eventos en un estado han sido vistas un número alto de veces.

b) Dependientes de evaluación de parámetros:

* *Descuento absoluto*:

- Método clásico de descuento.

- Tradicionalmente el descuento absoluto proporciona muy buenos resultados aunque la mayoría de las evaluaciones se dan en perplejidad (Ney et al, 1996).

* *Descuento lineal:*

- Método clásico de descuento.

- Tradicionalmente proporciona peores resultados que el descuento absoluto (Ney et al, 1996).

ii) Métodos basados en que el descuento se realiza sobre los eventos vistos peor estimados.

El descuento se realiza sólo sobre los eventos vistos menos de un determinado número de veces porque son las distribuciones de probabilidad que peor estimadas están. Por ello es necesario establecer el valor r que indica el número mínimo de veces que un evento debe ser visto para que su probabilidad no sufra ninguna atenuación.

* *Descuento de Katz* (Katz, 1987):

- Método clásico

- El desarrollo exige restricciones sobre las distribuciones de los eventos. En general estas restricciones no se cumplen para valores altos de *k*.

* *Descuento Acotado:*

- Nueva propuesta.

- Se puede desarrollar para cualquier valor de *k*.

- Depende de los valores que tome un parámetro.

A continuación se va a presentar una descripción de cada uno de ellos.

5.2.1.- Descuento sobre todos los eventos vistos

En primer lugar se van a presentar los descuentos que no son dependientes de parámetros: descuento de Witten-Bell (base de la experimentación previa) y descuento Sencillo. Después se presentan los descuentos dependientes de parámetros: descuento absoluto y lineal. Todos los descuentos se formulan utilizando la técnica de back-off bajo el formalismo sintáctico.

5.2.1.1.- Descuento de Witten-Bell

Esta aproximación aparece referenciada como back-off sintáctico en trabajos previos (Bordel, 1994) y coincide con el esquema de descuento de Witten-Bell que aparece referenciada en (Clarkson & Rosenfeld, 1997). Por esta razón recibe ese nombre para diferenciarlo del resto de suavizados que también se desarrollan bajo el formalismo sintáctico.

En el descuento de Witten-Bell, el factor de descuento (1-λ) de la ecuación 5.1 se aplica sobre las probabilidades de todos los eventos vistos de la siguiente forma:

$$1 - \lambda = \frac{N(q)}{N(q) + \left|\sum_q\right|} \qquad (5.5)$$

donde $|\Sigma_q|$ es el número de eventos distintos que siguen a un contexto particular en el estado q y $N(q)$ es el número de eventos totales vistos en el estado q, $N(q) = \sum_{\forall w \in \Sigma_q} N(w/q)$.

Siendo N(w/q) el número de veces que aparece la palabra w después de la cadena que etiqueta el estado q. En este caso el factor (1-λ) no es dependiente de N(w/q) y las

probabilidades de las cuentas más altas se ven menos alteradas, lo cual es lógico porque se supone que están mejor estimadas.

Por tanto, la probabilidad $P(w_i/q)$ asociada a la $\delta^k(q, w) = (q_d, P(w_i/q))$ se estima de acuerdo a la siguiente ecuación:

$$P(w/q) = \begin{cases} \dfrac{N(w/q)}{N(q)+\left|\sum_q\right|} & w \in \sum_q \\ \dfrac{\left|\sum_q\right|}{N(q)+\left|\sum_q\right|} \dfrac{P(w/b_q)}{1-\sum\limits_{\forall w_i \in \sum_q} P(w_i/b_q)} & w \in \left(\sum - \sum_q\right) \end{cases} \tag{5.6}$$

donde P(w/b_q) es la probabilidad estimada asociada al mismo evento w en el modelo (k-1)-EE.

5.2.1.2.- Descuento Sencillo

En este caso la cuenta N(q) se ha incrementado en una unidad para todo estado q. Como consecuencia el coeficiente de descuento (1-λ) que se aplica no depende tampoco de N(w/q) y se calcula como:

$$1-\lambda = \frac{N(q)}{N(q)+1} \tag{5.7}$$

Por lo que la probabilidad P(w/q) resultante se calcula como:

$$P(w/q) = \begin{cases} \dfrac{N(w/q)}{N(q)+1} & w \in \sum_q \\ \dfrac{1}{N(q)+1} \dfrac{P(w/b_q)}{1-\sum\limits_{\forall w_i \in \sum_q} P(w_i/b_q)} & w \in \left(\sum - \sum_q\right) \end{cases} \tag{5.8}$$

En este caso también las cuentas más bajas sufren una reducción de la probabilidad mayor. En cualquier caso todas las probabilidades sufren menos alteración que con el descuento de Witten-Bell, es decir, esta técnica otorga claramente menos probabilidad a repartir entre los eventos no vistos ya que:

$$\frac{1}{N(q)+1} \leq \frac{\left|\sum_q\right|}{N(q)+\left|\sum_q\right|} \tag{5.9}$$

Esta aproximación se basa en la aproximación que aplica el toolkit de la CMU cuando trabaja con el descuento de Katz para evitar el problema que se crea cuando todos los eventos son observados más de r veces en un estado (ver subsección 5.2.2.1). Con esta técnica se propone la extensión de esta misma idea a todos los estados del autómata.

5.2.1.3.- Descuento Absoluto

La idea básica que subyace es la dejar también las cuentas mas altas virtualmente inalteradas. Se introduce para todas las cuentas un factor que es independiente del número de elementos que se hayan visto. La aproximación consiste en restar una constante *b* de cada una de las cuentas de la siguiente forma:

$$1-\lambda = \frac{N(w/q)-b}{N(w/q)} \tag{5.10}$$

El descuento depende del número de veces que se observe la palabra *w* en el estado q N(*w*/*q*), por lo que la probabilidad P(*w*/*q*) se calcula como:

$$P(w/q) = \begin{cases} \dfrac{N(w/q)-b}{N(q)} & w \in \sum_q \\ \dfrac{b\left|\sum_q\right|}{N(q)} \dfrac{P(w/b_q)}{1-\sum\limits_{\forall w_i \in \sum_q} P(w_i/b_q)} & w \in \left(\sum - \sum_q\right) \end{cases} \tag{5.11}$$

5.2.1.4.- Descuento Lineal

Es el modo mas sencillo de resolver el problema de la asignación de la probabilidad a eventos no vistos en un estado. La aproximación consiste en aplicar un factor de descuento de la siguiente forma:

$$1-\lambda = 1-l \text{ con } 0<l<1 \tag{5.12}$$

por lo que la probabilidad P(*w*/*q*) se obtiene a partir de la siguiente ecuación:

$$P(w/q) = \begin{cases} (1-l)\dfrac{N(w/q)}{N(q)} & w \in \sum_q \\ l\dfrac{P(w/b_q)}{1-\sum\limits_{\forall w_i \in \sum_q} P(w_i/b_q)} & w \in \left(\sum - \sum_q\right) \end{cases} \tag{5.13}$$

En los experimentos reflejados en la literatura (Ney et al, 1996) y en experiencia propia muestran que este modelo no proporciona buenos resultados de perplejidad. Con este método de descuento tanto las cuentas más altas como las más bajas se reducen por el mismo factor (1-λ). Esto implica que hay una reducción importante de las probabilidades que en principio están bien estimadas.

5.2.2.- Descuento sobre los eventos peor estimados

Se presentan dos métodos, un método clásico que presenta restricciones sobre la distribución de los eventos (Descuento de Katz) y una nueva propuesta de este trabajo que eliminará las restricciones anteriores (Descuento Acotado).

5.2.2.1.- Descuento de Katz

Este método esta basado en la fórmula de Turing. La idea principal es la de suponer que las cuentas más altas están mejor estimadas y que por lo tanto se les aplica un descuento menor. A medida que disminuye la cuenta se incrementa el descuento de manera que los eventos vistos una sola vez son los peor estimados y se les aplica el descuento máximo.

En la formula original el factor de descuento (1-λ) depende de N(*w*/q) de la siguiente manera:

$$1-\lambda = \left[N(w/q)+1\right]\frac{n_{N(w/q)+1}}{n_{N(w/q)}} \tag{5.14}$$

donde n_i es el número de eventos que se han visto exactamente i veces. Por tanto $n_{N(w/q)+1}$ es el número de eventos distintos que se han visto tantas veces como las palabras *w* en el estado q y $n_{N(w/q)}$ es el número de eventos totales que se han visto en el estado q.

Según el descuento de Turing la cantidad de probabilidad asignada a eventos no vistos es igual al número de eventos distintos que se han visto una sola vez n_1 dividido entre el número total de eventos que se han visto en un estado q, N(q):

$$1-\sum_{\forall w_i \in \sum_q} P(w_i/q) = \frac{n_1}{N(q)} \tag{5.15}$$

Esta aproximación presenta algunas desventajas. En primer lugar, pone restricciones a los valores relativos de n_1, n_2,...,n_{r+1}, que no son siempre superados en todos los estados, por modelos con valores medios y altos de *k*. En la práctica estos estados son bastante frecuentes. Por grande que sea el número de muestras utilizadas las cuentas mayores se observan un número menor de veces que las menores. Katz eliminó este problema haciendo que para las cuentas más altas quedaran inalteradas y que por tanto no fuera necesario la aplicación de la formula de Turing sobre las probabilidades asociadas a las cuentas más altas. Sólo las cuentas inferiores a un valor *r* se decrementan (Katz, 1987) por lo que la probabilidad P(*w*/*q*) se obtiene de la siguiente forma:

$$P(w/q) = \begin{cases} \dfrac{N(w/q)}{N(q)} & w \in \sum_q \wedge N(w/q) > r \\ \left[1-\lambda\right]\dfrac{N(w/q)}{N(q)} & w \in \sum_q \wedge 1 \le N(w/q) \le r \\ \displaystyle\sum_{\substack{\forall w_i \in \sum_q \\ 1 \le N(w_i/q) \le r}} \left[\lambda \frac{N(w_i/q)}{N(q)}\right] \frac{P(w/b_q)}{1-\displaystyle\sum_{\forall w_i \in \sum_q} P(w_i/b_q)} & w \in \left(\sum - \sum_q\right) \end{cases} \tag{5.16}$$

Katz propuso la aplicación de un factor de descuento (1-λ) de forma que cuanto menor fuera el valor de la cuenta sufriera un descuento mayor, de la siguiente forma:

$$1-\lambda = \frac{\frac{(N(w/q)+1)n_{N(w/q)+1}}{N(w/q)n_{N(w/q)}} - \frac{(r+1)n_{r+1}}{n_1}}{1-\frac{(r+1)n_{r+1}}{n_1}} \qquad (5.17)$$

donde r y n_r tienen el mismo significado en la ecuación 5.16. A pesar de la introducción de esta nueva propuesta, cuando se trabaja con valores altos de *k* hay ocasiones en las que $n_{N(w/q)+1} \leq n_{N(w/q)}$ siendo $n_{N(w/q)}$<r. Si se dan estas condiciones, el descuento supone un aumento de la probabilidad para los eventos vistos en vez de un decremento. Además es necesario el uso de chequeos adicionales para los estados en los que todos los eventos son observados mas de r veces. Es decir que hacer cuando ocurre que N(*w*/q)>r para toda *w* visto a partir de q.

En el sistema desarrollado en el Toolkit de la CMU (Clarkson & Rosenfeld, 1997), la estrategia consiste en sumar 1 a N(q) en los estados en los se presenta el problema de forma que se quita una cierta probabilidad de todos los estados. La probabilidad alterada N(q)+1 se usa para ser distribuida entre los elementos no vistos. Lógicamente las probabilidades de los eventos que tengan cuentas más altas se verán menos alteradas. Esto en cierto modo contradice la idea asociada al descuento de Katz, que propone dejar las cuentas más altas totalmente inalteradas.

5.2.2.2.- Descuento Acotado

La distribución de ocurrencia de los eventos ($n_1..n_{r+1}$) no es adecuada sobre todo para valores altos de *k*, por tanto no es posible la aplicación directa del descuento de Turing. Teniendo en cuanta el problema se ha desarrollado un nuevo método de descuento que en filosofía es totalmente similar al método de Katz. De hecho, esta aproximación trata de simular totalmente el descuento de Katz pero evitando los problemas derivados de las restricciones asociadas a los valores n_1, n_2,...,n_{r+1},. Como en el modelo de Katz la operación de descuento se limita a las cuentas más bajas, es decir, $N(w/q) \leq r$ de acuerdo con la ecuación 5.6.

Cuentas mayores a r se estiman por el criterio de Máxima Verosimilitud y al resto se les aplica un descuento de la siguiente manera:

$$1-\lambda = d - \tau(r - N(w/q)) \quad \tau, d < 1, \tau <<< d. \qquad (5.18)$$

Por tanto sólo se descuenta sobre las cuentas $N(w/q) \leq r$ y el descuento es mayor para las cuentas más pequeñas. El descuento final depende de los valores de los parámetros *d* y τ que deben ser inferiores a la unidad. Cuanto mayor es $N(w/q) \leq r$ se aplica un descuento mayor. Por ejemplo cuando N(*w*/q)=r el descuento es el mínimo y sólo depende del parámetro *d*, y cuando N(*w*/q)=1, el descuento aplicado es el máximo *(d-*τ*(r-1))*.

Para solucionar el problema derivado de aquellos estados para los cuales todos los eventos se han visto más de r veces hay que hacer comprobaciones adicionales. Como alternativa

al método propuesto por la CMU (Clarkson & Robinson, 1997) se propone un nuevo método en el que, sólo las cuentas más pequeñas fueron decrementadas de la siguiente forma:

$$P(w/q)=\begin{cases}\dfrac{N(w/q)}{N(q)} & w\in\sum_q \wedge\ N(w/q)>min(N(w/q))\\ d\dfrac{N(w/q)}{N(q)} & w\in\sum_q \wedge\ N(w/q)=min(N(w/q))\\ \left(\displaystyle\sum_{\substack{\forall w_i\in\sum_q\\ 1\le N(w_i/q)\le r}}\left[[1-d]\frac{N(w_i/q)}{N(q)}\right]\right)\dfrac{P(w_i/b_q)}{1-\displaystyle\sum_{\forall w_i\in\sum_q}P(w_i/b_q)} & w\in\left(\sum-\sum_q\right)\end{cases} \quad (5.19)$$

Donde d es el parámetro del que depende el descuento, es decir $(1-\lambda)=d$ sólo afecta a los eventos que se han observado el menor número de veces.

Se utiliza esta aproximación ya que parece más lógica que la propuesta presentada por la CMU que supone una redistribución de todas las probabilidades cuando se está trabajando con un método que sólo aplica el descuento sobre las cuentas más bajas que un cierto valor r.

5.3.- Resultados experimentales

En este punto se van a evaluar los diferentes descuentos propuestos en la sección anterior a excepción del descuento de Katz que como ya se ha explicado anteriormente no se puede aplicar directamente debido a la distribución no adecuada de las muestras. El resto de los descuentos propuestos (Witten-Bell, Sencillo, absoluto, lineal y descuento acotado) se han evaluado tanto en perplejidad como en %PRE. Los descuentos Absoluto, Lineal y Acotado dependen del valor de distintos parámetros, en estos casos se ha realizado un barrido experimental sobre el valor del parámetro correspondiente para lograr el óptimo de error con la base de datos bajo estudio (BDGEO).

Todos los descuentos estudiados: Sencillo, absoluto, lineal y el acotado fueron evaluados tomando como referencia el descuento de Witten-Bell (Bordel et al, 1994), (Bordel et al, 1997), (Varona & Torres, 1999a), (Torres & Varona, 1999). El estudio se realiza primero con los métodos que realizan el descuento para todos los eventos vistos en entrenamiento (Sencillo, Absoluto, y Lineal) y posteriormente se evalúa la nueva propuesta: el descuento Acotado. En primer lugar, se presentan los resultados en perplejidad y %PRE sin reescalado de las probabilidades del ML y a continuación se ha estudiado el efecto de la aplicación del factor de escalado α sobre a las probabilidades del ML $(P(W))^{\alpha}$.

5.3.1.- Resultados sin reescalado de las probabilidades del ML

En este caso se va a realizar el estudio de los diferentes descuentos evaluados en perplejidad y en %PRE cuando los modelos se integran en el sistema de RAH y no hay

reescalado exponencial de las probabilidades del ML. Todos los resultados se van a obtener utilizando un elemento comparativo que es el descuento de Witten-Bell.

5.3.1.1.- Descuento Sencillo.

En ambas técnicas de descuento: descuento de Witten-Bell (dWB) y descuento Sencillo (dS) la masa de probabilidad que se asigna a los elementos no vistos en entrenamiento no depende de ningún factor que deba ser obtenido experimentalmente (ecuaciones 5.5 y 5.7).

Se puede ver en la figura 5.5 que los resultados que se obtienen en perplejidad cuando se aplica el descuento Sencillo son muy parecidos a los resultados obtenidos con el descuento de Witten-Bell. Se muestran los resultados que se obtienen con los modelos integrados de orden K (K modelos *k*-EE) con K=2...6, para el corpus de evaluación de texto (Prueba_Texto) y sobre el conjunto de evaluación de voz (Prueba_Voz) con el que se han obtenido los resultados de decodificación.

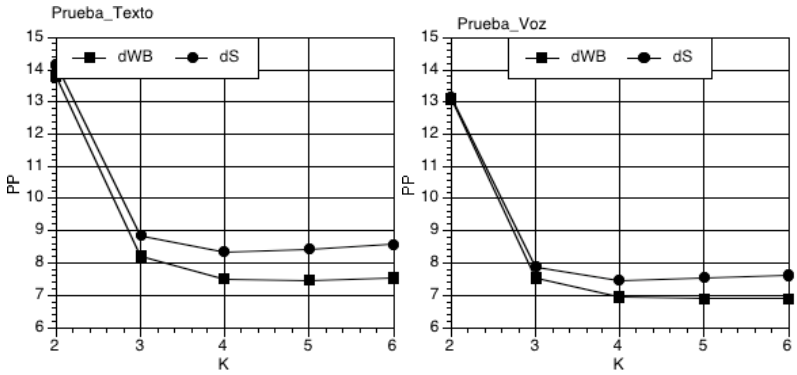


Figura 5.5.- Resultados de perplejidad obtenidos con los modelos integrados de orden K (K=2...6) suavizados utilizando el descuento de Witten-Bell y descuento Sencillo sobre los conjuntos Prueba_Texto y Prueba_Voz.

En ambos casos se obtienen valores de perplejidad casi constantes para valores de K superiores a 3. Sin embargo los valores de perplejidad que se obtienen con el descuento Witten-Bell son inferiores (mejores) para todos los valores de K. La masa de probabilidad que se asigna a los elementos no vistos en entrenamiento es inferior cuando se aplica el descuento Sencillo. En la mayor parte de los estados la normalización se realiza con una cantidad inferior $1<|\Sigma q|$ (ver ecuación 5.9). Esto implica un menor suavizado que se refleja sobre todo para valores altos de K. Como consecuencia la distribución de probabilidad en el modelo final resultante con el descuento Sencillo está menos suavizada (mayor rango de distribución de las probabilidades) y este hecho se refleja en un aumento de la perplejidad.

A continuación se pueden ver los resultados de %PRE obtenidos (ver figura 5.6) cuando se integran directamente las probabilidades acústicas y de ML en el esquema de decodificación utilizando la regla de Bayes. Como se había visto en el capítulo anterior, cuando las probabilidades del ML se aplican directamente no son lo suficientemente significativas para tomar decisiones en el desarrollo del trellis de Viterbi. Sin embargo, los modelos con distribuciones de probabilidad menos suavizadas (descuento Sencillo) presentan un menor

%PRE acompañado de una disminución del número medio de nodos activos (NMNA). Con el descuento Sencillo los modelos presentan un rango de dispersión de las probabilidades mayor y como consecuencia las probabilidades del ML tienen una influencia mayor en la sentencia final reconocida por el sistema.

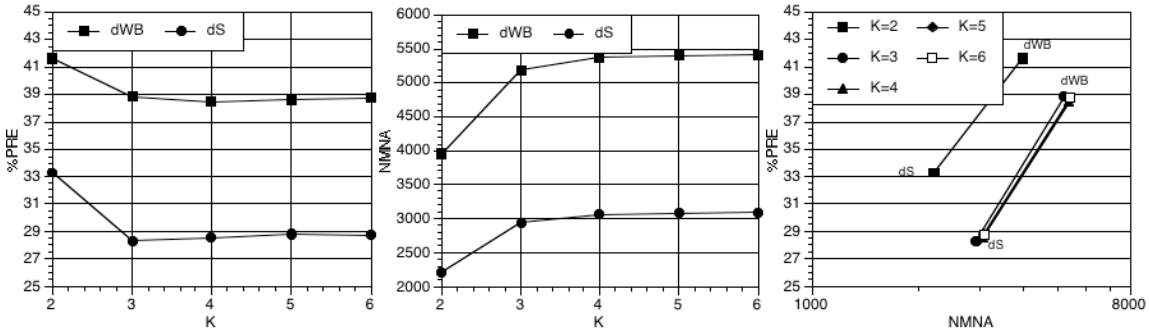


Figura 5.6.- a) Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) utilizando el descuento Sencillo y descuento de Witten-Bell. b) Representa los resultados del número medio de nodos activos (NMNA) necesarios para la decodificación de las frases en función de K. c) %PRE obtenido en función del NMNA.

A continuación vamos a estudiar en detalle la relación existente entre los valores de perplejidad y las tasas de error (%PRE) (figura 5.7). Se puede apreciar que un aumento de la perplejidad implica una disminución del error, por lo que la perplejidad no parece ser la medida más adecuada para predecir el comportamiento de los ML en el sistema de RAH.

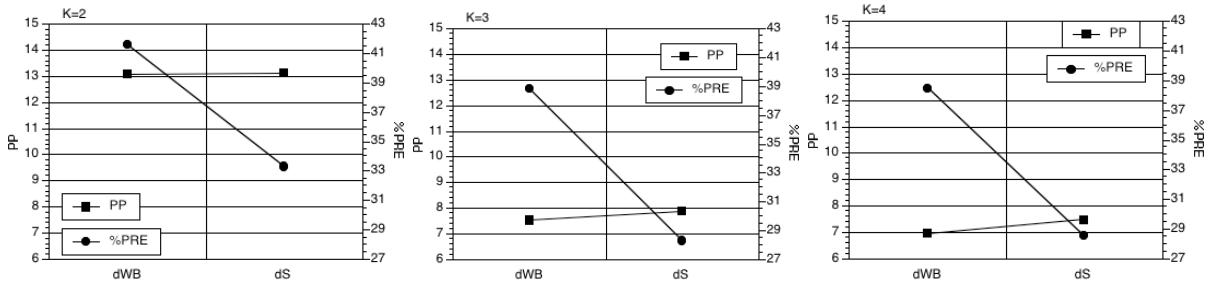


Figura 5.7.- Resultados de perplejidad (PP) y %PRE obtenidos para los modelos integrados de orden K (K=2...4) con los descuentos de Witten-Bell y descuento Sencillo.

5.3.1.2.- Descuento Absoluto

Con ambos métodos (descuento de Witten-Bell (dWB) y descuento absoluto (dAb)) las cuentas más altas se dejan prácticamente inalteradas, si bien este proceso se hace de dos formas completamente distintas. Se han evaluado diferentes valores del parámetro b que regula el suavizado de la técnica del descuento absoluto (ver ecuación 5.10). Cuanto más pequeño es el valor del parámetro b, el descuento que se aplica sobre los eventos vistos es menor, se distribuye una menor masa de probabilidad entre los eventos no vistos y como consecuencia se obtienen modelos menos suavizados. En la figura 5.8 se presentan los resultados de perplejidad obtenidos cuando se evalúan distintos valores del parámetro b para el conjunto de frases que serán posteriormente evaluadas en reconocimiento (Prueba_Voz). No se presentan los

resultados para el conjunto Prueba_Texto por ser redundantes. En la figura 5.8b se adjuntan también los resultados en detalle para notar las pequeñas diferencias existentes entre aquellos valores del parámetro b que presentan valores más similares (alrededor de 7).

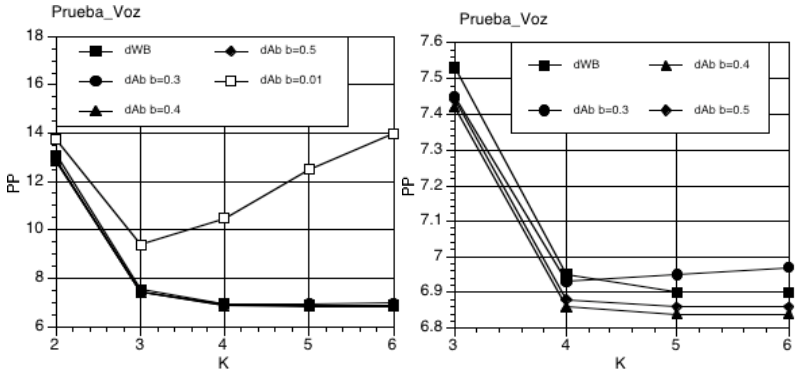


Figura 5.8.- Resultados de perplejidad (PP) obtenidos para los modelos integrados de orden K (K=2...6) suavizados con el descuento de Witten-Bell y descuento Absoluto. b) PP en detalle para aquellos descuentos que dan resultados más parecidos.

Los resultados muestran que cuánto mayor masa de probabilidad (b>>) se les asigna a los eventos no vistos en el entrenamiento (ver ecuación 5.10), hay un mayor suavizado y por tanto se obtienen mejores resultados de perplejidad (hasta un máximo b=0.4). A partir de b=0.4 la perplejidad aumenta, esto significa que el suavizado puede distorsionarse por ser incluso más probables los eventos no vistos en entrenamiento. En general cuanto más suavizada esté la distribución de probabilidades (menor rango de dispersión de las probabilidades) se obtienen mejores resultados de perplejidad (más bajos).

A continuación se muestran los resultados de error obtenidos. Se pueden ver en la figura 5.9, los resultados del %PRE y del número medio de nodos activos (NMNA) para los distintos descuentos.

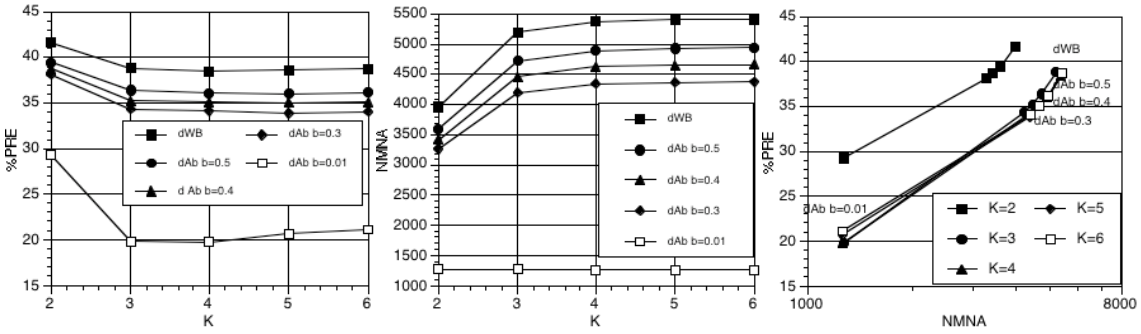


Figura 5.9.- a) Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) utilizando el descuento Absoluto (distintos valores del parámetro b) y descuento de Witten-Bell. b) Número medio de nodos activos (NMNA) necesarios para la decodificación de las frases en función de K. c) %PRE obtenido en función del NMNA.

Los resultados muestran que cuanto mayor masa de probabilidad (>>b) se les asigna a los elementos no vistos se obtuvieron mejores %PRE más altos junto con un mayor número de NMNA independientemente del valor de K. El descuento de Witten-Bell es el que peores resultados proporciona. Parece que los modelos con descuento de Witten-Bell son los que más

suavizados están (distribución con la menor dispersión de las probabilidades). En los resultados de perplejidad hay modelos con descuento Absoluto que presentaban mejores resultados de perplejidad (descuento Absoluto b=0.5 y b=0.4) que el descuento de Witten-Bell, si bien las diferencias no son estadísticamente significativas.

Con el descuento Absoluto b=0.01 se obtiene el mejor y mas significativo resultado del %PRE. En este caso la distribución de probabilidades entre los eventos vistos y no vistos está muy distribuida por lo que la técnica de búsqueda en haz actúa más para reducir el número medio de nodos activos y se logra el mejor resultado de error que se mantiene para todos los valores de K, a pesar de que los resultados de perplejidad obtenidos predecían un mal comportamiento de ese descuento sobre todo para valores altos de K. Parece que la perplejidad no fue capaz de predecir el comportamiento de los modelos ni siquiera frente a los diferentes valores de K utilizando un mismo descuento, en donde sí que se aprecian diferencias significativas en los valores de la perplejidad pero no en las tasas de error.

5.3.1.3.- Descuento Lineal

En este caso se han evaluado diferentes valores del parámetro l que regula el suavizado con la técnica de descuento Lineal (dL). Cuanto más alto es el valor de l se consigue un mayor suavizado (ver ecuación 5.12).

Con esta técnica de suavizado, se descuenta por igual sobre las probabilidades de los eventos vistos muchas veces (bien estimados) y los vistos pocas veces. De modo que las cuenta más altas también se ven seriamente alteradas. En primer lugar, vamos a ver los resultados de perplejidad obtenidos (ver figura 5.10) para el conjunto de frases que serán evaluadas en el sistema integrado de decodificación (Prueba_Voz).

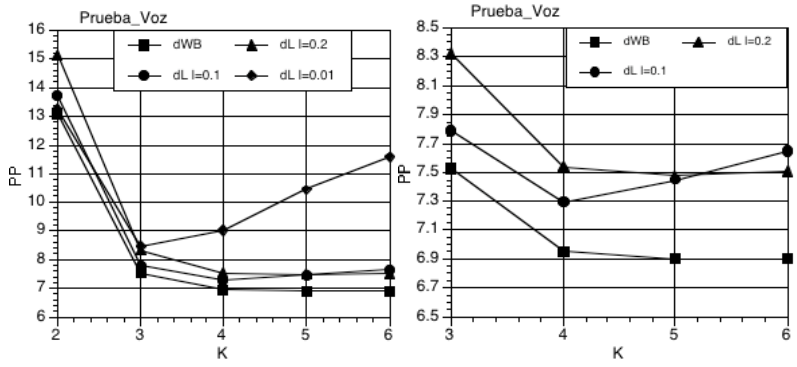


Figura 5.10.- Resultados de perplejidad (PP) obtenidos para los modelos integrados de orden K (K=2...6) suavizados con el descuento de Witten-Bell y descuento Lineal. b) PP en detalle para aquellos descuentos que dan resultados más parecidos.

Se observa que en todos los casos, los resultados son peores respecto al descuento de Witten-Bell. Con descuento Lineal l=0.1 la perplejidad tiende a aumentar para valores altos de K. Es el mismo efecto que se observa para el descuento Lineal l=0.01 pero mucho mas atenuado. Con descuento Lineal l= 0.2 el modelo está más suavizado y la perplejidad se mantiene para valores altos de K igual que con el descuento de Witten-Bell.

A continuación se muestran los resultados de error (figura 5.11). A medida que el valor del parámetro l es menor se distribuye menos probabilidad entre los eventos no vistos (ver ecuación 5.12). Con l=0.01 prácticamente no hay suavizado y por tanto a medida que aumenta el valor de K la perplejidad se incrementa mucho. Esto como ya sabemos por los experimentos previos significa un buen comportamiento de error (bajo %PRE y NMNA) cuando los modelos acústicos y de lenguaje se integran directamente en el sistema integrado.

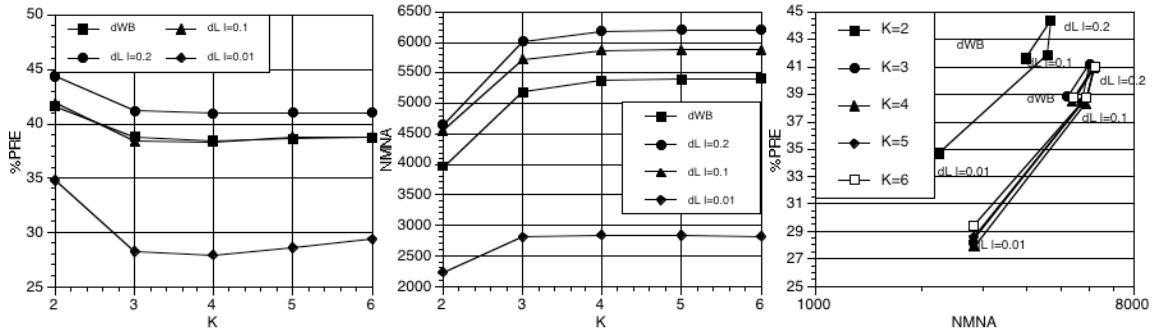


Figura 5.11.- a) Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) utilizando el descuento Lineal (distintos valores del parámetro l) y descuento de Witten-Bell. b) Número medio de nodos activos (NMNA) necesarios para la decodificación de las frases en función de K. c) %PRE obtenido en función del NMNA.

En general, los resultados muestran que cuanto mayor masa de probabilidad (>>l, (ver ecuación 5.12)) se asigna a los elementos no vistos (hasta un máximo l=0.1), se observan valores de error más altos junto con un mayor número medio de nodos activos por frame (NMNA), lo que implica tiempos de decodificación más altos.

5.3.1.4.- Descuento Acotado

Como ya se ha dicho, el descuento Acotado (dAc) está basado en el descuento de Katz, y supone dejar las cuentas más altas inalteradas. Los eventos observados más de r veces (r=7) se estiman por el criterio de Máxima Verosimilitud y al resto se les aplica un descuento de la forma presentada en la ecuación 5.16. Este descuento es mayor cuanto más baja sea el valor de la cuenta $N(w/q) \leq r$. Si todas las cuentas superan el umbral, el descuento sólo se realiza sobre los eventos de cuentas más bajas en cada estado (ver ecuación 5.19). En general, el descuento final depende de los valores de los parámetros d y τ. Se han probado distintos valores del parámetro d con un valor del parámetro τ fijo τ=0.01 (ver ecuación 5.18).

Cuanto mas bajo es el valor de d, se obtiene un mayor suavizado. La figura 5.12 muestra los resultados de perplejidad obtenidos junto con el detalle para aquellos valores de d con los que se obtienen resultados de perplejidad más similares. Los resultados muestran que cuanta más masa de probabilidad (<<d) se asigne a los eventos no vistos (hasta un máximo con d=0.70), mejores resultados de perplejidad se obtienen.

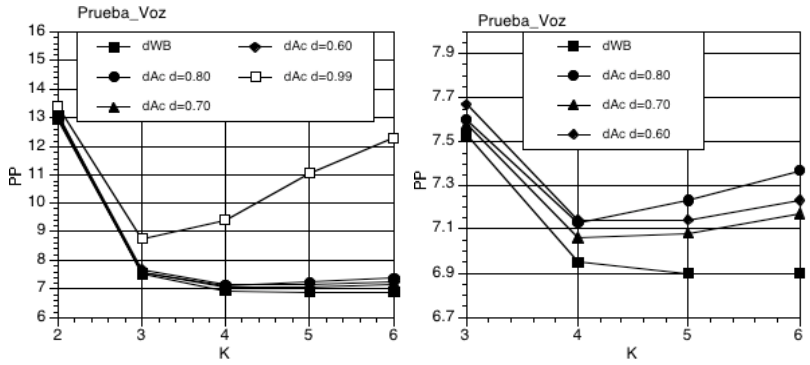


Figura 5.12.- Resultados de perplejidad (PP) obtenidos con los modelos integrados de orden K (K=2...6) suavizados utilizando el descuento de Witten-Bell y descuento Acotado. b) PP en detalle para aquellos descuentos que proporcionan resultados más parecidos.

Para todos los valores de K, el descuento de Witten-Bell presenta los mejores resultados de perplejidad (ver figura 5.12b)) y se mantienen constantes con K. Sin embargo, para el descuento Acotado para todos los valores de d, el valor de la perplejidad tiende a aumentar para modelos con valores altos de K. Si la perplejidad aumenta mucho para valores altos de K significa que se obtienen modelos menos suavizados.

La figura 5.13 muestra el %PRE que se obtiene junto con el número medio de nodos activos para los diferentes modelos integrados de orden K (K modelos *k*-EE con K=2... 6). Los resultados muestran que cuanto mayor masa de probabilidad (<<d) ha sido asignada a los elementos no vistos (hasta un máximo d=0.70), se observan mejores valores de perplejidad pero peores (más altos) %PRE. Sin embargo, no se observan las diferencias existentes en los resultados de perplejidad para valores altos de K ya que tanto los resultados de %PRE como NMNA se mantienen prácticamente constantes. El descuento de Witten-Bell que era el que presentaba una perplejidad más baja, es el que peores resultados de %PRE logra y NMNA más altos. El error es totalmente similar al observado con el descuento Absoluto. Con el descuento Acotado d=0.99 se obtienen los mejores resultados de error y NMNA (más bajos) independientemente del valor de K.

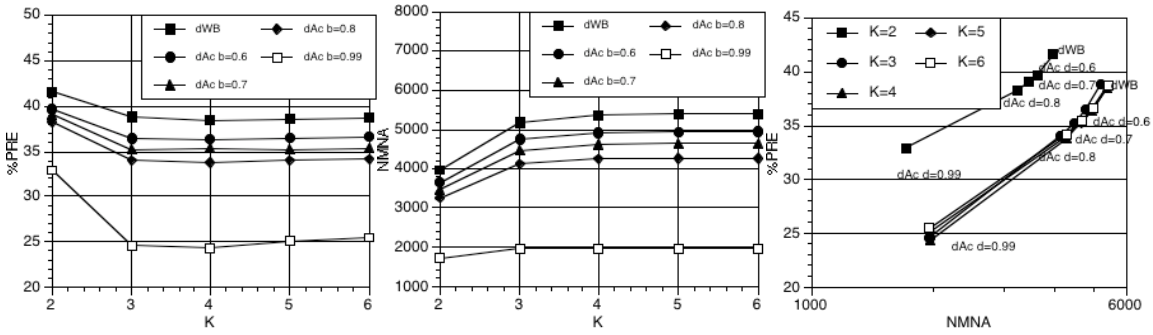


Figura 5.13.- a) Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) con el descuento Acotado con distintos valores del parámetro d y el descuento de Witten-Bell. b) Número medio de nodos activos (NMNA) necesarios para la decodificación de las frases en función de K. c) %PRE obtenido en función del NMNA.

5.3.1.5.- Resumen de resultados (α=1).

En este punto se presenta un resumen de los resultados más significativos obtenidos, tanto en perplejidad (PP) como en decodificación (%PRE y NMNA), para cada una de las técnicas de descuento estudiadas. En los casos en los que el descuento depende del valor de algún parámetro se ha elegido un modelo bastante suavizado (óptimo en perplejidad) y otro modelo poco suavizado (pobres resultados de perplejidad). En concreto se van a comparar los resultados obtenidos con el descuento Absoluto b=0.4 y b=0.01, descuento Acotado d=0.7 y d=0.99 y descuento Lineal l=0.1 y l=0.01. Primero se presentan los resultados de perplejidad (figura 5.14). En este caso se presentan de nuevo los resultados sobre el conjunto de evaluación de texto, para ver que los resultados son muy similares y la tendencia en el comportamiento se mantiene.

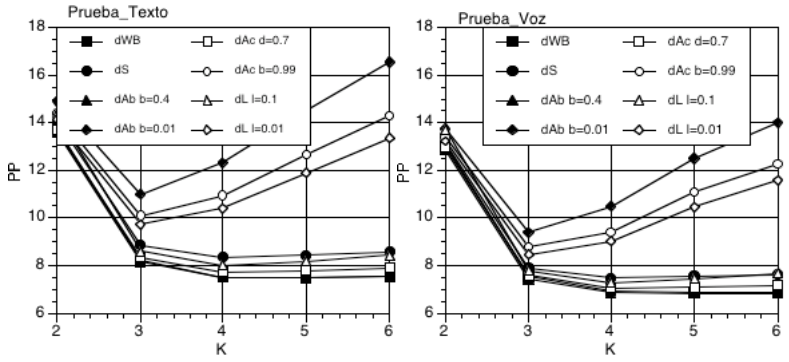


Figura 5.14.- Resumen de las perplejidades más significativas obtenidas con los modelos integrados de orden K (K=2...6) utilizando los descuentos de Witten-Bell, Sencillo, Absoluto, Lineal y Acotado sobre los conjuntos de evaluación Prueba_Texto y Prueba_Voz.

En las figura 5.15 se presentan los resultados en detalle para aquellos descuentos que presentan resultados más similares (alrededor de 7), también sobre los conjuntos Prueba_Texto y Prueba_Voz.

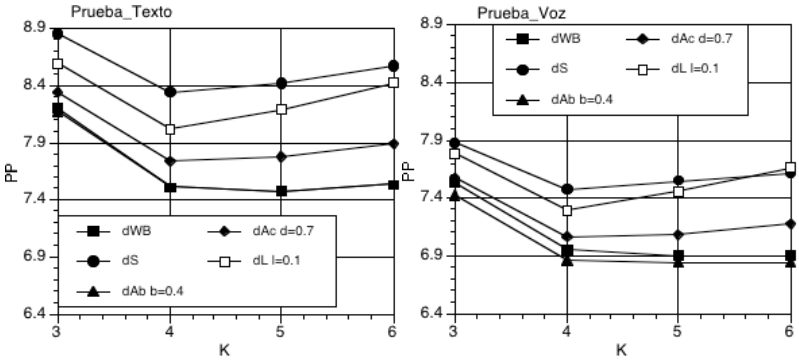


Figura 5.15.- Detalle de las perplejidades más significativas obtenidas con los modelos integrados de orden K (K=3...6) utilizando los descuentos de Witten-Bell, Sencillo, Absoluto, Lineal, y Acotado sobre los conjuntos de evaluación Prueba_Texto y Prueba_Voz.

Se puede observar que los mejores resultados de perplejidad se obtienen para el descuento Absoluto b=0.4 y descuento de Witten-Bell (en el conjunto de evaluación de texto (Prueba_Texto) las diferencias no se pueden distinguir), seguidos de los obtenidos con el

descuento Acotado d=0.7. Los resultados obtenidos utilizando el descuento Lineal l=0.1 y el descuento Sencillo se mantienen próximos pero con diferencias mayores. Los peores resultados de perplejidad se obtienen para el descuento Absoluto b=0.01 seguido del descuento Acotado b=0.99 y descuento Lineal l=0.01.

A continuación se presenta un resumen de los resultados %PRE y NMNA obtenidos cuando los modelos se integran en el sistema de RAH utilizando directamente la regla de Bayes (figura 5.16). Los mejores resultados se observan cuando se utiliza el descuento Absoluto b=0.01. A continuación se observan los mejores resultados con el descuento Acotado b=0.99 y el descuento Lineal l=0.01 (justo la relación inversa con los resultados de perplejidad obtenidos). Los resultados observados con descuento Lineal l=0.01 y descuento Sencillo son prácticamente los mismos, aunque en perplejidad existen diferencias notables sobre todo para valores altos de K.

Cuando se utiliza el descuento Absoluto b=0.4 y el descuento Acotado d=0.7 se obtienen prácticamente los mismos resultados a pesar de las pequeñas diferencias de perplejidad observadas entre ellos y el descuento de Witten-Bell. Por último, el descuento de Witten-Bell y descuento Lineal l=0.1 presentan prácticamente los mismos resultados de error aunque el descuento Lineal l=0.1 necesita mantener un mayor número de nodos activos. Sin embargo las diferencias en perplejidad son significativas entre ambos descuentos sobre todo para valores altos de K (ver figura 5.17).

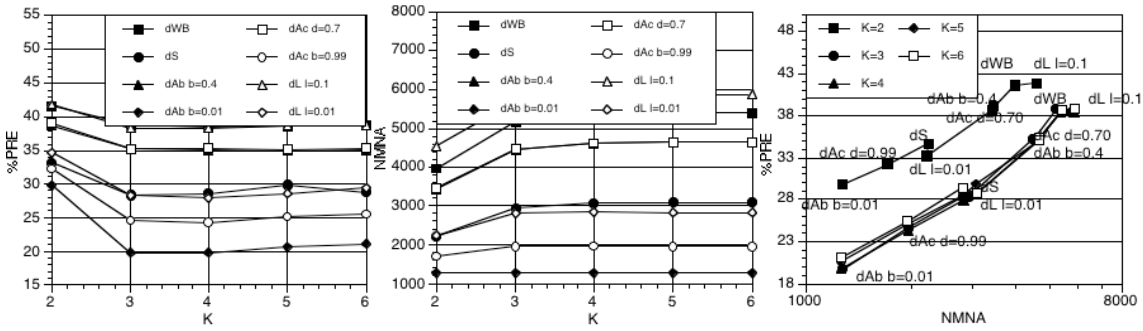


Figura 5.16.- a) Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) utilizando los descuentos de Witten-Bell, Sencillo, Absoluto (b=0.4 y b=0.01), Acotado (d=0.7 y d=0.99) y Lineal (l=0.1 y l=0.01). b) Número medio de nodos activos (NMNA) necesarios para la decodificación de las frases en función de K. c) %PRE obtenido en función del NMNA.

En general una menor masa de probabilidad asignada a los eventos no vistos por la técnica de descuento (hasta un mínimo), provoca una distribución de probabilidad más dispersa entre en el rango de valores permitidos, se obtienen peores resultados de perplejidad, pero una mejor eficiencia (menor %PRE y número medio de nodos activos) de los modelos de orden K integrados en el sistema de RAH.

Los resultados presentados en la figura 5.16 claramente difieren con los resultados de perplejidad presentados en las figuras 5.14 y 5.15. De hecho, los resultados de perplejidad parecen estar relacionados con los de error en orden inverso como puede apreciarse en detalle en la figura 5.17. Por tanto, parece cuestionable la habilidad de la perplejidad para predecir el

comportamiento de los distintos descuentos, aplicados para suavizar los ML que se integran en el sistema de RAH.

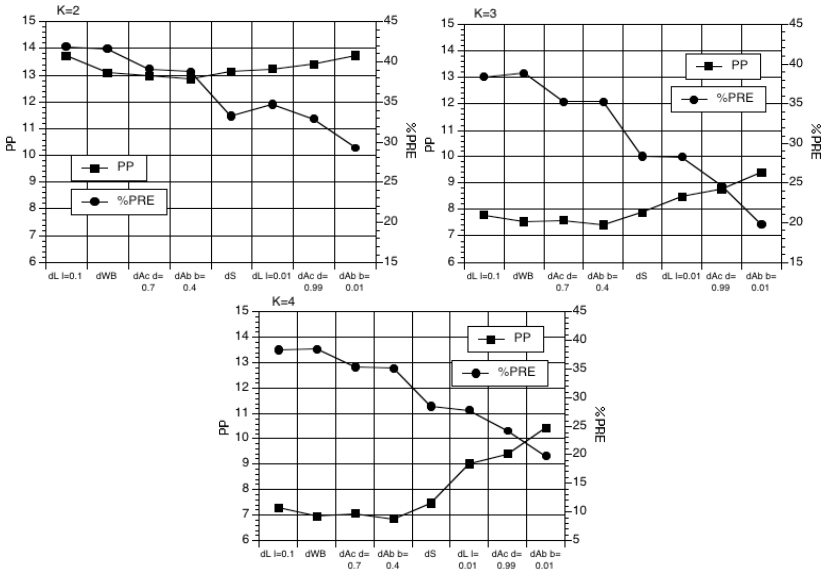


Figura 5.17.- Resultados de perplejidad (PP) y %PRE obtenidos para los modelos integrados de orden K (K=2...4) con los distintos descuentos evaluados: Lineal (dL l=0.1), de Witten-Bell (dWB), Acotado (dAc d=0.7), Absoluto (b=0.4), Sencillo (dS), Lineal (dL l=0.01), Acotado (dAc d=0.99) y Absoluto (dAb b=0.4).

5.3.2.- Resultados reescalando las probabilidades del ML

A continuación vamos a ver el efecto del reescalado exponencial de las probabilidades del ML (α). Como se ha visto en el capítulo 4, con ello se logra una disminución del error y del número medio de nodos activos. En este punto se van a evaluar los efectos de la aplicación de la función exponencial sobre las probabilidades de los modelos integrados de orden K suavizados con el descuento Sencillo, Absoluto (b=0.4 y b=0.01), Acotado (d=0.70, d=0.99) y Lineal (l=0.1 y l=0.01). De nuevo los resultados se van a comparar con los obtenidos cuando se utiliza el descuento de Witten-Bell (ver capítulo 4). En primer lugar, la figura 5.18 muestra el %PRE cuando los modelos integrados de orden K (K modelos *k*-EE) con el descuento de Witten-Bell y el descuento Sencillo se integran en el sistema de RAH. Se presentan los resultados para K=2, K=3 y K=4 (mínimo porcentaje de error) en la tarea BDGEO.

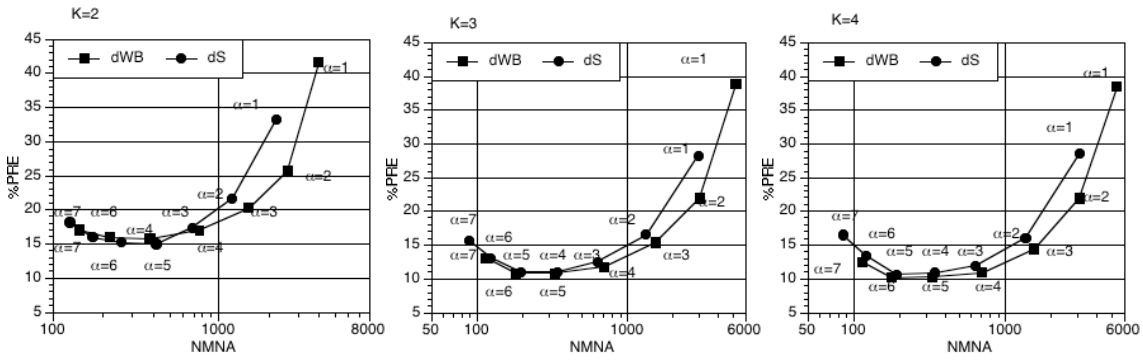


Figura 5.18.- Resultados de %PRE obtenidos en función del NMNA para los modelos integrados de orden K (K=2, 3 y 4) con el descuento Sencillo y el descuento de Witten-Bell.

Los puntos situados el margen inferior izquierdo de cada figura son los que representan la mayor eficacia del sistema, es decir el %PRE más bajo junto con un menor número medio de nodos activos (NMNA).

En ambos casos a medida que aumenta el valor de α (hasta un máximo) hay una notable disminución del error junto con una disminución del número de nodos activos por frame, lo que coincide exactamente con los resultados presentados en el capítulo 4. Con K=2 el %PRE se sitúa alrededor del 15% en ambos casos (resultados ligeramente mejores para el modelo con el descuento Sencillo). Los resultados obtenidos para K=3 y K=4 son muy similares, un poco mejores para K=4 con alrededor del 10% del %PRE, tanto con el descuento de Witten-Bell como con el descuento Sencillo, cuando se aplica el valor α adecuado (α=6 para el descuento de Witten-Bell y α=5 para el descuento Sencillo). La aplicación de la función exponencial sobre las probabilidades del ML, supone una redistribución de la distribución de probabilidad para conseguir un mayor grado de dispersión entre las probabilidades altas y bajas. Como los modelos con el descuento Sencillo presentan una distribución de la probabilidad de partida con un mayor grado de dispersión (ver ecuación 5.9), se necesita un valor menor de α para conseguir el optimo.

Por otra parte, con el descuento Sencillo las diferencias entre el peor %PRE obtenido, (alrededor del 33% en K=2 y 28% en K =3 y K =4) y el mejor (15% in K=2 y 10% en K =3 y K =4) son significativamente más pequeñas. A pesar de que el descuento Sencillo es una técnica de suavizado extremadamente simple y presenta peores resultados en PP para todos los valores de K, se obtuvieron mejores resultados en %PRE cuando α=1 y muy similares para los valores de α que proporcionan los mejores resultados en cada caso.

El la figura 5.19 se comparan los resultados de PP y de %PRE obtenidos cuando se trabaja con el valor óptimo de α. Aunque las diferencias en los resultados no son muy significativas parece que la perplejidad es capaz de predecir la bondad de los resultados de decodificación siempre y cuando se localice el valor del parámetro α adecuado**.**

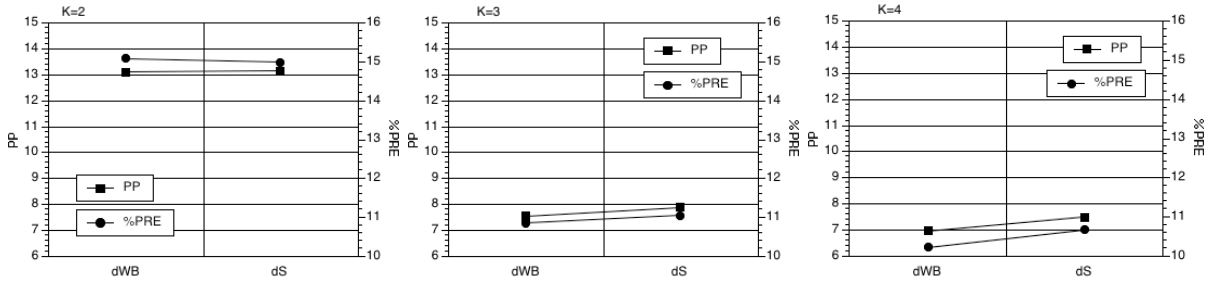


Figura 5.19.- Resultados de perplejidad (PP) y %PRE obtenidos para los modelos integrados de orden K (K=2...4) utilizando los descuentos de Witten-Bell (α=6) y descuento Sencillo (α=5).

A continuación se presentan los resultados obtenidos para las otras tres técnicas de descuento, en las que los resultados dependen de los valores que tomen los respectivos parámetros. Se va a realizar el estudio en cada caso para dos valores concretos, por una parte, el que proporciona resultados óptimos de perplejidad y por otra, el que proporciona resultados

bastante pobres. En definitiva se van a evaluar el descuento Absoluto (b=0.4 y b=0.01), Acotado (d=0.70 y d=0.99) y Lineal (l=0.1 y l=0.01).

El comportamiento observado cuando se utiliza el descuento Absoluto y el descuento Acotado es totalmente similar (ver figura 5.20 y figura 5.21). Con valores pequeños del parámetro α, los modelos menos suavizados (b=0.01 en el descuento Absoluto y d=0.99 en el descuento Acotado) obtienen los mejores resultados de decodificación y se logra el óptimo para un valor del parámetro α bastante bajo, α=3 con K>2. Los resultados obtenidos con estos modelos se degradan alrededor del óptimo, sobre todo para valores altos de K a partir de K=3. Con K=2 se obtienen los mejores resultados y también para un valor alto del parámetro α.

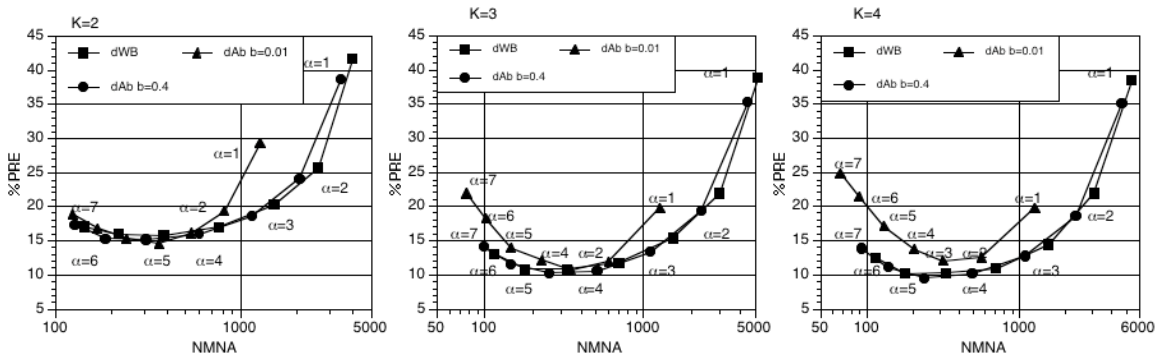


Figura 5.20.- Resultados de %PRE obtenidos en función del NMNA para los modelos integrados de orden K (K=2, 3 y 4) utilizando el descuento Absoluto con distintos valores del parámetro b y el descuento de Witten-Bell.

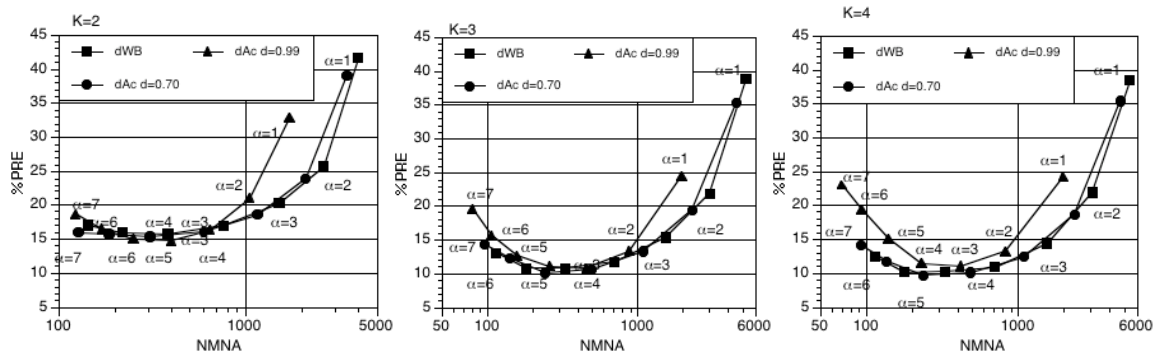


Figura 5.21.- Resultados de %PRE obtenidos en función del NMNA para los modelos integrados de orden K (K=2, 3 y 4) utilizando el descuento Acotado con distintos valores del parámetro d y el descuento de Witten-Bell.

Para los modelos más suavizados utilizando el descuento Absoluto (b=0.4) y el descuento Acotado (d=0.70) se obtienen resultados muy similares alrededor del óptimo, coincidiendo con los obtenidos con el descuento de Witten-Bell. La diferencia es que el valor del parámetro de atenuación de las probabilidades del ML que logra el óptimo es α=5 para el descuento Acotado y descuento Absoluto y α=6 para el descuento de Witten-Bell.

Los resultados obtenidos con el descuento Lineal son algo diferentes (figura 5.22). En este caso el comportamiento de los modelos con el descuento Lineal l=0.1 alrededor del óptimo corresponde prácticamente con el observado aplicando el descuento de Witten-Bell (en ambos

casos el óptimo del %PRE se obtiene para α=6). Para el descuento Lineal l=0.01 se obtiene el óptimo para un valor del parámetro α menor, pero las diferencias no son muy significativas aunque aumente el valor de K.

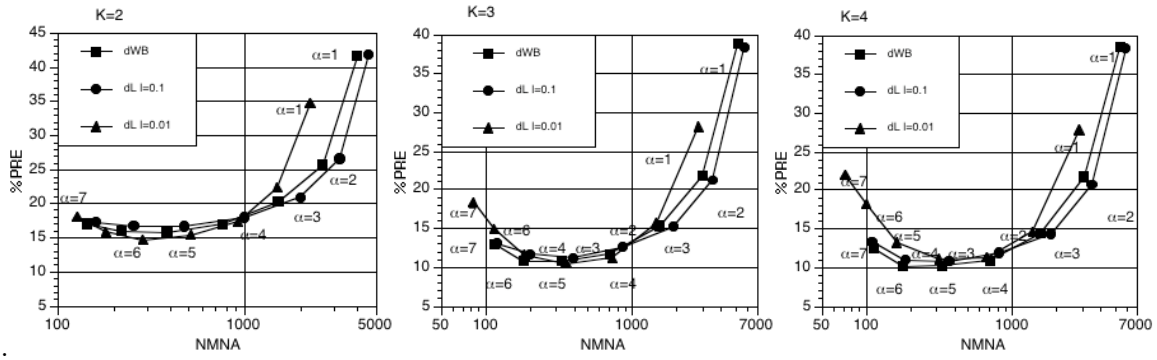


Figura 5.22.- Resultados de %PRE obtenidos en función del NMNA para los modelos integrados de orden K (K=2, 3 y 4) utilizando el descuento Lineal con distintos valores del parámetro l y el descuento de Witten-Bell.

5.3.2.1.- Resumen de resultados (α>1).

A continuación se presenta (ver figura 5.23) un resumen de los resultados de decodificación obtenidos más significativos alrededor de óptimo (α=3, 4, 5, 6) cuando los modelos integrados de orden K con (K=2, 3, 4, 5 y 6) se integran en el sistema de decodificación. En este caso se han añadido los resultados obtenidos con valores mayores de K, para mostrar que la decodificación es posible y que los resultados obtenidos son similares.

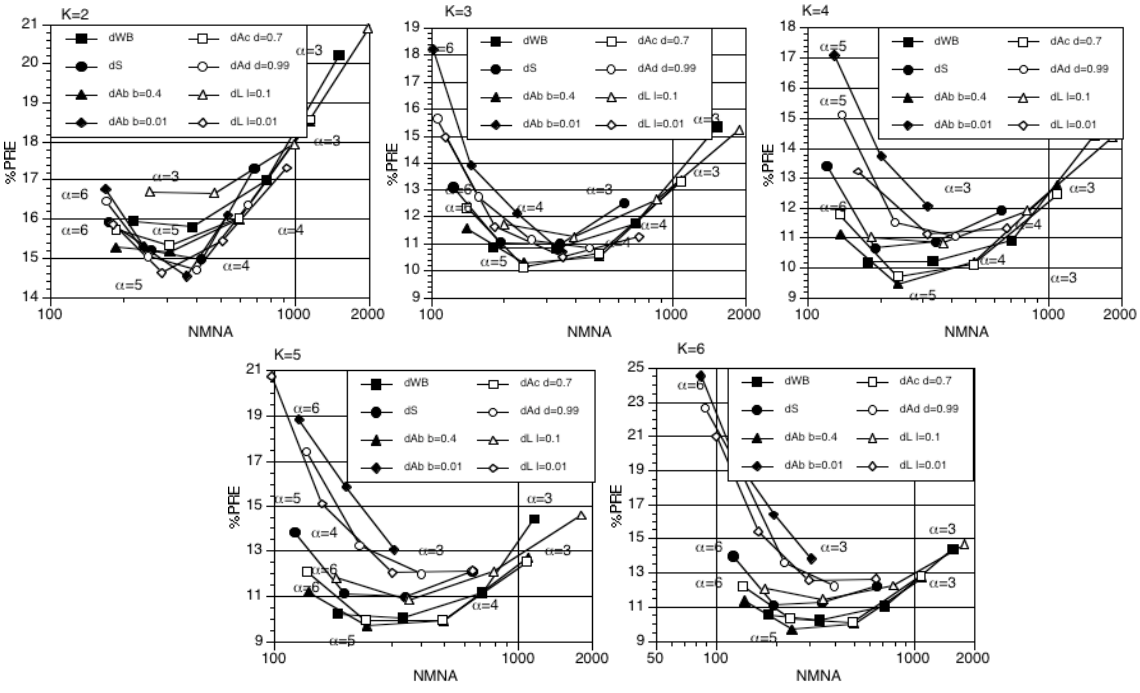


Figura 5.23.- Detalle de los resultados del %PRE más significativos obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) utilizando los descuento de Witten-Bell, Sencillo, Absoluto (b=0.4, b=0.01), Lineal (l=0.1, l=0.01) y Acotado (d=0.70, d=0.99) y α=3, 4, 5, 6.

Se observa que los ML con los descuentos que peores resultados presentan en perplejidad (descuento Absoluto b=0.01, descuento Acotado b=0.99 y descuento Lineal l=0.01), sobre todo para valores altos de K (ver figura 5.14), llegan a resultados óptimos de error más altos que los demás modelos suavizados para K>2, aunque las diferencias no son demasiado grandes, sí que son significativas. El modelo con descuento Absoluto b=0.01 claramente es el que peores resultados presenta. En los tres casos el valor de α necesario para alcanzar el óptimo es relativamente bajo (α=3) para valores de K>2 debido a que se parte de una distribución de probabilidad más dispersa en el rango de valores. Para K=2 los modelos con estos descuentos presentan los mejores resultados siendo necesarios valores altos del parámetro α (α=5 y α=6) para conseguirlos.

Para el resto de los descuentos, se observa que los modelos con descuento Sencillo y descuento Lineal l=0.1 presentan los peores resultados en perplejidad (ver figura 5.15) y ese mismo comportamiento se manifiesta en los resultados de error (peor o más alto %PRE), sobre todo para valores altos de K.

Sin embargo, para los otros tres descuentos, las diferencias observadas en perplejidad (ver figura 5.14 y 5.15) que eran muy poco significativos no se manifiestan en los resultados de error, donde los modelos con descuento Absoluto con b=0.4 y descuento Acotado con d=0.7 presentan resultados muy similares y mejores que los observados con el descuento de Witten-Bell como se puede ver en la figura 5.23 (el óptimo para los descuentos Absoluto y Acotado se obtiene para α=5 y para el descuento de Witten-Bell para α=6). Estos tres descuentos tratan de dejar las cuentas más altas inalteradas, siendo el descuento Acotado el que las deja totalmente inalteradas.

Todas estas apreciaciones se pueden observar con mayor claridad en la figura 5.24 donde se comparan los resultados de PP y del mejor %PRE obtenido para cada uno de los descuentos utilizando el valor adecuado del parámetro α en cada caso.

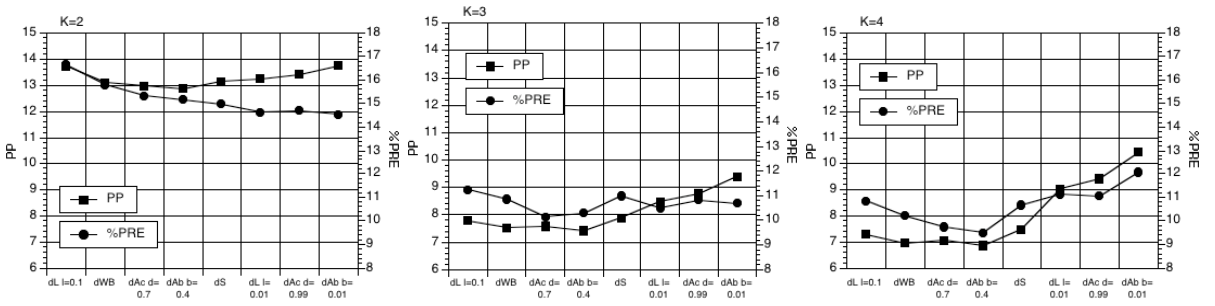


Figura 5.24.- Resultados de perplejidad (PP) y %PRE obtenidos para los modelos integrados de orden K (K=2...4) con los distintos descuentos evaluados: Lineal (dL l=0.1), de Witten-Bell (dWB), Acotado (dAc d=0.7), Absoluto (b=0.4), Sencillo (dS), Lineal (dL l=0.01), Acotado (dAc d=0.99) y Absoluto (dAb b=0.4).

Cuando K=2, la perplejidad no logra predecir en absoluto los resultados de decodificación que se obtienen al integrar los modelos en el sistema de decodificación. Con K=3 las predicciones son mejores, aunque no fiables sobre todo para los modelos muy poco suavizados. Con valores superiores de K las predicciones parecen ser más adecuadas siempre y

cuando se localice el valor adecuado del parámetro de ponderación de las probabilidades del ML (α).

En general, parece que sólo cuando las diferencias en perplejidad son muy notables, sus predicciones se manifiestan en los resultados de error, aunque siempre de una forma mucho más atenuada siendo las diferencias en la mayoría de los casos no significativas estadísticamente. Este hecho aparece también reflejado en (Rosenfeld, 1994), donde los resultados experimentales les confirmaba, la regla de la "raíz cuadrada", que dice que la mejora del error es aproximadamente la raíz cuadrada de la mejora de la perplejidad. La pregunta que se plantea inmediatamente, es ¿porqué el impacto es tan pequeño?. El ML afecta al error de reconocimiento a través de su poder discriminante, es decir, su habilidad para asignar probabilidades más altas a aquellas hipótesis que ocurren un mayor número de veces, y probabilidades más bajas para aquellas que se ven pocas veces. Pero la perplejidad afecta solamente a las probabilidades asignadas por el ML a sentencias correctas. Por eso, un ML que sobrestime las probabilidades de las sentencias no correctas no está directamente penalizado por la perplejidad. La única penalización es indirecta, puesto que asignar una probabilidad muy alta a algunas hipótesis significa una reducción en la probabilidad total asignada al resto de las hipótesis. Si la subestimación se reduce a una pequeña porción del espacio de probabilidades, la medida de la perplejidad podría ser totalmente incorrecta. Incluso ese modelo podría incrementar notablemente el porcentaje de errores, porque las probabilidades altas, asignadas a sentencias no correctas, podrían causar que fueran seleccionadas por el reconocedor.

En resumen, los experimentos muestran que los modelos integrados de orden K (K=2...6) consiguen una buena eficiencia dentro del sistema cuando se aplica el valor de α adecuado que es totalmente dependiente del tipo de descuento utilizado y que un incremento de la perplejidad no significa obligatoriamente una degradación de la eficiencia del sistema (%PRE más altos). Se podría decir que cuanta más masa de probabilidad distribuye el método de descuento entre los eventos no vistos (hasta un máximo), la distribución de probabilidad está menos dispersa, se obtienen buenos resultados de perplejidad y se mantienen cuando se incrementa el valor de K, se obtiene la peor eficiencia (alta %PRE y NMNA) con α=1 y se necesita un valor de α mayor par conseguir la máxima eficiencia (muy similar para todos los suavizados estudiados).

5.3.3.- Experimentación con una base de datos menor (tarea MLA)

Hasta ahora se han presentado los resultados obtenidos para la tarea BDGEO, observando que las técnicas de descuento y el valor del factor de ponderación de las probabilidades del ML α necesario para alcanzar el óptimo están estrechamente relacionadas, siendo la perplejidad una evaluación poco "fina" para determinar cual es el mejor descuento que se puede aplicar. Vamos a ver si este comportamiento se sigue manifestando para la tarea MLA (ver apéndice A).

En primer lugar se presentan los resultados de perplejidad obtenidos (ver figura 5.25) para el conjunto de evaluación de texto (Prueba_Texto) y para el conjunto de frases que realmente se a van a utilizar en reconocimiento (Prueba_Voz). En la figura 5.26 se muestran los resultados en detalle para aquellos descuentos que presentan resultados más similares.

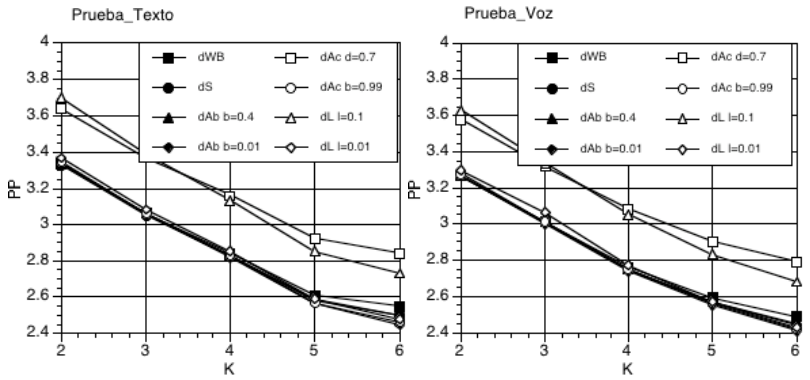


Figura 5.25.- Resumen de las perplejidades más significativas obtenidas con los modelos integrados de orden K (K=2...6) utilizando el descuento de Witten-Bell, Sencillo, Absoluto, Lineal y Acotado para los conjuntos de evaluación Prueba_Texto y Prueba_Voz con la tarea MLA.

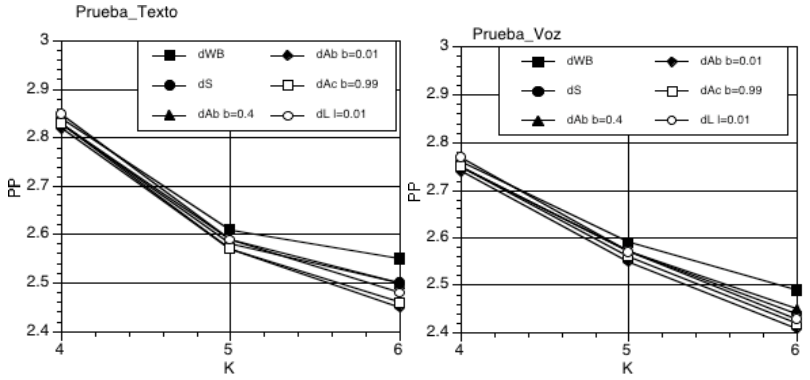


Figura 5.26.- Detalle de las perplejidades más significativas obtenidas con los modelos integrados de orden K (K=2...6) utilizando el descuento de Witten-Bell, Sencillo, Absoluto, Lineal y Acotado para los conjuntos de evaluación Prueba_Texto y Prueba_Voz con la tarea MLA.

Al contrario que para la tarea BDGEO se observa un comportamiento descendente de la perplejidad a medida que aumenta el valor de K. Los peores resultados se obtienen para el descuento Lineal l=0.1 y el descuento Acotado d=0.7. El resto de los resultados son prácticamente similares obteniéndose los peores resultados para el descuento de Witten-Bell aunque las diferencias son apenas apreciables.

5.3.3.1.- Resultados sin reescalado de las probabilidades del ML

En primer lugar vamos a ver los resultados de error que se obtienen cuando los modelos integrados de orden K (K=2...6) se integraron en el sistema de RAH aplicando directamente la regla de Bayes. (figura 5.27).

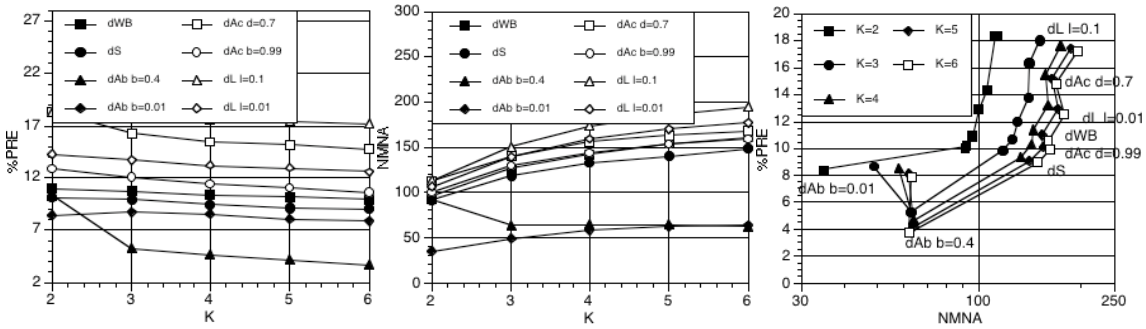


Figura 5.27.- a) Resultados de error obtenidos al integrar en el sistema de RAH los modelos integrados de orden K (K=2...6) utilizando el descuento de Witten-Bell, Sencillo, Absoluto(b=0.4 y b=0.01), Acotado (d=0.7 y d=0.99) y Lineal (l=0.1 y l=0.01). b) Número medio de nodos activos (NMNA) necesarios para la decodificación de las frases en función de K. c) %PRE obtenido en función del NMNA.

Se observan resultados contrarios a los vistos en perplejidad como viene siendo habitual y bastante distintos a los vistos para la tarea BDGEO, la tasa del error sigue disminuyendo ligeramente con K aunque también aumenta el número medio de nodos activos por frame. Se observa que los modelos con descuento Lineal l=0.1 y descuento Acotado d=0.70 presentan los peores resultados de perplejidad y son los que peores %PRE obtienen a pesar de ser dos descuentos totalmente distintos (ver ecuaciones 5.12 y 5.18).

5.3.3.2.- Resultados reescalando las probabilidades del ML

Vamos a ver ahora los resultados de error que se obtienen cuando se modifica el valor del parámetro de ponderación α de las probabilidades del ML (ver figura 5.28). En este caso el error también sigue disminuyendo con K, obteniéndose los mejores resultados para K=6. Los modelos con el descuento de Witten-Bell y el descuento Sencillo tienen un comportamiento similar (no demasiado bueno alrededor del óptimo con α=4). Los mejores resultados se obtienen con el descuento Absoluto b=0.4 (α=4), descuento Acotado d=0.99 (α=6) y descuento Lineal l=0.01 (α=6).

El descuento Absoluto b=0.4 parece que proporciona la mejor distribución de probabilidad: el óptimo se logra para un valor bajo de α (α=3) y con α=1 se obtienen también los mejores resultados. Por el contrario, con el descuento Absoluto b=0.01, los resultados obtenidos son bastante pobres aunque los valores de perplejidad no lo manifiestan. En la tarea MLA la mayor parte de los eventos vistos se repiten muchas veces por lo que están bastante bien estimados. El descuento Absoluto con un valor relativamente alto del parámetro b, b=0.4, no supone demasiado descuento sobre los eventos que han aparecido muchas veces y por eso es posible obtener el óptimo del %PRE. Por el contrario, un valor del parámetro b=0.01, supone un descuento tan sumamente pequeño que lleva a muy malos resultados de %PRE, aunque con la perplejidad no se manifiesta este hecho.

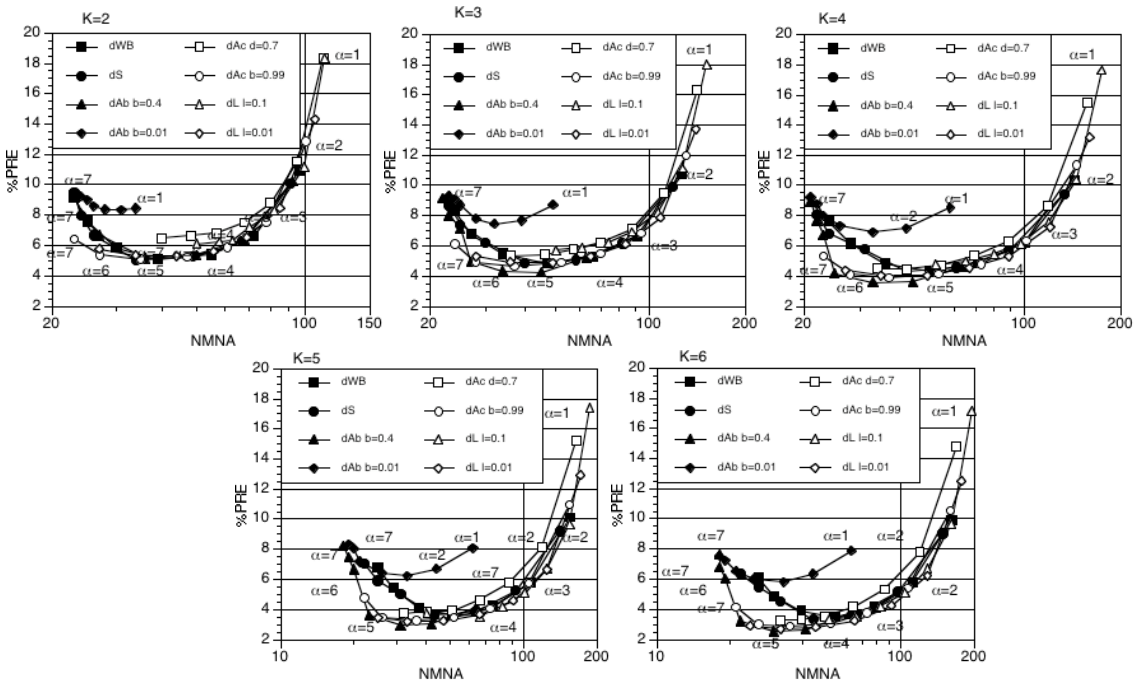


Figura 5.28.- Resumen de los resultados del %PRE más significativos obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) utilizando los descuento de Witten-Bell, Sencillo, Absoluto (b=0.4, b=0.01), Lineal (l=0.1, l=0.01) y Acotado (d=0.70, d=0.99).

El descuento Lineal resta por igual una parte de probabilidad a los eventos vistos muchas y pocas veces. Un descuento con l=0.1 es exageradamente alto y la distribución de probabilidad puede que incluso se haya invertido, otorgando más probabilidad a los eventos no vistos. Por eso los valores de perplejidad obtenidos son tan altos y se necesita un valor pequeño del parámetro α para lograr el óptimo. Por otro lado, el descuento Lineal con l=0.01 proporciona una distribución de probabilidad más suavizada y se logran resultados de %PRE próximos al óptimo para valores altos de α.

El descuento Acotado con d=0.7 también supone un descuento excesivo, puede que incluso más que el descuento Lineal l=0.1. Por eso se obtienen también valores altos de perplejidad y se necesita un valor de α pequeño para logra unos resultados cercanos al óptimo. Con d=0.99 se logra una distribución de probabilidad más suavizada y se obtienen buenos resultados de %PRE con valores altos de α.

Capítulo 6

Evaluación de los modelos de lenguaje k-EE podados

A pesar de que los modelos *k*-EE necesitan relativamente poca cantidad de memoria para ser representados, los requisitos espaciales con valores altos de *k* empiezan a ser significativamente importantes (ver capítulo 3), por lo que el uso de tales modelos podría ser cuestionado.

Además, a medida que aumenta el valor de *k*, existen cada vez más combinaciones de (*k*-1) palabras que se han visto un número muy pequeño de veces, o incluso que sólo se hayan observado en una ocasión en la muestra de entrenamiento por lo que la estimación de sus probabilidades no es en absoluto significativa.

Teniendo en cuenta este hecho, un criterio muy simple para podar los modelos consiste en exigir un número mínimo de ocurrencia de un evento para mantener su estado asociado. En este caso, la poda se lleva a cabo por el criterio de robustez. Se establece un valor umbral de ocurrencia de un evento a partir de un estado para mantener su transición a un nuevo estado. Una vez que un estado es eliminado del modelo, todas las transiciones a partir de ese estado son también eliminadas, ya que representan combinaciones de palabras de longitud aún mayor y por tanto aún más específicas.

Se pueden aplicar otros criterios de poda como por ejemplo el criterio de discriminación (Bonafonte, 1995) que consiste en fusionar en un único estado, todos aquellos que presenten una distribución de probabilidad muy similar de manera que no suponga ninguna ventaja mantenerlas por separado. Para ello, se establece un umbral de una medida de similitud como por ejemplo la distancia de Kullback-Lieber (Blah, 1987).

En este capítulo se va a estudiar el efecto que tiene sobre la tasa de error la poda de los modelos *k*-EE utilizando el criterio de robustez. Como se ha dicho anteriormente, es una técnica muy simple y consiste en establecer un valor umbral, llamado factor de poda *fp*, de modo que aquellos estados que cumplen que $N(w/q)<fp$ se eliminan del modelo. Este criterio hace que se eliminen muchos estados asociados a historias especificas que aparecen muy pocas veces y que sólo se modelen las historias que aparecen con una relativa regularidad. Es decir, los *k*-gramas poco frecuentes $w_{i-k}w_{i-(k-1)}...w_{i-1}$, con frecuencia de ocurrencia inferior a *fp*, se eliminan del modelo. Se trata de buscar el valor del parámetro *fp* adecuado para reducir el tamaño de los

modelos tratando de mantener su efectividad en el sistema integrado de RAH. El estudio se realizará para valores de K>2, ya que la poda de los modelos con K=2 significaría la eliminación de algunas de las palabras propias del vocabulario.

En el capítulo anterior se había visto la fuerte relación existente entre los distintos descuentos aplicados a los modelos *k*-EE que dan como resultado diferentes distribuciones de probabilidad y el valor del factor de ponderación de las probabilidades del ML α necesario par alcanzar el óptimo en reconocimiento. El parámetro α, $P(ML)^{\alpha}$, supone una redistribución de las probabilidades obtenidas como consecuencia de la aplicación de la técnica de suavizado. Como en los modelos podados habrá cada vez más situaciones no vistas, las distribuciones de probabilidad serán completamente distintas dependiendo de la técnica de descuento que se emplee. Por tanto, parece interesante el estudio del comportamiento del %PRE con el descuento aplicado en función del parámetro de ponderación de las probabilidades del ML.

La evaluación experimental de los modelos integrados de orden K (K modelos *k*-EE) podados se ha llevado a cabo sobre el corpus BDGEO (ver apéndice A). En primer lugar, y como viene siendo habitual se ha estudiado la degradación del modelo en base a medidas de perplejidad (PP). Después el modelo suavizado se ha integrado en el sistema de reconocimiento manteniendo el parámetro que regula el ancho de haz fijo (*ah*=0.5) para que los resultados no se vieran influidos por este factor adicional. Veremos igual que en el capítulo anterior la fuerte relación existente entre las distribuciones de probabilidad posteriores a la poda y el valor del parámetro de ponderación de las probabilidades del ML adecuado para alcanzar el mínimo de error en decodificación.

6.1.- Evaluación experimental utilizando el descuento de Witten-Bell

El objetivo de la poda es encontrar un buen compromiso entre la cantidad de memoria necesaria para representar el modelo de lenguaje y la correcta respuesta del sistema de RAH. Vamos a estudiar el efecto de la poda sobre los modelos *k*-EE utilizando el descuento de Witten-Bell (ver capítulo anterior).

La tabla 6.1 muestra la evaluación de los modelos integrados de orden K (K modelos *k*-EE con K=2...6) usando diferentes factores de poda (*fp*). Los factores de poda (*fp*) representan la cuenta umbral bajo la cual los *k*-gramas fueron desestimados (eliminados), por tanto *fp*=1 representa al modelo no podado. La tabla 1.1 muestra también el número de estados (Q^K) de los modelos integrados de orden K, y el tamaño de la memoria necesario para representarlos. Además se muestra el Número Medio de Nodos Activos (NMNA) por frame en el trellis de Viterbi (incluyendo los nodos acústicos y de ML) y el tiempo medio necesario para la evaluación de cada frame (*TMF*) cuando no hubo renormalización de las probabilidades del ML (α=1).

Tabla 6.1.- Evaluación de los modelos integrados de orden K podados (K=2...6) con diferentes factores de poda (fp). Se representa el número de estados (QK), la memoria, la perplejidad (PP), el Numero Medio de Nodos Activos por frame (NMNA), el Tiempo Medio de decodificación por frame (TMF) y el porcentaje real de error (%PRE).

K	***fp***	$\mathbf{Q^K}$	***memoria (Mb)***	***PP***	***NMNA***	***TMF(msec)***	***%PRE***
2	1	1,213	0.13	13.10	3,397	69	41.62
3	1	7,479	0.43	7.53	4,270	91	38.85
	2	3,854	0.20	8.51	2,755	59	36.44
	3	2,845	0.14	9.76	2,286	43	35.46
	4	2,336	0.11	11.03	1,960	35	34.83
	5	1,999	0.09	12.14	1,753	30	34.93
4	1	21,551	0.95	6.95	4,392	96	38.5
	2	9,360	0.38	8.07	2,809	60	36.29
	3	6,366	0.25	9.62	2,325	46	35.22
	4	4,993	0.19	10.08	1,988	36	34.75
	5	4,139	0.16	12.41	1,776	31	35.24
5	1	42,849	1.69	6.90	4,411	98	38.63
	2	16,086	0.58	8.10	2,818	60	36.26
	3	10,260	0.36	9.80	2,332	46	35.18
	4	7,795	0.26	11.28	1,994	36	35.17
	5	6,308	0.22	12.63	1,781	32	35.34
6	1	69,616	2.55	6.90	4,418	99	38.77
	2	22,839	0.77	8.18	2,822	61	36.25
	3	13,784	0.45	9.91	2,333	46	35.19
	4	10,179	0.33	11.47	1,996	37	35.07
	5	8,089	0.26	12.78	1,783	32	35.32

A continuación en la tabla 6.2 se muestra un resumen de los resultados mas significativos para los modelos no podados (*fp*=1) y los modelos podados que logran la mayor eficiencia (*fp*=4). Además se ha obtenido el intervalo de confianza tanto para los modelos podados como no podados mostrando que las diferencias en el error del sistema son estadísticamente significativas.

Tabla 6.2.- Resumen de los resultados de %PRE obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5, 6). Se añaden los valores del intervalo de confianza $[PRE^-,PRE^+]_{95\%}$ para medir la significancia estadística de los resultados (5655 palabras).

K	***fp***	***%PRE***	$[PRE^-,PRE^+]_{95\%}$
2	1	41.62	[40.25, 42.92]
3	1	38.85	[37.59, 40.13]
	4	34.83	[33.60, 36.07]
4	1	38.50	[37.24, 39.78]
	4	34.75	[33.53, 36.01]
5	1	38.63	[37.37, 39,91]
	4	35.17	[33.94, 36.43]
6	1	38.77	[37.51, 40.05]
	4	35.07	[33.84, 36.29]

Cuando se usan los modelos integrados de orden K podados (*fp*>1), se observan fuertes reducciones en el número de estados y como consecuencia en la memoria requerida para su

utilización. Estas reducciones son más significativas para valores altos de K. Se observa que hay un incremento de la perplejidad (PP) cuando se incrementa el valor de *fp* mostrando la degradación normal que sufre la estructura de los modelos podados. La figura 6.1 muestra los valores de la perplejidad (conjunto de evaluación de texto (Prueba_Texto) y en el conjunto de evaluación del sistema integrado (Prueba_Voz)) de los modelos integrados de orden K podados y no podados. Se observa que los valores de la perplejidad son prácticamente constantes para valores de K superiores a 3 independientemente del factor umbral de poda que se esté utilizando y que se produce un aumento notable de la perplejidad a medida que se incrementa el valor del *fp*.

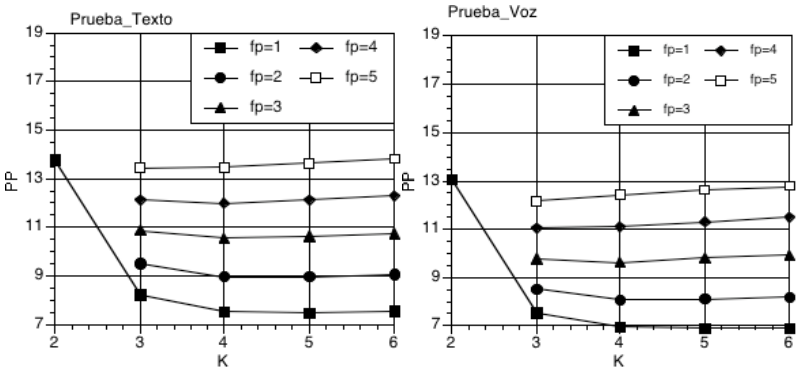


Figura 6.1.- Perplejidad obtenida con los modelos integrados de orden K (K=2...6) no podados (*fp*=1) y podados (*fp*>1).

La figura 6.2 representa el %PRE para los modelos integrados de orden K podados. Observando la tabla 6.1 y las figuras 6.1 y 6.2, se puede ver que los modelos podados consiguen mejores resultados (hasta el máximo con *fp*=4). Además el número medio de nodos disminuye de forma casi constante y como consecuencia el tiempo por frame necesario para decodificar cada sentencia.

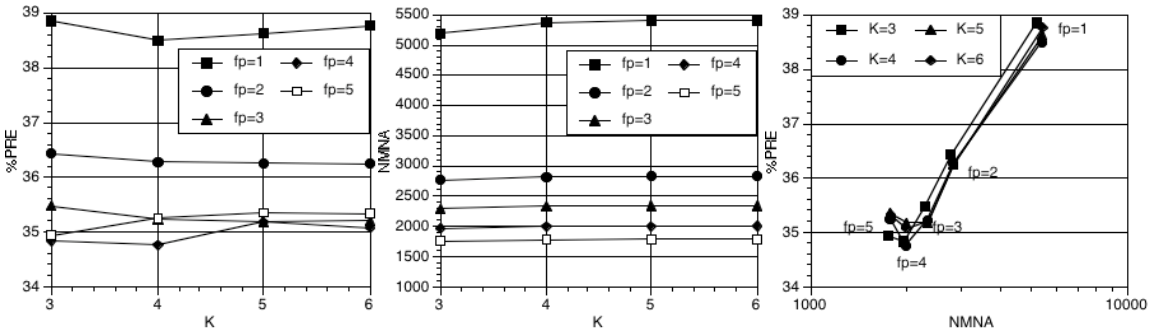


Figura 6.2.- a) %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, b) Numero medio de nodos activos en función de K, c) %PRE en función del NMNA.

Esto significa que los modelos podados logran con respecto a los modelos no podados un porcentaje de error menor, siendo necesario un gasto inferior de memoria y un tiempo menor para decodificar cada sentencia como puede verse en la figura 6.2c. Este sorprendente comportamiento se puede explicar analizando la técnica de descuento utilizada, descuento de Witten-Bell. Cuando se aplican los factores de poda *fp*, los K-gramas más infrecuentes

desaparecen del modelo. Esos K-gramas pasan ahora a ser eventos no vistos y por tanto el número de eventos vistos en cada estado ($|\Sigma_q|$) disminuye notablemente. El valor de $N(q)$ decrece también pero no de forma tan significativa. Por tanto, la distribución de la masa de probabilidad entre los eventos vistos y no vistos que otorga la técnica de descuento se ha modificado seriamente. El coeficiente de descuento ($1-\lambda$) es ahora menor (ver ecuación 5.1) y por tanto la masa de probabilidad que se asigna a la transición por el estado de back-off, es decir, a los eventos no vistos, es más pequeña lo que significa que hay menos suavizado y como consecuencia una distribución más dispersa de la probabilidad. Como habíamos visto ya en los capítulos previos eso significa un mejor comportamiento del sistema cuando no se aplica la ponderación exponencial sobre las probabilidades del ML.

Sin embargo, la perplejidad no muestra este comportamiento. La figura 6.3 representa la perplejidad y los porcentajes de error obtenidos para algunos de los modelos integrados de orden K evaluados en estos experimentos (tabla 6.1). Como ya se ha visto en el capítulo anterior un incremento de la perplejidad no siempre significa una degradación en la calidad del sistema (Ferreti et al, 1990) (Rubio te al, 1994) (Clarkson & Robinson 1999). La figura 6.3 muestra que la perplejidad incrementa su valor mientras que %PRE decrece cuando el *fp* aumenta de valor. El valor mínimo de la perplejidad y del %PRE no se consiguieron con los modelos no podados (*fp*=1). Los mejores valores se obtuvieron con *fp*= 3 y 4. Sin embargo, los valores de *fp* superiores a 4 implican una excesiva degradación de los modelos y como consecuencia un peor rendimiento del sistema aunque el rendimiento sigue siendo superior al obtenido con los modelos no podados.

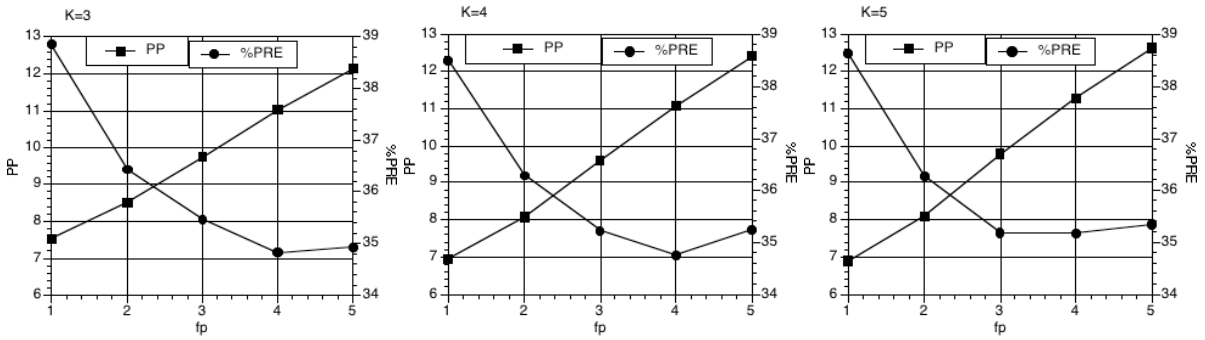


Figura 6.3.- Evolución de la perplejidad y %PRE para los modelos integrados de orden K (K=3, 4 y 5) con distintos valores del parámetro *fp* (descuento de Witten-Bell).

6.1.1.- Escalando las probabilidades del ML: P(ML)$^{\alpha}$

A continuación se presenta una nueva evaluación experimental de los modelos podados pero ponderando las probabilidades del ML con un factor exponencial α, $(P(\Omega))^{\alpha}$ al utilizar la regla de Bayes. Se han evaluado diferentes valores del parámetro α para minimizar el porcentaje de error.

La figura 6.4 muestra los resultados experimentales obtenidos usando diferentes modelos integrados de orden K (K=3...5) con diferentes valores del factor de poda (*fp*) y

diferentes valores del parámetro α. Al igual que en los capítulos precedentes los puntos en el margen inferior izquierdo de cada figura muestran el mejor funcionamiento del sistema (menor %PRE y menor número medio de nodos activos (NMNA)).

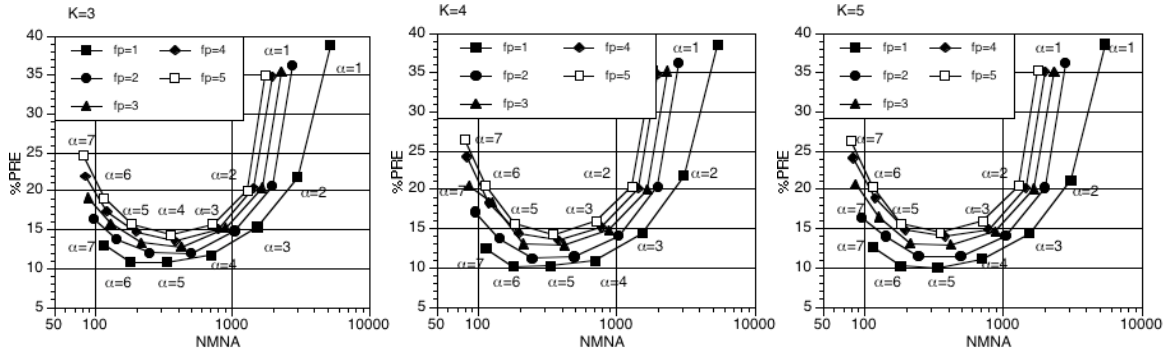


Figura 6.4.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) para distintos valores del factor de poda (*fp*) y distintos valores del parámetro α.

Para cualquiera de los modelos integrados de orden K se logró un descenso importante del %PRE junto con un notable decremento de la cantidad media de nodos implicados cuando el factor α incrementa (hasta un valor máximo). Queda de nuevo patente la importancia de utilizar el parámetro α para mejorar el comportamiento de los modelos integrados de orden K cuando se integran en un sistema completo de RAH.

Se puede resumir que:

a) el valor de α que optimiza el %PRE es ligeramente superior para los modelos no podados (6) frente a los modelos podados (4 o 5), lo cual corresponde exactamente con las conclusiones del capítulo anterior que decían que los modelos más suavizados necesitan un valor mayor del parámetro α para conseguir el óptimo. En este caso la distribución de probabilidad tras aplicar el descuento de Witten-Bell después de podar los modelos se asemeja a la distribución de probabilidad de los modelos no podados tras aplicar un valor de $\alpha>1$ en el esquema de reconocimiento. Con los modelos no podados la distribución de partida está menos concentrada en un rango concreto de valores y ese es precisamente el efecto de la aplicación del parámetro $\alpha>1$.

b) los modelos podados proporcionan mejores resultados que los no podados cuando el factor α se mantiene por debajo de 5 (ver figura 6.4). Este comportamiento está en sintonía con el caso particular que se observó en los experimentos presentados anteriormente con $\alpha=1$. Sin embargo para valores altos de α (>4), valores altos del *fp* significan un incremento del %PRE. La tabla 6.3 muestra los resultados experimentales para $\alpha=5$ (en líneas generales, es cuando se obtiene la mejor respuesta del sistema). Los modelos no podados en estas condiciones logran mejores resultados como era de esperar.

Tabla 6.3.- Evaluación de los modelos integrados de orden K (K=3...5) podados para α=5 (mejor respuesta del sistema).

K	*fp*	Q^K	TMF (msec)	%PRE	$[PRE^- PRE^+]_{95\%}$
3	1	328	16	**10.82**	[10.04 11.66]
	2	245	10	**12.07**	[11.25 12.94]
	3	214	8	**13.38**	[12.52 14.29]
	4	197	7	**14.75**	[13.85 15.70]
	5	183	6	**15.77**	[14.84 16.74]
4	1	328	16	**10.24**	[9.48 11.06]
	2	242	10	**11.24**	[10.44 12.09]
	3	210	8	**13.08**	[12.23 13.98]
	4	194	7	**14.45**	[13.56 15.39]
	5	180	6	**14.28**	[13.39 15.22]
5	1	332	17	**10.06**	[9.30 10.87]
	2	242	10	**11.48**	[10.68 12.34]
	3	211	8	**13.22**	[12.36 14.13]
	4	194	7	**14.81**	[13.91 15.76]
	5	181	6	**15.76**	[14.83 16.73]

La figura 6.5 representa la perplejidad y los porcentajes de %PRE para algunos modelos podados cuando se aplica un parámetro α=5.

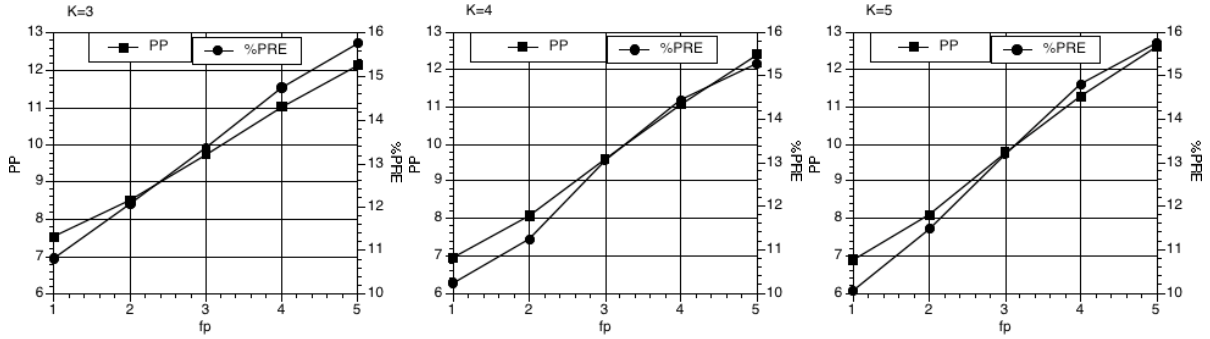


Figura 6.5.- Evolución del porcentaje de %PRE y PP para los modelos integrados de orden K (K=3, 4 y 5) en función del valor del factor de poda (*fp*) (α=5).

En este caso se observa que un incremento en la perplejidad (obtenido con modelos podados *fp*>1) significa un incremento del porcentaje de error cuando se aplica el factor α adecuado. Con el descuento estudiado (Witten-Bell), se logra una distribución de probabilidad bastante concentrada en un rango de valores para los modelos no podados. Al podar los modelos la distribución de probabilidades se hace más dispersa. Por tanto, para valores de α<4 se obtienen mejores resultados con los modelos no podados, mejores cuanto más podados para los valores de α más pequeños.

Parece que la perplejidad si que es capaz de predecir el comportamiento de los modelos podados siempre y cuando se localice el valor de α adecuado. En general, la degradación del error no es excesiva para los podas con *fp* bajos una vez que se encuentra el valor óptimo del parámetro de ponderación de las probabilidades del ML.

6.2.- Evaluación de los distintos descuentos

La poda de los modelos está directamente relacionada con las técnicas de descuento. Al podar los modelos habrá necesariamente menos situaciones vistas en entrenamiento. Sin embargo estas situaciones vistas estarán bastante bien estimadas, es decir, se habrán visto un número relativamente alto de veces. Esto implica varias cosas, entre ellas que la distribución de probabilidad antes del suavizado estará más concentrada alrededor de un rango de valores y que estará mejor estimada a medida que aumente el valor del *fp*. Dependiendo de la probabilidad que cada una de las técnicas de descuento otorgue a los eventos no vistos los modelos lograrán mejores o peores resultados al ser integrados en el sistema de RAH.

En este punto vamos a ver la evolución del error para los descuentos estudiados en el capítulo anterior: descuento Sencillo (dS), descuento Absoluto (dAb b=0.4 y b=0.01), descuento Lineal (l=0.1, l=0.01) y descuento Acotado (dAc d=0.70 y dAc d=0.99). En primer lugar, se realizará una evaluación experimental del descuento en sí mismo aplicando directamente la regla de Bayes en el esquema de decodificación. A continuación se estudiará la evolución de las tasas de error en función del factor de ponderación de las probabilidades del ML: $P(ML)^{\alpha}$.

6.2.1.- Resultados sin reescalado de las probabilidades del ML

A continuación se muestran los resultados de perplejidad y de decodificación obtenidos cuando no se aplica ningún reescalado sobre las probabilidades del ML.

6.2.2.1.- Descuento Sencillo

En la figura 6.6 se ven los resultados de perplejidad que se han obtenido cuando se utiliza el descuento Sencillo (dS) para el conjunto de frases que constituyen el corpus de evaluación de voz (Prueba_Voz) (ver apéndice A).

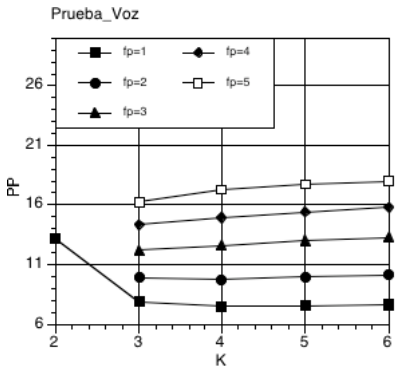


Figura 6.6.- Resultados de perplejidad obtenidos con los modelos integrados de orden K (K=2, 3, 4, 5 y 6) suavizados con el descuento Sencillo.

Si comparamos los resultados con los resultados obtenidos con el descuento de Witten-Bell (ver figura 6.1) vemos que la perplejidad se degrada bastante más, al incrementarse el valor de *fp* (no hay que olvidar que los resultados de perplejidad de los modelos no podados utilizando el descuento Sencillo son inferiores a los obtenidos con el descuento de Witten-Bell

para cualquier valor de K). En la figura 6.7 se muestran los resultados del %PRE y NMNA obtenidos cuando los modelos se integraron en el sistema de RAH aplicando la directamente regla de Bayes.

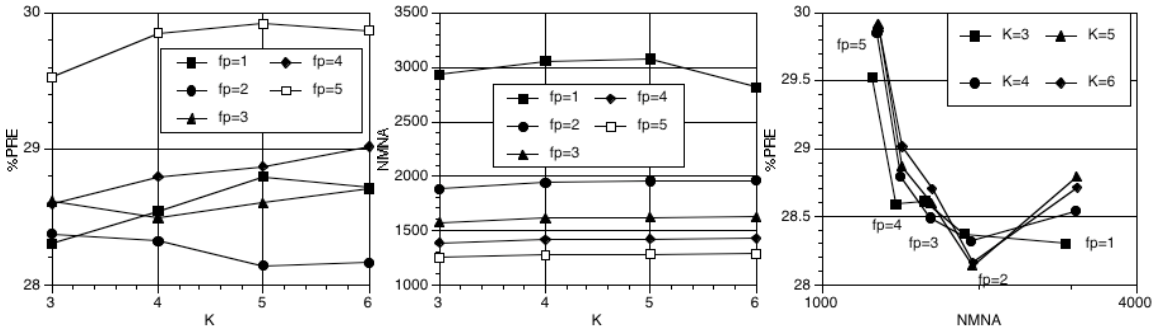


Figura 6.7.- a) %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Sencillo b) Numero medio de nodos activos en función de K, c) %PRE en función del NMNA.

Se observa que el óptimo de error se obtiene con *fp*=2, siendo la diferencia en los %PRE obtenidos mucho menos significativa que la obtenida con el descuento de Witten-Bell. Se observa también una fuerte reducción del número medio de nodos activos a medida que aumenta el *fp* igual que la observada con el descuento de Witten-Bell. El distinto comportamiento observado con respecto al descuento de Witten-Bell es lógico ya que con este método de descuento se reparte menor cantidad de probabilidad entre los eventos no vistos (ver ecuación 5.9) que son cada vez más a medida que aumenta el valor de *fp*.

6.2.2.2.- Descuento Absoluto

Vamos a presentar a continuación los resultados que se obtienen cuando se aplica el descuento Absoluto (dAb) con un valor del parámetro b=0.4 (modelo bastante suavizado con óptimo en perplejidad para los modelos no podados) y b=0.01 (resultados pobres de perplejidad para los modelos no podados y distribución final de probabilidad poco suavizada).

En la figura 6.8 se pueden apreciar los resultados de perplejidad obtenidos utilizando el corpus de evaluación de voz (Prueba_Voz).

Con el descuento Absoluto b=0.4 los resultados de perplejidad obtenidos son similares a los obtenidos con el descuento de Witten-Bell, aunque algo peores a medida que aumenta el valor del parámetro que regula la poda. Para *fp*=5 los resultados son incluso peores que los obtenidos para el modelo K=2 no podado. Se observa una ligera tendencia al incremento de la perplejidad con K.

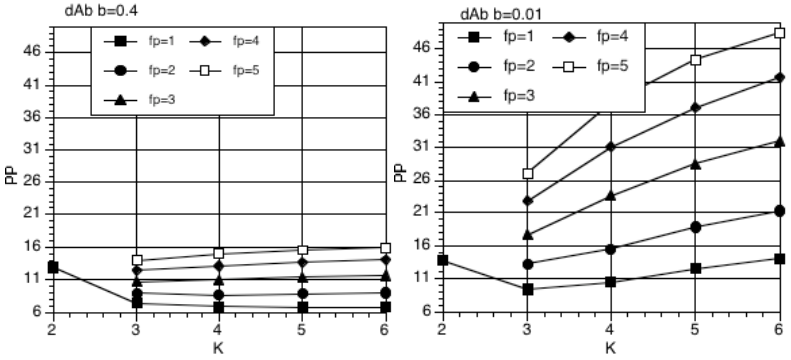


Figura 6.8.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5, 6) suavizados con el a) descuento Absoluto b=0.4, b) descuento Absoluto b=0.01.

En la figura 6.8b se observan los resultados obtenidos con el descuento Absoluto b=0.01. En este caso la degradación en perplejidad es muy notable, y esta degradación es mayor para valores altos de K. Cuando se aplica la poda hay que repartir la probabilidad entre mas estados y siempre se va a llegar a una distribución de probabilidad con valores más dispersos. En este caso sólo los eventos vistos un número relativamente alto de veces (dependiente de *fp*) mantienen una probabilidad alta. A continuación se van a ver los resultados de error obtenidos cuando los modelos se integran en un sistema de RAH y se aplica la regla de Bayes directamente. En la figura 6.9 se muestran los resultados cuando se utiliza el descuento Absoluto b=0.4 y en la figura 6.10 cuando b=0.01.

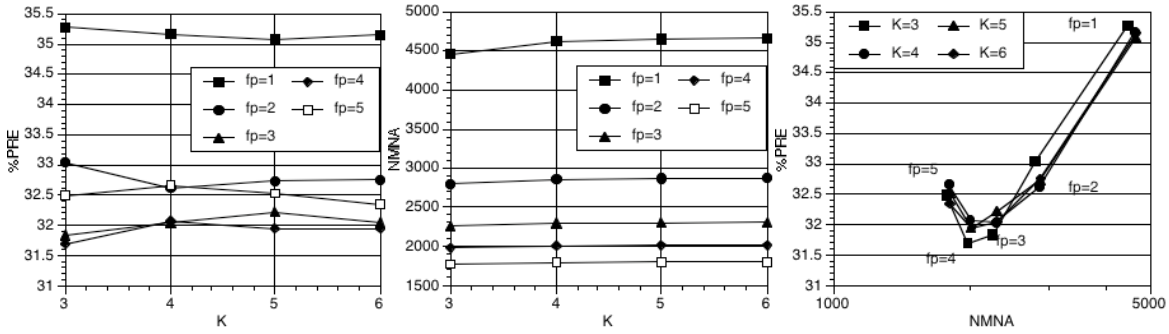


Figura 6.9.- a) %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Absoluto b=0.4 b) Numero medio de nodos activos en función de K, c) %PRE en función del NMNA.

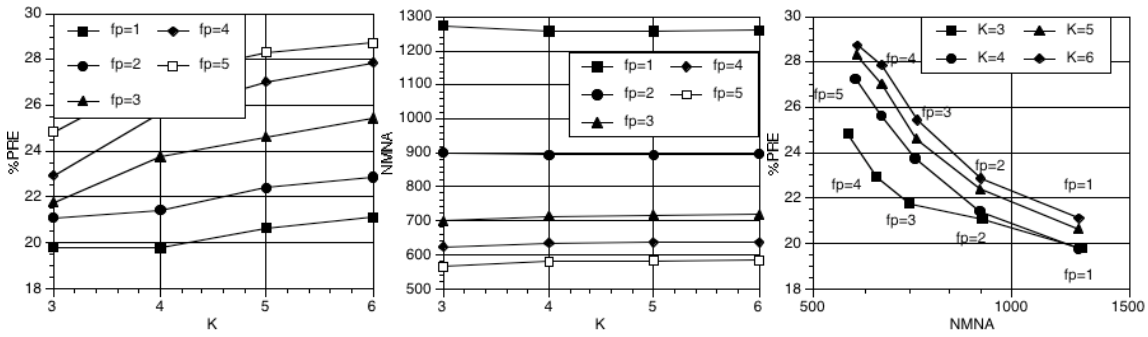


Figura 6.10.- a) %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Absoluto b=0.01 b) Numero medio de nodos activos en función de K, c) %PRE en función del NMNA.

Cuando b=0.4 el comportamiento que se observa es muy similar al observado con el descuento de Witten-Bell, el %PRE que se obtiene para α=1 disminuye con el factor *fp* hasta *fp*=4, siendo los porcentajes en números absolutos más bajos que los obtenidos para el descuento de Witten-Bell (menor %PRE).

Para el descuento Absoluto b=0.01, el aumento de *fp* nos lleva a una reducción de estados pero también se produce un aumento muy notable de las tasas de error. Los resultados ponen de manifiesto que los modelos podados poco suavizados logran malos resultados al ser integrados directamente en un sistema de RAH. Ésto es porque cada vez hay un mayor número de transiciones con una probabilidad excesivamente pequeña.

6.2.2.3.- Descuento Lineal

Se van a presentar los resultados obtenidos cuando se aplica el descuento Lineal l=0.1 (distribución final de probabilidades bastante suavizada y resultado bastante bueno de perplejidad para los modelos no podados) y el descuento Lineal l=0.01 (modelo muy poco suavizado con pobres resultados de perplejidad sobre todo para valores altos de K).

A continuación (figura 6.11) se muestran los resultados de PP que se obtienen para el descuento Lineal con valores del parámetro l=0.1 y l=0.01.

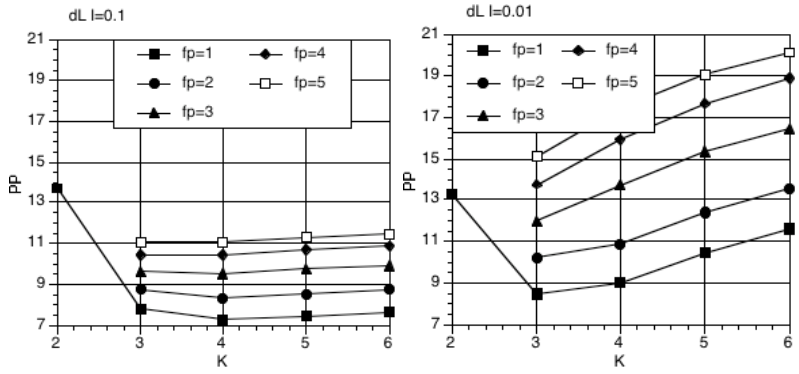


Figura 6.11.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5, 6) suavizados con el a) descuento Lineal l=0.1 b) descuento Lineal l=0.01.

Los modelos integrados de orden K con el descuento Lineal l=0.1 son los únicos modelos en los que las perplejidades de los modelos podados a partir de K>3 se mantienen por debajo del valor de la perplejidad para el modelo no podado con K=2. Cuando se aplica una técnica de poda siempre resultan modelos menos suavizados (una distribución de probabilidad más dispersa en el rango de valores permitidos) porque hay menos estados a los que sustraer una cierta masa de probabilidad y además hay que repartir la probabilidad entre mas estados. El resultado es que la degradación de la perplejidad es muy notable. Con el descuento Lineal se realiza un descuento sobre todos los eventos vistos que es proporcional a su probabilidad, es decir, a los eventos vistos muchas veces se les resta la mayor probabilidad. Como consecuencia, a pesar de estar podados aun se reserva más masa de probabilidad para los eventos no vistos que en el resto de descuentos estudiados y por ello la degradación de la probabilidad es menor. Este

mismo hecho se manifiesta en los resultados obtenidos para el descuento Lineal l=0.01, la degradación en perplejidad es muy notable pero mucho menor que la observada para el descuento Absoluto b=0.01. A continuación se muestran los resultados de error cuando los modelos integrados de orden K e integran en el sistema de reconocimiento utilizando directamente la regla de Bayes. En la figura 6.12 se muestran los resultados para el descuento Lineal l=0.1 y en la figura 6.13 para l=0.01.

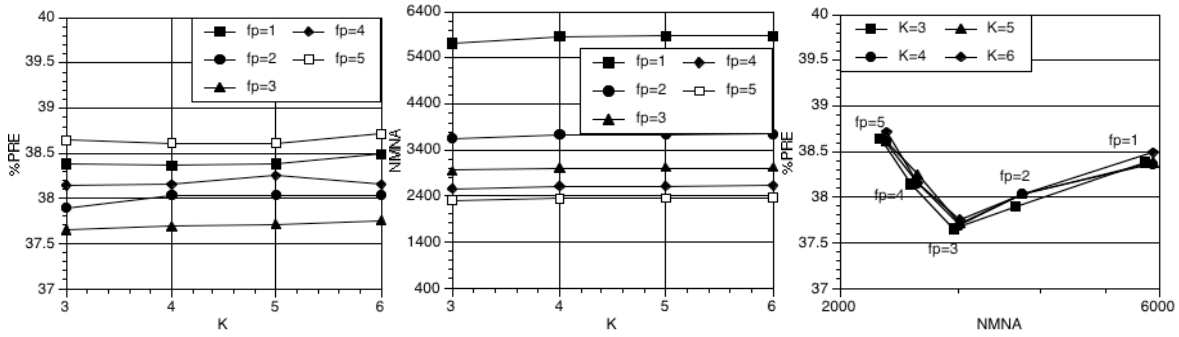


Figura 6.12.- a) %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Lineal l=0.1 b) Numero medio de nodos activos en función de K, c) %PRE en función del NMNA.

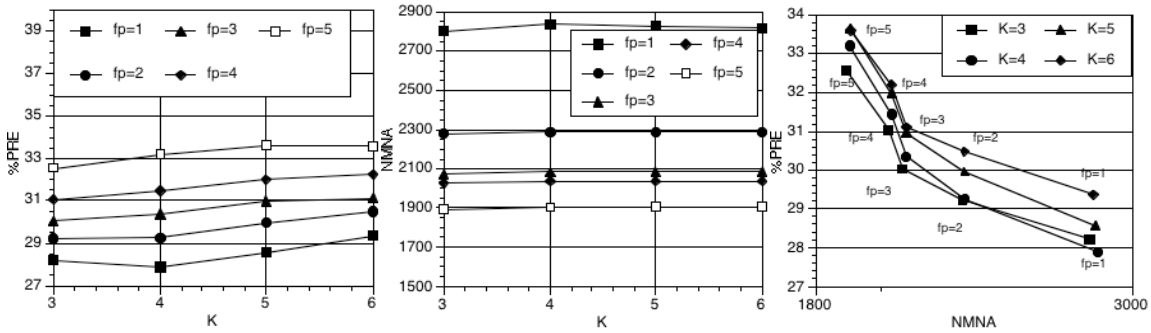


Figura 6.13.- a) %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Lineal l=0.01, b) Numero medio de nodos activos en función de K, c) %PRE en función del NMNA.

Cuando l=0.1 se observa un óptimo de error con *fp*=3. Sin embargo, para l=0.01 el aumento de *fp* nos lleva a una reducción de estados pero también al aumento notable de las tasas de error.

6.2.2.4.- Descuento Acotado

Se presentan a continuación los resultados que se obtienen cuando se aplica el descuento Acotado (dAc) con un valor del parámetro d=0.7 (con distribución de probabilidad bastante suavizada y resultado bastante bueno en perplejidad para los modelos no podados y) y d=0.99 (resultado bastante pobre en perplejidad para los modelos no podados y distribución de probabilidad poco suavizada). En la figura 6.14 se presentan los resultados de perplejidad.

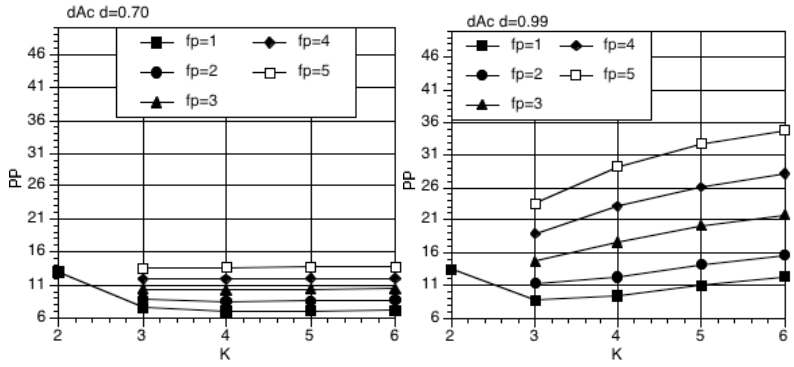


Figura 6.14.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5, 6) suavizados con el a) descuento Acotado d=0.70, b) descuento Acotado b=0.99.

Cuando se utiliza el descuento Acotado d=0.7 se observa un aumento de la perplejidad parecido al observado con el descuento de Witten-Bell y el descuento Absoluto b=0.4 y que se mantiene con valores altos de K. Para el descuento Acotado d=0.99 la degradación en los valores de la perplejidad es considerable. A continuación se muestran los resultados obtenidos cuando los modelos se integraron en el sistema de RAH utilizando directamente la regla de Bayes (ver figuras 6.15 y 6.16).

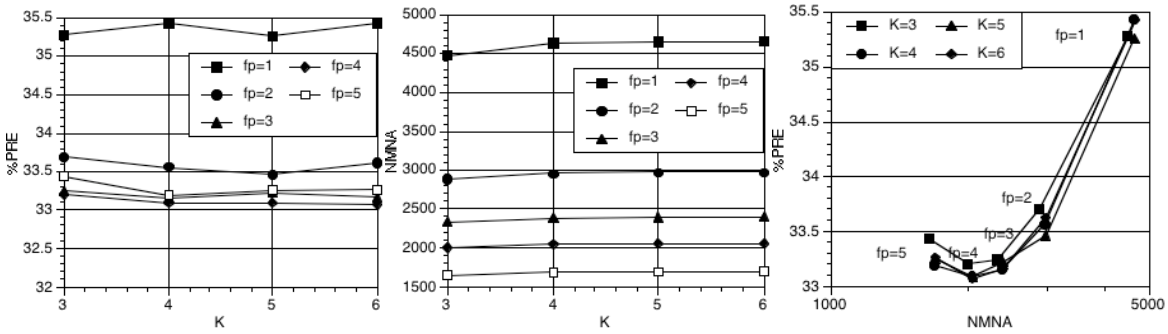


Figura 6.15.- a) %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Absoluto d=0.7 b) Número medio de nodos activos en función de K, c) %PRE en función del NMNA.

Para el descuento Acotado d=0.7 se vuelven a observar resultados muy parecidos a los que se obtuvieron con el descuento de Witten-Bell y Absoluto b=0.4. Se obtienen los mejores resultados de reconocimiento para un valor del factor *fp*=4.

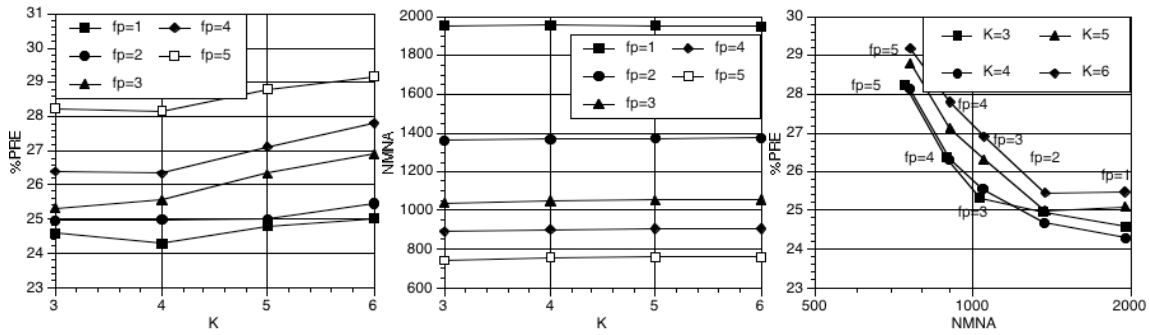


Figura 6.16.- a) %PRE que se obtienen para los modelos integrados de orden K (K=3...6) podados, y suavizados con el descuento Lineal l=0.01 b) Numero medio de nodos activos en función de K, c) %PRE en función del NMNA.

Para el descuento Acotado d=0.99 al igual que con el descuento absoluto b=0.01 y descuento Lineal l=0.01 la degradación con el factor de poda (*fp*) es muy fuerte.

6.2.2.5.- Resumen de resultados (α=1).

En primer lugar, se presenta un resumen de los resultados de perplejidad obtenidos con los modelos no podados *fp*=1 y con los distintos modelos podados (*fp*>1) (ver figura 6.17a y figura 6.18). Además se muestran también los resultados más en detalle para aquellas situaciones en las que los resultados de perplejidad son más parecidos (ver figura 6.17b y 6.19).

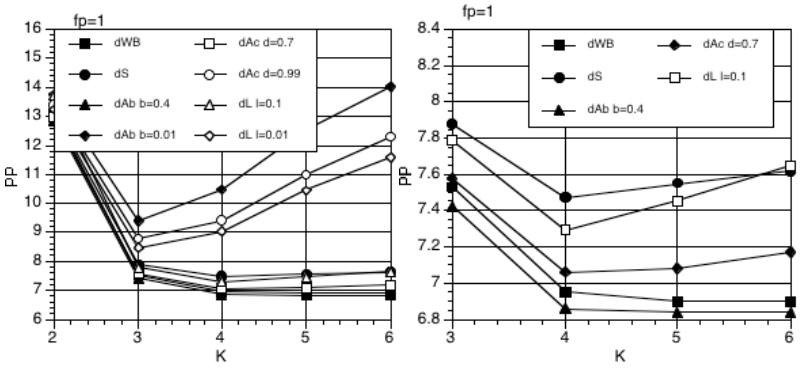


Figura 6.17.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) no podados, *fp*=1, a) en general y b) en detalle para aquellos modelos que presentan resultados más similares.

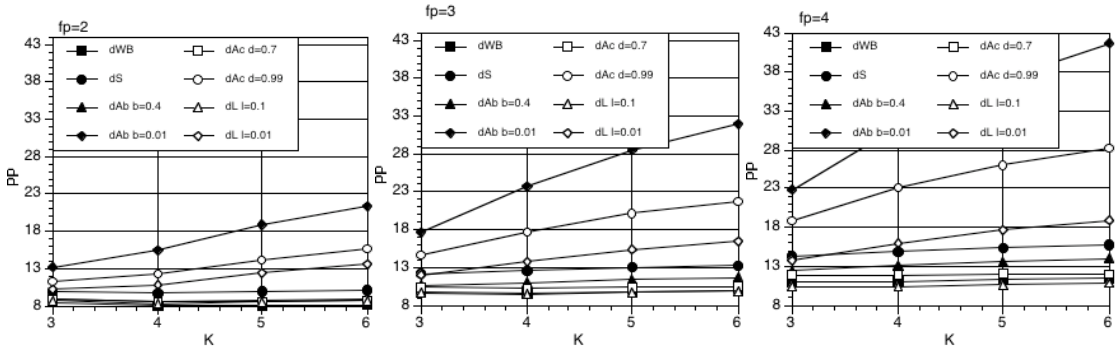


Figura 6.18.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) con distintos factores de poda *fp*>1.

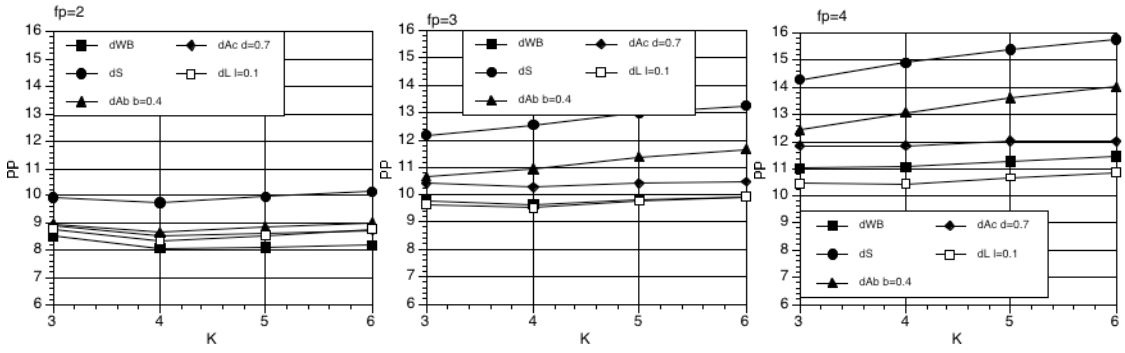


Figura 6.19.- Detalle de los resultados de perplejidad presentados en la figura 6.18.

A medida que se aumenta el valor de *fp* se hacen más notables las diferencias en perplejidad. Claramente los peores resultados se obtienen para los descuentos que proporcionan modelos muy poco suavizados (Absoluto b=0.01, Acotado d=0.99 y Lineal l=0.01). Se observa que para los modelos no podados (*fp*=1) el descuento Lineal l=0.1 no presenta un buen comportamiento (bastante peor que el observado con el descuento de Witten-Bell, descuento Absoluto b=0.4 y descuento Acotado d=0.7). A medida que aumenta el valor del parámetro *fp* el comportamiento del descuento Acotado b=0.4 cambia notablemente y pasa a ser uno de los descuentos que peores resultados de perplejidad presenta, y los mejores resultados se obtienen para descuento Lineal l=0.1.

A continuación, se va a presentar un resumen de los resultados obtenidos cuando los modelos se han integrado en el sistema de reconocimiento aplicando directamente la regla de Bayes. En la figura 6.20a y 6.21 se muestra los %PRE obtenidos en función de K, para los distintos valores de fp y en la figura 6.20b y 6.22 se muestra el número medio de nodos activos (NMNA).

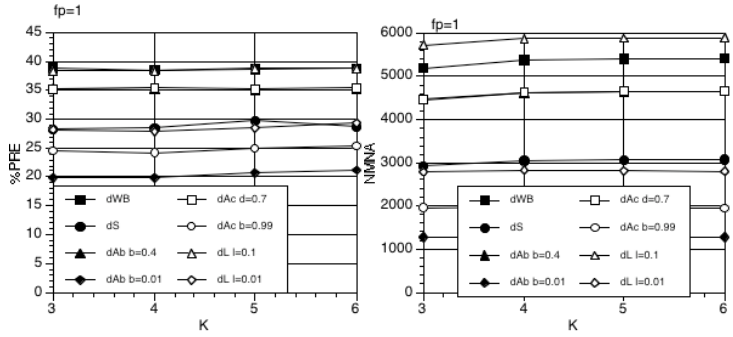


Figura 6.20.- Resultados del %PRE y NMNA obtenidos con los modelos integrados de orden K (K=2, 3, 4, 5 y 6) no podados para α=1.

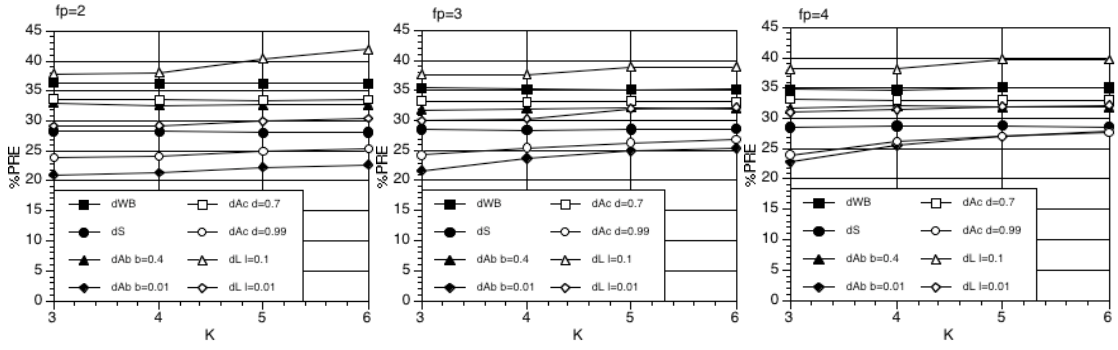


Figura 6.21.- Resultados del %PRE obtenidos con los modelos integrados de orden K (K=2, 3, 4, 5 y 6) para α=1 con distintos valores de *fp*.

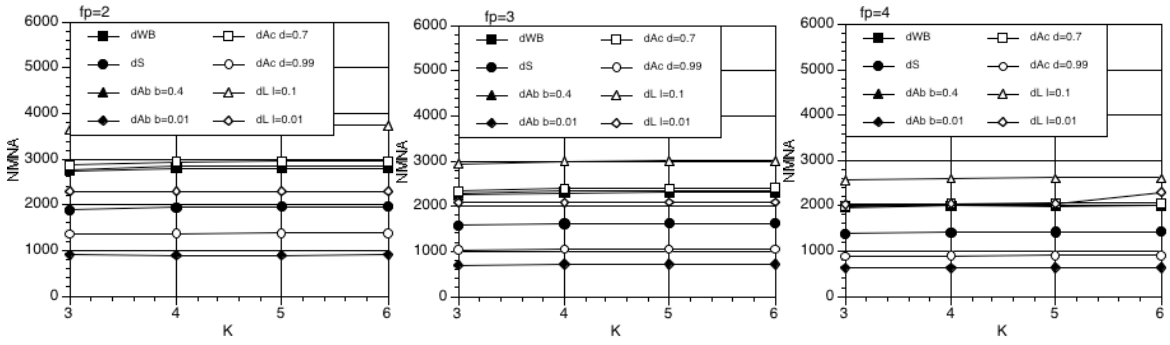


Figura 6.22.- NMNA obtenidos con los modelos integrados de orden K (K=2, 3, 4, 5 y 6) para α=1 con distintos valores de *fp*.

Hay que destacar que los descuentos que presentan peor %PRE, también presentan el NMNA más alto. Aunque algunos valores de la perplejidad van cambiando a medida que aumenta el valor del *fp*, los resultados de decodificación muestran exactamente las mismas relaciones que las observadas para los modelos no podados (*fp*=1) lo que significa que el comportamiento de los distintos suavizados sigue siendo exactamente el mismo a medida que aumenta el factor de poda.

En la figura 6.23 se muestra la relación existente entre la perplejidad y las tasas de error para los distintos modelos podados cuando K=4 (óptimo de decodificación). Estos resultados son totalmente similares a los obtenidos con los modelos no podados (*fp*=1) (ver figura 5.17c). De nuevo se manifiesta la incapacidad de la perplejidad de predecir el comportamiento de los modelos de lenguaje en el sistema de decodificación.

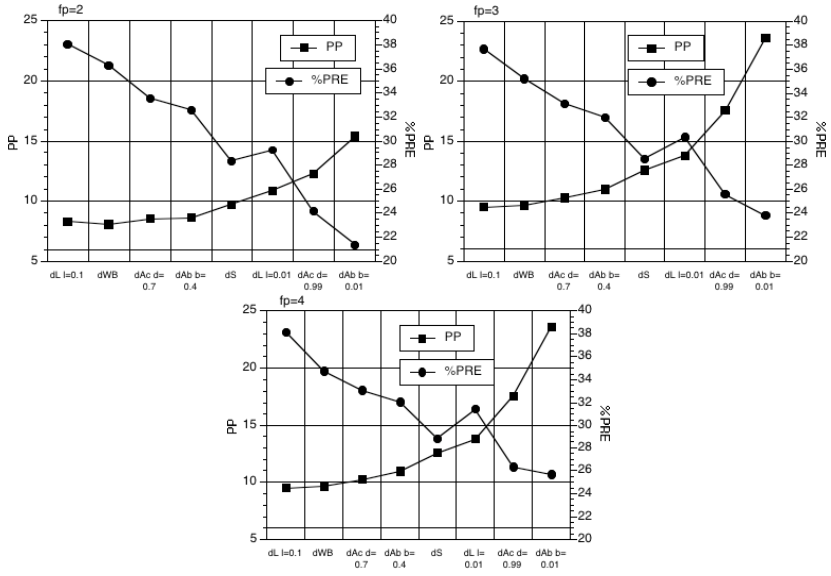


Figura 6.23 .- Resultados de perplejidad (PP) y %PRE obtenidos con el modelo integrado de orden K=4 utilizando los descuentos: Lineal (dL l=0.1), de Witten-Bell (dWB), Acotado (dAc d=0.7), Absoluto (b=0.4), Sencillo (dS), Lineal (dL l=0.01), Acotado (dAc d=0.99) y Absoluto (dAb b=0.4).

6.2.2.- Resultados reescalando las probabilidades del ML

Vamos a ver a continuación el efecto de aplicar la función exponencial de atenuación de las probabilidades del ML sobre los modelos podados y no podados, suavizados con los descuentos presentados previamente.

En primer lugar, vamos a ver el efecto de la aplicación de la ponderación exponencial (α) sobre las distribuciones de probabilidad que han sufrido una menor degradación en perplejidad (descuento Sencillo, descuento Absoluto (b=0.4), descuento Acotado (d=0.7) y descuento Lineal (l=0.1)).

De estos cuatro descuentos el descuento Sencillo es el que produce modelos menos suavizados. Vamos a ver a continuación la evolución de las tasas de error de los modelos integrados de orden K a los que se les ha aplicado el descuento Sencillo con la variación del parámetro α (figura 6.24).

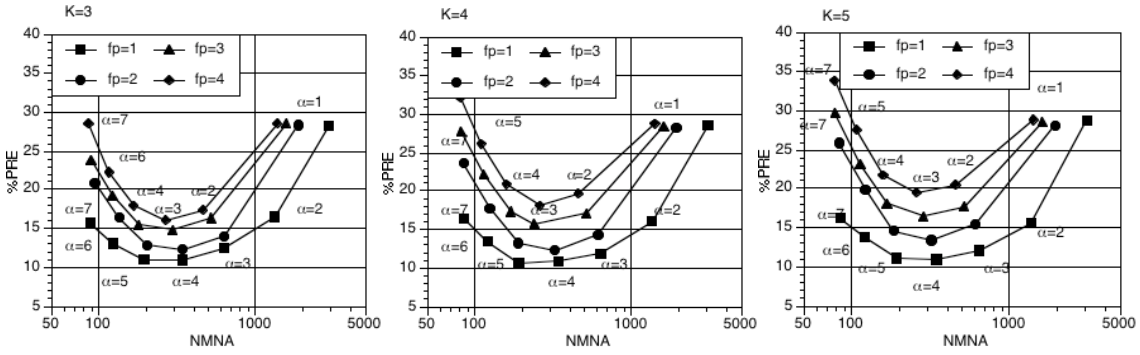


Figura 6.24.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Sencillo, para distintos valores del factor de poda (*fp*) y distintos valores del parámetro α.

En los resultados se puede apreciar que con los modelos podados aumenta el error alrededor del óptimo en números absolutos en todos los casos y el deterioro es muy notable con valores de *fp* altos. En los resultados obtenidos se ve que los modelos podados llegan al óptimo de error con un valor de α bastante menor α=3 frente a α=5 del modelo no podado.

Cuando se ha utilizado el descuento Sencillo se obtiene un modelo claramente menos suavizado que el modelo obtenido con el descuento de Witten-Bell. Esto significa que se reserva menos probabilidad para los eventos no vistos. A medida que se podan los modelos hay más situaciones no vistas, por lo que la distribución está cada vez menos concentrada alrededor de un valor concreto. Esta es la razón por la que es necesario un valor de α menor para alcanzar el óptimo. Al podar los modelos con α=1, se observan un menor NMNA, pero no hay una disminución notable del error. Solo se observa una ligera mejora con respecto al modelo no podado para *fp*=2 (modelo no muy podado y para α<3). Con *fp* =3 y 4 la degradación en el porcentaje de error es muy notable.

A continuación vamos a ver los resultados observados para el descuento Absoluto b=0.4 y descuento Acotado d=0.7 que presentan unos resultados muy similares como puede apreciarse en las siguientes figuras 6.25 y 6.26 respectivamente.

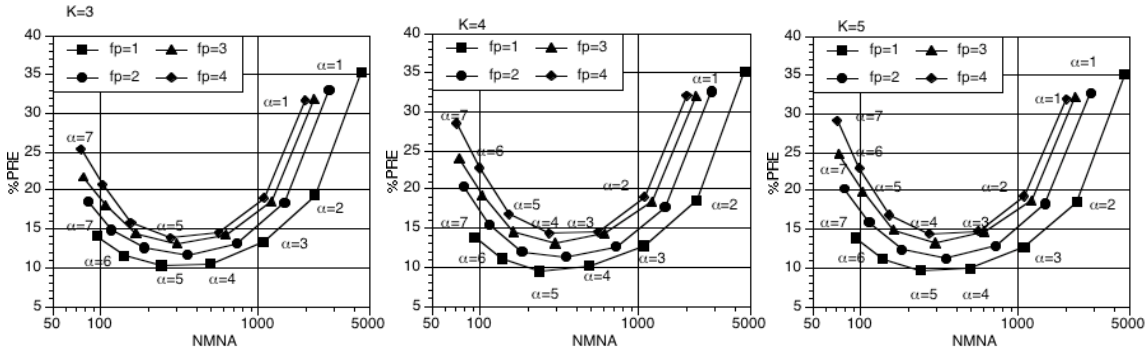


Figura 6.25.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Absoluto b=0.4, para distintos valores del factor de poda (*fp*) y distintos valores del parámetro α.

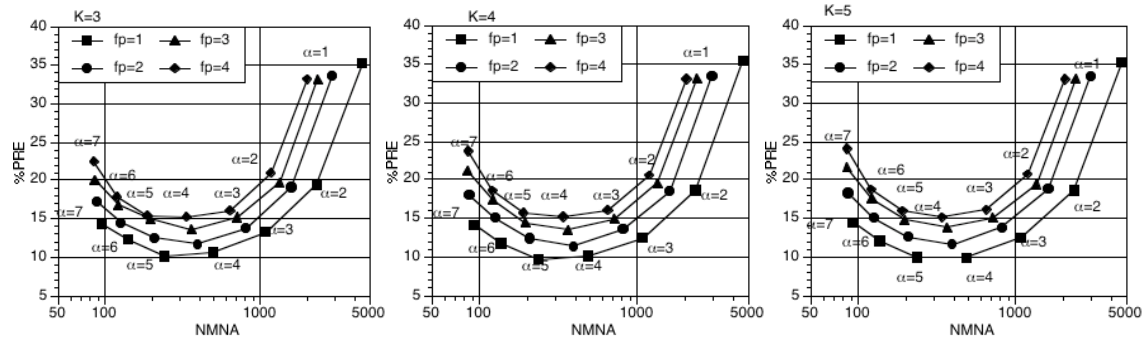


Figura 6.26.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Acotado d=0.7, para distintos valores del factor de poda (*fp*) y distintos valores del parámetro α.

Se observa de nuevo que para los modelos podados el óptimo del error se alcanza para un valor de α inferior (α=4) frente a los modelos no podados (α=5) y se obtienen unos resultados del %PRE ligeramente inferiores. Al contrario que con el descuento de Witten-Bell con α=2 prácticamente solo se disminuye el número de estados pero no hay mejora apreciable en el %PRE.

En la figura 6.27 vamos a ver los resultados obtenidos para el descuento Lineal l=0.1 que es el modelo que menos degradación ha sufrido en los valores de perplejidad.

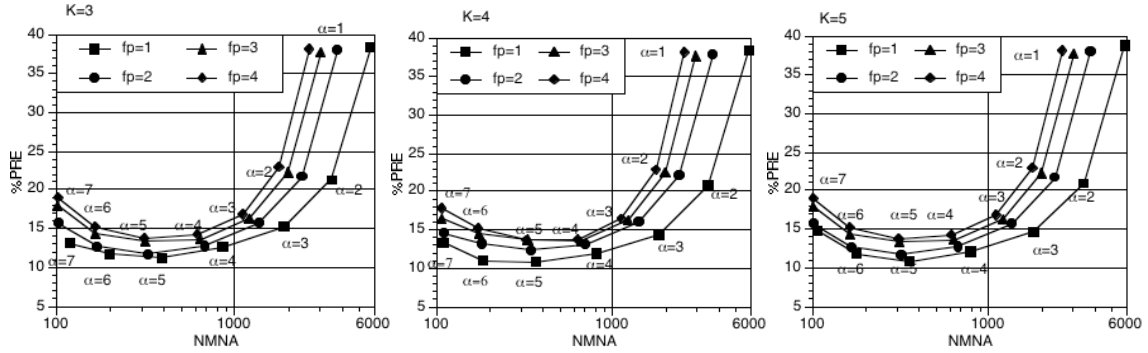


Figura 6.27.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Lineal l=0.1, para distintos valores del factor de poda (*fp*) y distintos valores del parámetro α.

Para el descuento Lineal l=0.1 se observa un comportamiento algo distinto a los observados hasta el momento con los otros suavizados. El hecho de podar los modelos no significa que se necesite un valor de α distinto (inferior) para alcanzar las tasas de error óptima. No se observa una excesiva degradación al usar los modelos podados. Hay que señalar que este tipo de suavizado para los modelos no podados en general no es muy bueno porque resta probabilidad por igual tanto a los eventos observados muchas veces como pocas. Por tanto, al podar los modelos aunque el descuento total es menor como se realiza bastante descuento sobre los eventos observados muchas veces no es un cambio tan radical como el obtenido con otros descuentos (las cuentas más altas a las que se les restaba la mayor probabilidad no han sido eliminadas del modelo y se les sigue restando la mayor probabilidad).

Se observa que el valor de α con el que se obtiene el optimo se mantiene igual a 6 lo cual indica que la distribución mantiene la distribución de probabilidad concentrada en un rango concreto de valores con la poda. También se ve que para α=1 el %PRE se mantiene y lo que disminuye es el número medio de estados En resumen, el descuento Lineal se mantiene sin una excesiva degradación con los modelos podados.

A continuación vamos a ver los resultados que se obtienen sobre aquellos suavizados que sufren una gran degradación en perplejidad. En la figura 6.28 se muestran los resultados para el descuento Absoluto b=0.01 que es la mayor degradación sufre. En la figura 6.29 se muestran los resultados para descuento Absoluto b=0.01 y en la figura 6.30 para el descuento Lineal l=0.01 que es el que menor degradación en perplejidad ha sufrido de los tres.

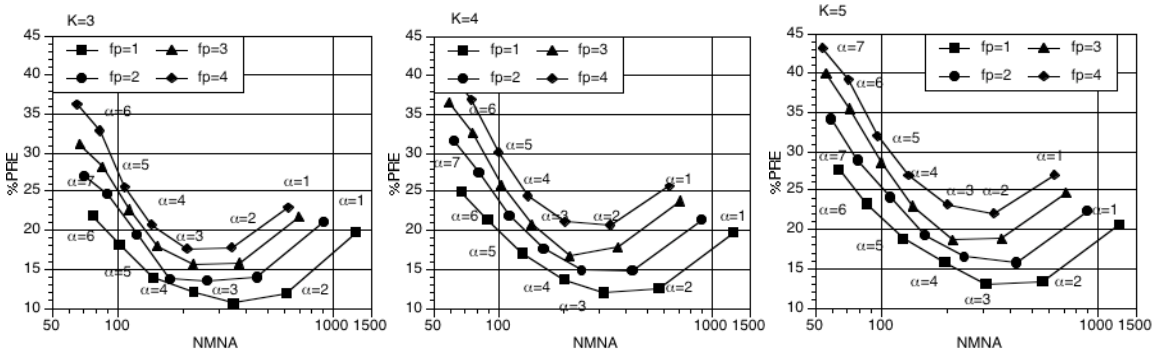


Figura 6.28.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Absoluto b=0.01, para distintos valores del factor de poda (*fp*) y distintos valores del parámetro α.

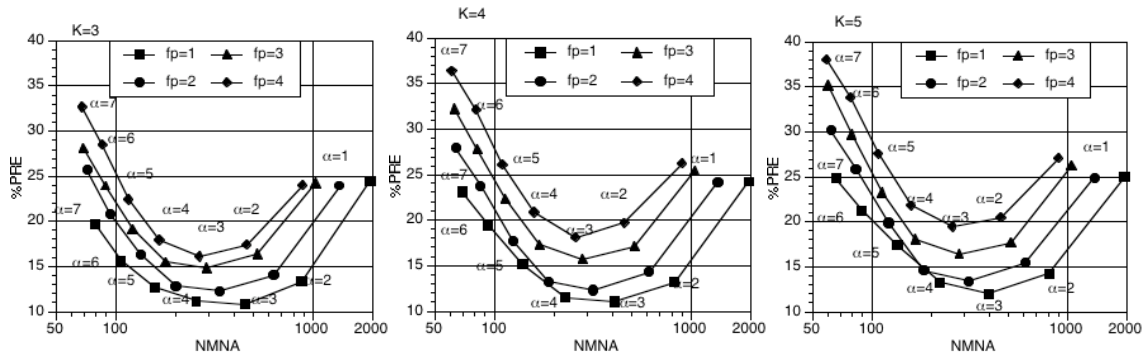


Figura 6.29.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Acotado d=0.99, para distintos valores del factor de poda (*fp*) y distintos valores del parámetro α.

En general para los modelos no podados los mejores resultados se obtenían para α=4. Con los modelos podados el óptimo se logra para α=3, aunque este valor óptimo en cada caso está totalmente alejado del óptimo real sobre todo para el descuento Absoluto b=0.01. El descuento Lineal l=0.01 alcanza el óptimo para α=4, obteniéndose el mismo efecto que el observado para el descuento Lineal l=0.1.

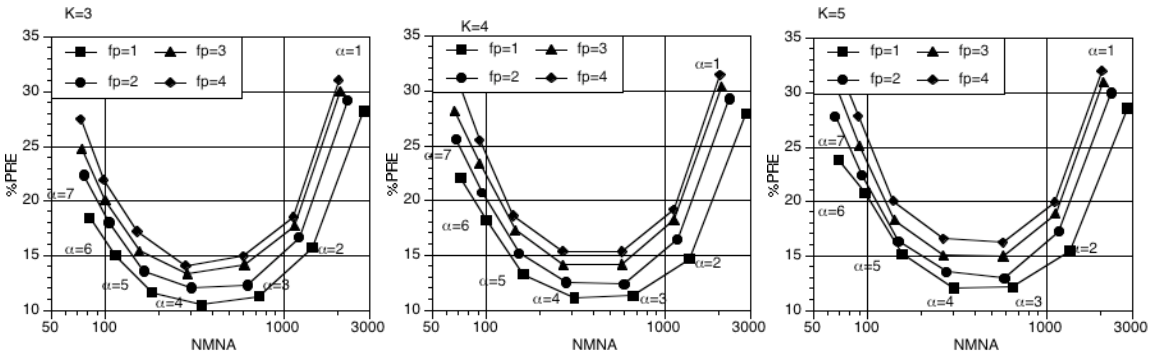


Figura 6.30.- Evaluación de los diferentes modelos integrados de orden K (K=3, 4, 5) suavizados con el descuento Lineal l=0.01, para distintos valores del factor de poda (*fp*) y distintos valores del parámetro α.

6.2.2.1.- Resumen de resultados (α>1).

En primer lugar se presenta una comparación de los resultados de decodificación obtenidos cuando los modelos integrados de orden K, se integran en el sistema de RAH y se introduce la ponderación de las probabilidades del ML mediante la utilización de la función exponencial α.

Los resultados se presentan para K=4 ya que es con el valor con el que se obtienen los resultados óptimos de decodificación, aunque con diferencias muy poco significativas sobre el restos de los modelos con distintos valores de K. En la figura 6.31 se presentan los resultados para los modelos no podados (*fp*=1) y en la figura 6.32 para los modelos podados (*fp*>1). La figura 6.33 muestra en detalle los resultados de la figura 6.32 alrededor del óptimo.

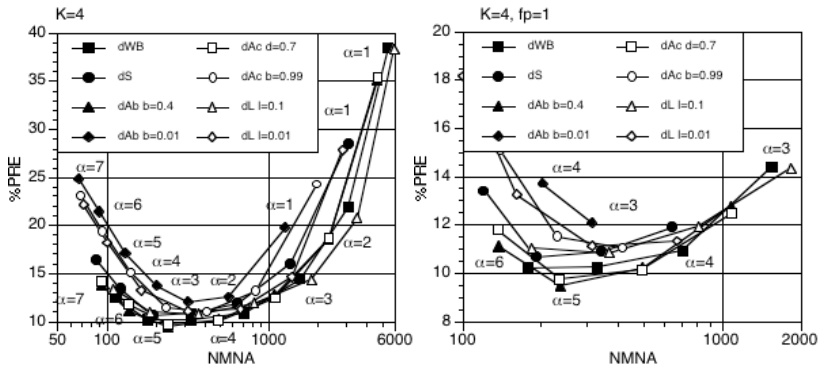


Figura 6.31.- Resultados de %PRE obtenidos cuando se integra en el sistema de RAH el modelo integrado de orden K=4 y *fp*=1 A la derecha se muestran los resultados más en detalle.

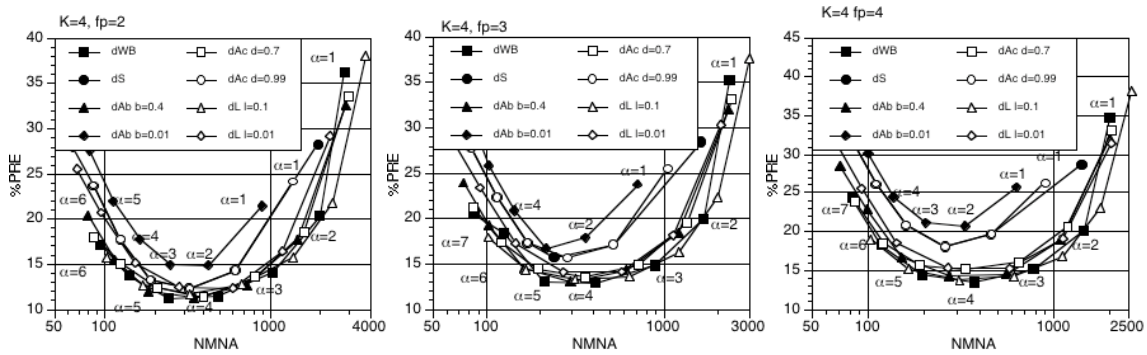


Figura 6.32.- Resultados de %PRE obtenidos cuando se integra en el sistema de RAH el modelo integrado de orden K=4 y distintos valores de *fp*>1.

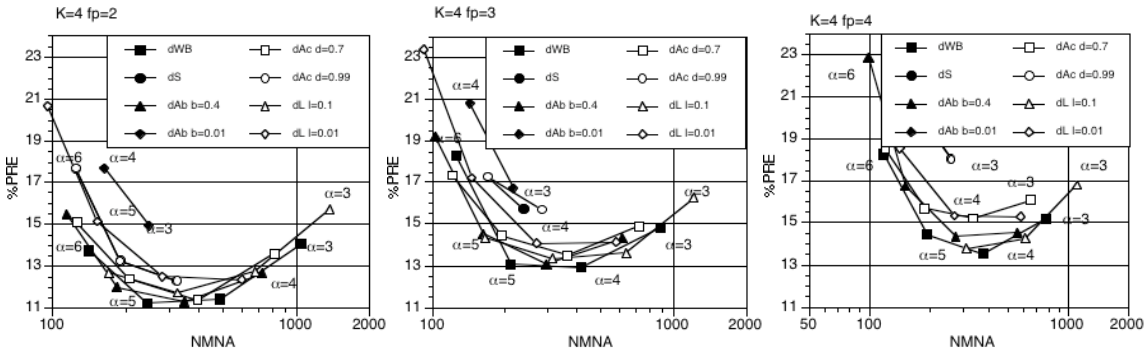


Figura 6.33.- Detalle de los resultados de la figura 6.32.

El descuento de Witten-Bell, descuento Acotado d=0.7, descuento Absoluto b=0.4 y descuento Lineal l=0.1 presentan los mejores resultados. Es decir se produce la menor degradación en números absolutos del %PRE tras podar los modelos. Al haber menos estados hay un descuento menor lo que implica que hay menor masa de probabilidad para ser repartida entre mayor número de estados. Esto implica una distribución de probabilidad un poco más dispersa en el rango de valores sin que llegue a ser perjudicial.

Sin embargo el uso del descuento Sencillo, descuento Lineal l=0.01, descuento Acotado d=0.99 y descuento Absoluto b=0.01 degrada bastante los resultados. En general son técnicas basadas en restar poca probabilidad a los elementos que se han visto muchas veces y cuando se trabaja con modelos podados cada vez hay mas situaciones de este tipo. Por tanto, como a

medida que aumenta el valor del *fp* sólo quedan los elementos vistos muchas veces, la probabilidad final a repartir entre los elementos no vistos es muy pequeña.

Una vez visto los resultados anteriores se puede extraer la siguiente conclusión: si se quiere trabajar con modelos podados para ahorrar memoria habrá que usar técnicas que proporcionen relativamente alta probabilidad a los elementos no vistos porque de lo contrario se degradan bastante los resultados.

En la figura 6.34 se presentan los resultados de PP y %PRE para los distintos modelos podados (*fp*>1) cuando K=4. Los resultados para los modelos no podados (*fp*>1) se presentaron en la figura 5.24c.

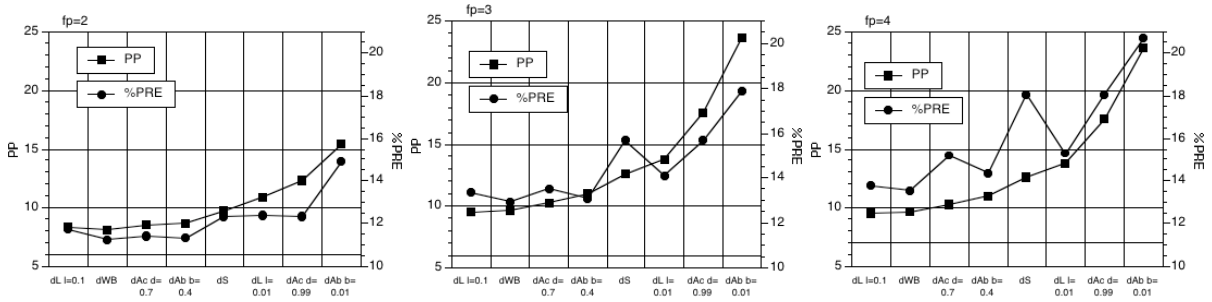


Figura 6.34 .- Resultados de perplejidad (PP) y %PRE obtenidos con el modelo integrados de orden K=4 (K=2...4) utilizando distintos descuentos: Lineal (dL l=0.1), de Witten-Bell (dWB), Acotado (dAc d=0.7), Absoluto (b=0.4), Sencillo (dS), Lineal (dL l=0.01), Acotado (dAc d=0.99) y Absoluto (dAb b=0.4).

De nuevo se manifiesta que la perplejidad no siempre es capaz de determinar cuál es el descuento que mejores resultados proporciona en el sistema completo de reconocimiento. De hecho, a medida que los modelos están más podados los resultados de decodificación para el descuento Acotado d=0.7 y sobre todo para el descuento Sencillo se degradan mucho, y este hecho no se manifiesta en los resultados de perplejidad. Evidentemente la mejor evaluación de los modelos de lenguaje resulta de observar su funcionamiento dentro del sistema integrado.

Capítulo 7

Conclusiones y desarrollos futuros

Una de las etapas fundamentales en un sistema de RAH es la incorporación de conocimiento sintáctico y/o semántico mediante un Modelo de Lenguaje (ML). Para el desarrollo de tal ML se pueden utilizar dos aproximaciones fundamentalmente: estadística o gramatical. Tradicionalmente se han utilizado modelos estadísticos (N-gramas) porque el uso de gramáticas ha sido considerado como una aproximación muy difícil para ser integrada junto con los modelos acústicos en los sistemas de RAH (Segarra, 1993). Uno de los principales objetivos de este trabajo ha sido aplicar una aproximación gramatical al modelado de lenguaje, basada en la utilización de una subclase de las gramáticas regulares, conocidas como gramáticas k-Explorables en Sentido Estricto (k-EE).

Además, la construcción y evaluación de los ML se ha realizado tradicionalmente de forma independiente al sistema de decodificación para el que habían sido construidos. En este trabajo la evaluación de cualquier modificación practicada a los ML se ha realizado en base a resultados de reconocimiento, que se han comparado con las medidas de evaluación clásicas, perplejidad del corpus de test.

Con la integración directa de las probabilidades acústicas y de ML no se logran los resultados de decodificación óptimos. En este trabajo se ha estudiado y evaluado los heurísticos que se utilizan tradicionalmente (ponderación lineal y exponencial sobre las probabilidades del acústicas y de ML) para mejorar el comportamiento del sistema, junto con nuevas propuestas (aplicación de una función escalón, combinación de heurísticos, etc.).

El uso de las gramáticas k-EE permite integrar en un sólo modelo K autómatas k-EE deterministas (k=1...K) y la técnica de suavizado por back-off. La formalización de este modelo ha sido también objetivo del trabajo. Además el suavizado de los ML es fundamental para evitar las probabilidades igual a cero de las combinaciones de palabras que no aparecen en las muestras de entrenamiento. En este trabajo se han evaluado, dentro del sistema de RAH, varios descuentos usando la técnica de back-off sintáctico: propuestas clásicas (Witten-Bell, Absoluto y Lineal) y nuevas propuestas (Sencillo y Acotado).

Para reducir el tamaño de los modelos se ha evaluado también el efecto que tiene en decodificación la poda de los modelos k-EE suavizados con los distintos descuentos evaluados

previamente. Se compara el diferente comportamiento de los modelos podados frente a los no podados.

7.1.- Conclusiones y aportaciones

Las conclusiones y aportaciones de este trabajo se pueden resumir como sigue:

1. Revisión de las diferentes propuestas de modelización del lenguaje aplicadas al Reconocimiento Automático del Habla.
2. Definición de un modelo integrado para representar el ML en el sistema de RAH utilizando las gramáticas k-Explorables en Sentido Estricto. A partir de un esquema de construcción recursivo se obtiene un único modelo que integra K autómatas k-EE (k=1...K) y la técnica de suavizado (back-off sintáctico) en su estructura. Para permitir la integración con los modelos acústicos, se propone una representación eficiente de la estructura del modelo en la que sólo las situaciones vistas en entrenamiento necesitan ser explícitamente desarrolladas.
3. Se comprueba la viabilidad del uso de gramáticas regulares para generar de forma automática los ML necesarios en un sistema de reconocimiento de discurso continuo. Bajo la aproximación sintáctica todo el sistema consiste en un único autómata integrando el conocimiento sintáctico, léxico y fonético de acuerdo a la regla de Bayes. Además, se ha demostrado la fácil integración de los modelos K-EE con los modelos acústicos, incluso para valores altos de K.
4. Puesta a punto del prototipo de RAH, TORNASOL:

- Configuración acústica:

  - Los modelos acústicos semicontinuos aportan ligeras mejoras enmascaradas por el importante incremento del coste computacional, por lo que se propone el uso de modelos acústicos discretos con 4 *codebooks*.

  - El uso de unidades acústicas dependientes del contexto logra decrementar muy ligeramente la tasa de error por lo que se usan principalmente unidades incontextuales.

- Configuración del sistema integrado:

  - Se utilizan modelos de lenguaje gramaticales: modelos integrando K modelos k-EE con K=2...6 y una representación del léxico lineal. Se realiza una implementación eficiente del módulo de búsqueda integrando los autómatas k-EE con los modelos acústicos utilizando el algoritmo de Viterbi y limitando la búsqueda a un haz.

- Se comprueba la necesidad de los parámetros de ponderación de las distribuciones de probabilidad acústicas o del ML para lograr los porcentajes óptimos de error y de tiempos de decodificación.
- La búsqueda de los valores de dichos parámetros de forma teórica no resulta factible ya que se trata de heurísticos que dependen de la tarea y de las relaciones entre las condiciones

experimentales de integración (parámetro que regula el ancho de haz, número de componentes acústicas, técnicas de suavizado, etc.).

5. Tras el estudio de diferentes métodos de descuento (propuestas clásicas: descuentos de Witten-Bell, Absoluto, Lineal y nuevas propuestas descuentos Acotado y Sencillo) se comprueba que los métodos que proporcionan modelos menos suavizados (los que reservan menos probabilidades para los eventos vistos) son los que mejores resultados proporcionan (tasas de error y tiempos de decodificación inferiores), en contra de los resultados obtenidos utilizando como medida la perplejidad.

- Se establece la fuerte relación existente entre los métodos de descuento y el factor de ponderación de las probabilidades del ML necesario para alcanzar el óptimo de reconocimiento. Si se parte de una distribución de probabilidad menos suavizada, simplemente se necesita un valor del parámetro de ponderación de las probabilidades del ML inferior para alcanzar el óptimo. En las tareas bajo estudio, con el valor del parámetro adecuado es posible alcanzar prácticamente los mismos resultados independientemente de la técnica de descuento que se utilice.

6. El uso de modelos K-EE podados supone, además de un ahorro importante de memoria, un decremento de las tasas de error y del tiempo de decodificación ya que la técnica de búsqueda en haz necesita mantener un menor número de caminos activos.

- Cuando se obtienen los valores óptimos del parámetro de ponderación de las probabilidades del ML para los distintos métodos de descuento se observa que las diferencias del porcentaje de error alrededor del óptimo son más significativas que las observadas para los modelos no podados. De hecho, con los modelos no podados se necesita hacer uso de técnicas de descuento que proporcionen modelos bastante suavizados para lograr resultados de decodificación próximos a los obtenidos con los modelos no podados.

7. Tanto si se trabaja con los modelos podados o no podados, la perplejidad no parece ser la medida más adecuada para predecir el comportamiento de los ML en el sistema de RAH puesto que en la mayor parte de los resultados experimentales presentados, un descenso de los valores de perplejidad no significa un mejor resultado de decodificación. De hecho, los resultados de decodificación son más dependientes de los valores de los parámetros de ajuste entre modelos acústicos y de ML. La perplejidad permite discriminar entre la bondad de los modelos, sólo cuando las diferencias en los resultados obtenidos son realmente significativas y se trabaja con los valores adecuados de los parámetros de ponderación. En estos casos se observan también las diferencias en los resultados de decodificación pero mucho más atenuadas.

7.2.- Líneas de trabajo futuras

El sistema de RAH TORNASOL ha sido desarrollado para trabajar con aplicaciones de vocabularios grandes, ya que la implementación desarrollada con la aproximación gramatical permite reducir el espacio de almacenamiento de la información adecuadamente. La mayor parte de los resultados presentados en esta memoria se han obtenido sobre una tarea de talla media (BDGEO), aunque también se han realizado pruebas comparativas sobre una tarea de vocabulario menor para validar algunas de las conclusiones obtenidas. En el momento en que dispongamos de una tarea de mayor vocabulario para la lengua castellana, se podrían evaluar de nuevo alguno de los descuentos aplicados para obtener los modelos K-EE suavizados podados y no podados, teniendo en cuenta que alguno de los resultados obtenidos pueden ser en cierto modo dependientes de la tarea.

Por otra parte los ML gramaticales se han construido para modelar la tarea desde el punto de vista sintáctico. Una nueva vía de actuación sería construir los modelos k-EE de clases de palabras y utilizarlos como un nuevo autómata de grado superior a ser integrado en el sistema de RAH.

Dentro del proyecto titulado "Desarrollo de un Sistema de Diálogo para Habla Espontánea en un dominio Semánticamente Restringido" (TIC98-0423-C06-03) en el que participan 6 grupos de investigación españoles entre ellos el Grupo de Reconocimiento Automático del Habla de la Universidad del País Vasco se está adquiriendo un corpus en Castellano para la construcción de un sistema de diálogo. En concreto, nuestro grupo de investigación está interesado en la caracterización y procesamiento de los fenómenos de habla espontánea que aparece en los sistemas de diálogo. En primer lugar, una vez etiquetadas las disfluencias, se abordará el reconocimiento de las sentencias directamente usando los modelos de lenguaje presentados en esta memoria, lo cual nos permitirá valorar las diferencias y limitaciones existentes. Posteriormente, y gracias al esquema de construcción recursivo de los modelos k-EE se podría abordar la integración de los modelos de lenguaje k-EE a distintos niveles (sintáctico, semántico, etc).

Referencias bibliográficas

Adda, G. Jardino, M. Gauvain, J.L. (1999). “Language modeling for broadcast news transcription”. *Procc of EUROSPEECH´99*. Vol 4. pp 1759-1762.

Aibar, P. (1997) “Diseño de un decodificador acústico-fonético mediante una aproximación basada en distancias”. *Tesis doctoral,* Universidad Politécnica de Valencia.

Austin, S. Schwarz, R. Placeway. P. (1991). "The Fordward-Backward Search Algorithm". *Procc. ICASSP-91*. pp 697-700.

Bahl, L.R. Jelinek, F. Mercer, R. (1983). “A maximun Likelihood Approach to Continuous Speech Recognition”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol PAMI-5 N-2 pp 179-190.

Bahl, L.R. Gopalakrishnan, P. S. Kanevsky, D S. Nahamoo. D (1989) "Matrix Fast Match: A Fast Method for Identifying a Short List of Candidate Words for Decoding". *Procc. ICASSP*. Vol I. pp 345-348

Bahl, L.R. Bakis, R. , Jelinel, F. Mercer L.R. (1990) “Language-model/acoustic channel balance mechanism”. *IBM Technical Disclouse Bulletin*, 23(7B). pp 3464-3465, Dec,.

Bahl, L.R. de Souza, P. V. Gopalakrishnan, P.S. Nahamoo, D. Picheny. M.A. (1992) "A Fast Match For Continuous Speech Recognition using Allophonic Models". *Procc. ICASSP´92*. Vol I. pp. 17-20.

Bahl, L. R. de Genaro S. V. Gopalakrishnan, P.S. Mercer L.R. (1993). "A Fast Approximate Acoustic Match for Large Vocabulary Speech Recognition". *IEEE Transactions on Speech and Audio Processing,* Vol I. No 1. January

Bahl, L.R. Souza, V.P. Gopalakrishnan, P.S. Nahamoo,D. Picheny M.A. (1994) “Decision Trees for Phonological Rules in Continuous Speech Recognition”. *Procc IEEE ICASSP´94,* pp 533-536

Baker, J.K. (1975). “Stochastic Modelling as a Means of Automatic Speech Recognition”. Ph. D. *Thesis in Dept. of Computer Science fo Carnegie- Mellon University.*

Baker, J.K. (1975). “The Dragon System: an overview”. *IEEE Transactions an Automatic Speech and Sygnal Processing*. Vol 23. n 1. pp 24-29. February.

Baker, J.K. (1979). “Trainable Grammars for Speech Recognition”. *Speech Communications Papers for the 97th Meeting of the Acoustical Society of America.* Klatt & Wolf. Eds. pp 547-550

Baum. L.E. (1971). “Inequalities”, *capítulo* “An inequality and associated maximization techniques in statistical estimation for probabilistic functions of Markov processes”. pp 1-8. *Academic Press.*

Blah, R.E. (1987). “Principles of Practice of Information Theory”. *Addison-Wesley.*

Bonafonte, A. (1995). “Comprensión del Habla en Tareas semánticamente restringidas”. Tesis doctoral, Barcelona.

Bonafonte, A. Mariño, J.B. Lleida, E. (1995) "Semantic decoding of speech in constrained domain". *Proceedings of EUROSPEECH´95* Vol 1. pp 559-562.

Bonafonte, A. Mariño, J.B. Nogueiras, A. (1996). "Sethos- the UPC understanding system". *Procc of ICSLP96*. Philadelphia.

Bonafonte, A. Mariño, J.B.(1996). "Language modelling using X-gramas". *Procc of ICSLP96*. Philadelphia.

Bordel, G., Torres, I. Vidal, E. (1994). "Back-off smoothing in a syntactic approach to Language Modelling". *Proc. International Conference on Speech and Language Processing*, pp. 851-854.

Bordel, G., Torres, I. Vidal, E. (1995). "QWI: A Method for Improved Smoothing in Language Modelling". *Procc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Detroit (Michigan, USA).

Bordel, G. (1996) "Aprendizaje Automático de Modelos k-Explorables estocásticos en Reconocimiento Automático del Habla". *Tesis Doctoral*. Departamento de Electricidad y Electrónica (UPV/EHU). Marzo

Bordel, G., Varona, A. Torres, I. (1997). "K-TLSS(S) Language Models for Speech Recognition". *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol 2, pp 819-822.

Bourlard, H. (1995). "Towards increasing speech recognition error rates," *Proceedings of EUROSPEECH 95*. pp 883-894

Bourlard, H., Hermansky, H, Morgan, N. (1996). "Towards increasing speech recognition error rates." *Speech Communication* 18, pp 205-231.

Brugnara, F. Federico, M. (1997). "Dynamic Language Models for Interactive Speech Applications". *Procc EUROSPEECH 97*, pp 2751-2754.

Casacuberta, F. Vidal, E. (1987). "Reconocimiento Automático del Habla". *Ed. Marcobo.*

Casacuberta, F. Mariño, J.B. Torres, I. Lleida (1995-1998). "Construcción de sistemas robustos de reconocimiento de habla continua y gran vocabulario mediante técnicas de aprendizaje automático". *Proyecto concedido por la CICYT (TIC95-0884-C04-03).*

Castro, M.J , Llorents, D. Sanchez, J.A., Casacuberta, F., Aibar, P. Segarra, E. (1999) "A Fast Version of the Atros System". *Procc of EUROSPEECH, 99*. Vol 3. pp 1239-1242.

Castro, M.J. (1999) "Modelado acústico de unidades subléxicas mediante una aproximación basada en métodos estructurales-conexionistas". *Tesis doctoral*. Universidad politécnica de Valencia.

Castro, M.J. Casacuberta, F. (1999) "Hybrid connectionist-structural acoustical modeling". *Proceedings of the VIII Symposium Nacional de Reconocimiento de Formas y Análisis de Imágenes*. Vol I pp 69-76. Bilbao

Chandhuri, R. and Booth, T. L. (1986). "Approximating Grammar Probabilities: Solution of a Conjecture," *Journal ACM*, vol 33, nº 4, pp. 702-705.

Chen, S. F., Goodman, J. (1999). "An empirical study of smoothing techniques for language modeling". *Computer Speech and Language*. Vol 13. pp359-394

Chollet, G. (1995). "Evaluation of ASR systems, algorithms and databases". *In Speech Recognition and Coding: New advances and Trends*. pp 32-40. *Springer-Verlag*

Clarkson, P. Rosenfeld, R. (1997a) "Statistical language modeling using the CMU-CAMBRIDGE toolkit", *Proceedings of EUROSPEECH 97* pp- 2707-2710.

Clarkson, P.R., Robinson, A. J. (1997b). " Language model adaption using mixtures and an exponentially decaying cache". *In IEEE ICASSP-97* pp 799-802, Munich, Germanny.

Clarkson, P. Robinson, T. (1999) ." Towards improved language model evaluation measures". *Procc of EUROSPEECH.99*. Vol 5. pp 1927-1930.

Cravero, M. Pieraccini, R. Raineri, F. (1986). "Definition and Evaluation of Phonetic units for Speech Recognition by Hidden Markov Models". *Procc. IEEE Int. Conf. Acoust. Speech and Signal processing*. pp 2235-2238

Cremelie,N. Martens, J.P. (1993) "Heuristic Search Methods for a Segment Based Continuous Speech Recognizer". *New Advances and Trends in Speech Recognition and Coding. NATO-ASI*. pp 1-4. Granada,

Davis, S.H.,. Mermelstein, P. (1980) "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *IEEE Trans. on ASSP*, vol. 28, n. 4, pp. 357-366. August.

Deligne, S., Bimbot, F. (1995) "Language modelling by variable length sequences: theoretical formulation and evaluation of multigrams". *Procc ICASSP*. Vol 1 pp 169-172. Detroit

Deroault, A. M. Merialdo, B. "Natural language Modelling for phoneme-to-text transcription". *Vol PAMI*-8, n 6, pp 742-749, noviembre.

Diaz, J. E., Rubio, A. J., Peinado, A. M., Segarra, E., Prieto, N. and Casacuberta, F. (1993); "Development of Task Oriented Spanish Speech Corpora," *Proceedings of EUROSPEECH 93*

Digilakis, V. Tsakalidis, S. Neumeyer, L. (1999). "Reviving discrete HMMS. The myth about the superiority of continuous HMMs". *Procc of EUROSPEECH´99*. Vol 6. pp 2463-2466.

Duda, R. O. Hart. P.E. (1973). "Pattern classification and scene analysis". John Wileyan Sons.

Dupont, P. (1993). "Efficient integration of context-free grammars based languaje models in continuous speech recognition". *In new Advances and Trends in Speech Recognition and Coding*. pp 179-189. NATO-ASI

Feldman, J.A., Lakoff, A., Stolcke, A., Weber, S.H., (1990) "Miniature Language Acquisition: a touch-stone for cognitive science". *Technical Report, Tr-90-009. ICSI*, Berkeley, California. April

Ferretti, M. Maltese, G. Scarci, S. (1990), ."Measuring information provided by language model and acoustic model in probabilistic speech recognition: Theory and experimental results". *Speech Communication* 9. 531-539.

Flanagan, J.L. (1972). "Speech Analysis, Synthesis, and Perception". Segunda Edición. New York. *Springer-Verlag*.

Fu, K.S. Huang, X. (1972). "Stochastic grammars and language". *International Journal of Computer and Information Sciences*. Vol 1. n. 2.

Fu, K.S. (1974). "Syntactic Methods in Pattern Recognition". *Serie Mathematics in Science and Engeneering. Academic Press. N. Y.*

Fu, K.S. (1982). "Syntactic Pattern Recognition an Applications". *Prentice Hall.*

Galiano, I.(1992) "Decodificación acústico-fonética del habla continua mediante metodología de Inferencia Gramatical basada en Generadores Mórficos". *Tesis Doctoral, DSIC*. Universidad Politécnica de Valencia.

García, P. and Vidal, E. (1990). "Inference of k-testable languages in the strict sense and application to syntactic pattern recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, nº 9, pp. 920-925

Gildea, D. Hofmam, T. (1999). "Topic-Based Language models using EM". *Procc of EUROSPEECH´99*. Vol 5. pp 2167-2170.

Gold, E. M. (1967). "Languaje Identification in the Limit". *Information and Control*. Vol 10. pp 447-474

Gonzalez, R. C. Thomason, M. G. (1978). "Syntactic Pattern Recognition: An Intoduction". *Addison-Wesley*.

GuoDong, Z, KimTeng, L. (1999)., "Interpolation of n-gram and mutual-information based trigger pair language models for Mandarin speech recognition". *Computer Speech and Language*. 13, 125-141.

Hopcroft, J. Ullman, J (1979) "Introduction to automata theory, languages and computation". *Ed. Addison Wesley*. 2988.

Hu, J. Turi, W. Brown, M. K. (1997). " Language modeling using stochastic automata with variable length contexts". *Computer Speech and Language*. Vol 11. pp 1-16

Huang, H., Alleva, F. Hon, H. W.,Hwang, M. Y., Lee, K. F., Rosenfeld, R. (1993) "The SPHINX-II speech recognition system: an overview". *Computer Speech and Language*. 7(2). pp 137-148 Abril

Huang, X, D., Ariki, Y., Jack, MA. (1990) "Hidden Markov Models for speech recognition". *Edinburgh Information Technilogy Series,* Eds S. Michelson and M. Steedman, Edinburgh University Press,

Huang, X. D. Jack, K. A. (1989). "Semi-continuous Hidden Markov Models for Speech Sygnals". *Computer Speech and Languaje*. N 3.pp 239-251

Huang, X.D. (1992). "Phoneme Classification Using Semicontinuos Hidden Markov Models". *IEEE Trans. Acoust. Speech and Sygnal processing*. Vol I. pp 357-360.

Huang, X. Belin, H. Alleva, F. Hwang, M. (1993) "Unified Stochastic Engine (USE) for Speech Recognition". *Procc. ICASSP*. Vol 2 pp. 636-639.

Ito, A. Kohda, M. Ostendorf, M. (1999) ." A new metric for stochastic language model evaluation". *Procc of EUROSPEECH.99*. Vol 4. pp 1591-1594.

Jardino, M. (1996). "Multilingual Stochastic n-gram class language models". *Procc of ICASSP´96*. pp 161-163.

Jelinek, F. (1969) "Fast Sequential Decoding Algorithm Using a Stack". *IBM J. Res. Develop*.pp 675-685. November.

Jelinek, F. (1976) "Continuous Speech Recognition by Statistical Methods". *Proceedings of the IEEE*, Vol. 64, No. 4, pp 532-556. April.

Jelinek, F. Mercer, R. L.(1980). "Interpolated Estimation of Markov Source Parameters form Sparse Data". *Procc. of the Workshop on Pattern Recognition in Practice. E. S.* Gelsema and L. N. Kanal editores, Amsterdam. pp. 381-402.

Jelinek, F. (1985). "Markov source modeling fo text generation". NATO-ASI. In the Impact of Processing Techniques on Communications". *Maritnus Nijhff Eds*. Amsterdan. pp 569-598

Jelinek, F. (1990) "Self-organized language modelling for speech recognition", *in Readings in Speech Recognition, Morgan Kaufmann Publishers, Inc*.

Jelinek, F. Lafferty. J.D. (1991). "Computation of the Probability of Initial Substring Generation by Stochastic Context-Free Grammars". *Computational Linguistic*, Vol 17. n3

Jelinek, J. (1991). "UP FROM TRIGRAMS! the strugle for improved language models". *Procc of EUROSPEECH´91*. Vol III. pp. 1037-1040.

Jelinek, F. (1996). "Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky and N. Morgan". *Speech Communication* 18, pp 242-246.

Juang, B. H. Katagiri, S. (1992). "Discriminative Learning for Minimum Error Classificatin". *IEEE Trans. ASSP*. Vol 40. N0 12. pp3043-3054.

Jurafsky, D. Wooters, C. Tajchman, G. Segal, J, Stolcke, A, Fosler, E. Morgan, N. (1995). "Using context-free grammar as a language model for speech recognition". *In IEEE ICASSP-95* pp 189-192.

Katz, S. M. (1987). "Estimation of Probabilities from Sparce Data for The Language Model Component of a Speech Recognizer". *IEEE Trans. on Acoustics, Speech and Signal Processing,. vol. ASSP-35*, n 3, pp. 400-401

Kenny, P. Hollan, R., Gupta, V. N. Lennig, M Mermelstein, P O´Shaughnessy D (1993). "A*-Admissible Heuristics for Rapid Lexical Access". *IEEE Transactions on Speech and Audio Processing*, Vol. I, No 1. January

Kenny, P. Labute P, Li, Z. O´Shaughnessy. D. (1994) "New Graph search Techniques For Speech Recognition". *Procc. ICASSP´94*. Vol 1. pp 553-556. Australia

Khudanpur, S. Wu, J. (1999) "A maximun entropy language model integrating n-grams and topic dependencies for conversational speech recognition". *Procc of ICASSP´99*. Vol I. pp 553-556.

Kilian, U. Class, F. Kaltenmeier, P. Regel-Brietsmann, P. (1995) "Representation of a Finite state grammar as bigram language model for continuos speech recognition". *Procc of EUROSPEECH. 95*. pp 1241-1244.

Kilian, U. Class, F. (1999) "Combining Syntactical and statistical Language constrains in context-dependent language models for interactive Speech applications". *Procc of EUROSPEECH. 99*. Vol 5. pp 1935-1938.

Kneser, R, S. Ney, H. (1995). "Improved backing-off for m-gram language modeling". *Procc of the IEEE International Conference on Acoustic, Speech and Signal Processing*. Detroit. Vol 1. pp 181-184

Kuhn R. de Mori, R. (1992). "A cache based natural language model for speech recognition". *IEEE Transactions on Pattern Analysis and Machine Inteligence*, 14:570-583.

Kuhn, R. Lazarides. Normandin. I. (1995) "Improving Decision Trees for phonetic modelling", *ICASSP95,* pp, 552-555.

Lee, K, F. (1988) "Large Vocabulary Speaker-Independent continuous Speech Recognition : the SPHINX System". Tesis Doctoral, C -S Dept. CMU,

Lee, K. F. (1989a). "Automatic Speech Recognition. *The development of a SPHINX system*. Kluwer Academic Publishers.

Lee, K. F. (1989b). "Hidden Markov Models: Past, Present and Future". *Procc of EUROSPEECH´89*. pp 148-159. Paris.

Lee, C. H., Rabiner, L, R., Pirtaccini, R,. Wilpon, J, G. (1990) "Acoustic modeling for large vocabulary speech recognition ". *Computer Speech and Language*. Vol 4. pp 127-165

Lee, K, F, Aller, F (1991). "Continuous Speech Recognition", *Ch Advances in Speech Signal processing*.

Levinson, S. E. (1985). "Structural Methods in Automatic Speech Recognition". *Procc. of the IEEE,* Vol-75, N. 11, pp. 1625-1650.

Llisterri, J. (1991). "Criterios para la elaboración de una base de datos para el reconocimiento del Habla en Español". Informe interno del proyecto (TIC91-1488).

Llorents, D. Casacuberta, F. Segarra, E. Sanchez, J.A., Aibar, P. Castro, M.J. (1999) "Acoustic and Syntactical modeling in the ATROS system". *Procc of ICASSP, 99.* Vol II, pp 641-644.

Llorents, D. Casacuberta, F. (1999) "An experimental study of histogram pruning in speech recognition". *Proceedings of the VIII Symposium Nacional de Reconocimiento de Formas y Análisis de Imágenes.* Vol I, pp 5-6. Bilbao

Lopez de Ipiña, K. Rodriguez, L.J. Varona, A. Torres, I. (1999a) "Decision tree-based context dependent sublexical units for Spanish continuous speech recognition tasks". *Proceedings of the VIII Symposium Nacional de Reconocimiento de Formas y Análisis de Imágenes.* Vol I pp 53-60. Bilbao

López de Ipiña, K. Varona, A. Torres, I. Rodriguez, L.J. (1999b)" Decision Trees for Inter-Word Context Dependencies in Spanish Continuous Speech Recognition Tasks". *Procc of EUROSPEECH, 99.* vol 2. pp 899-902

Lowerre, B. Reddy. R. (1980). " The Harpy Speech Understanding System". pp 340-360. from the book, "Trends in Speech Recognition". *Wayne A. Lea. Editor. Englewood Cliffs. N. J. Prentice- Hall.*

Lowerre, B. T. (1976) "The HARPY Speech Recognition System". *Ph. D. Thesis in Dept. of Computer Science fo Carnegie- Mellon University.*

Mariño. J.B. Nogueiras, A. Bonafonte, A. (1997). "The dimiphone: an efficient subword unit for Continuous Speech Recognition". *Procc of EUROSPEECH´97.* pp 1215-1218.

Martin, S. C. Liermann, J. Ney, H. (1997). "Adaptive topic-dependent language modelling using word-based varigrams". *Procc of EUROSPEECH´97.* pp 1447-1450.

Martin, S. Lierman, J. Ney, H. (1998) "Algorithms for bigram and trigram word clustering". *Speech Communications.* 24 pp 19-37.

Martin, S. C. Ney, H. Zaplo. J. (1999) "Smoothig methods in maximum entropy language modeling". *Procc of ICASSP´99.* Vol I. pp 545-548.

Meeter, M. Rohlicek, R. (1993). "Statistical Language Modelling Combining N-gram and Context Free Grammars". *Procc of ICASSP 93.* Vol 2. pp 37-40.

Moreno, A. (1993). "EUROM.1 Spanish Database". *Informe Técnico D6, Document SAM-A/UPC3, Esprit Technology Assesment in Multilingual Applications. Esprit Project 6919.*

Müller, L. Psuka, J. (1999). "Using various Language Model Smoothig techniques for the transcription of a weather Forecast Broadcasted by the Czed radio". *Procc of EUROSPEECH´99.* Vol 4 .pp 1783-1786

Nasr, A. Esteve, Y. Bechet, F. Spriet, T. De Mori, R. (1999). " A Language model combining N-grams and stochastic finite state Automata". *Procc of EUROSPEECH. 99.* Vol 5. pp 2175- 2178.

Ney, H. Essen, U. (1991) "On smoothing Techniques for Bigram-Based Natural language modeling". *Procc of ICASSP.* pp. 825-828

Ney, H. (1991). "Dynamic programming parsing for context-free grammars using the inside-outside algorithm". *Computer, Speech and Language.* pp 237-257.

Ney, H. . Mergel, D. , Noll, A. and Paeseler, A. (1992). "Data driven search organization for continuous speech recognition". *IEEE transactions on signal processing*, Vol 40. No2. February.

Ney. H. (1992a) "A comparative study of two search strategies for conected word recognition: dynamic programming and heuristic search". *IEEE transactions PAMI*, V. 14(5). pp. 586-594.

Ney. H. (1992b) "Stochastic grammars and pattern recognition". In P. Laface and R. *De Mori, editors, Speech Recognition and Understanding. Recent Advances*, pp 319-344.

Ney, H. (1993). "Architecture and Search Strategies for Large-Vocabulary Continuous-Speech Recognition". *New Advances and Trends in Speech Recognition and Coding. NATO-ASI.* Lecures pp 59-84. Granada.

Ney, H. Essen, U. (1993). "Estimating ´Small´ probabilities by leaving-one-out". *Procc of the Eurospeech.* pp 2125-2129.

Ney, H. Aubert. X. (1994a) "A Word Graph Algorithm for large Voabulary, Continuous Speech Recognition". *Procc. ICSLP´94.* Vol 3. pp 1355-1358.Yokohama.

Ney, H. Aubert X. (1994b) "A Word Graph Algorithm for large Voabulary, Continuous Speech Recognition". *Procc. ICSLP´94.* Vol 3. pp 1355-1358.Yokohama.

Ney, H. Essen, U. Kneser, R. (1994c). "On structuring probabilistic dependencies in stochastic language modeling". *Computer Speech and Language.* 8 1:38.

Ney, H., Steinbiss, R. Haeb-Umbach, B. H., Essen, V. (1994b) "An overview of the Philips Research System for large vocabulary continuous speech recognition". *International Journal of Pattern Recognition and Artificial Intelligence.*

Ney, H. Martin, S.C. Wessel, F. (1997) "Statistical Language modelling Using Leaving-One-Out". pp 174-207. In S.Young, G. Bloothooft: "Corpus-Based Methods in Language and Speech Processing", *Kluwer Academic Publishers, Dordrecht*, The Netherlands.

Niesler, T. R. Woodland, P.C. (1999)."Variable-length category n-gram language models". *Computer Speech and Language,* 13 99-124.

Odell, J. (1995) "The use of Context in Large Vocabulary Speech Recognition". *Doctoral Thesis. Cambridge University.* March

Ogawa, A. Takeda, K. Itakura, F. (1998). "Balancing Acoustic and linguistic Probabilities". *In Procc ICASSP´98.* Vol1, pp. 181-185

Ortmanns. S. Eiden, A. Ney, H. (1994). "Improved lexical trie search for large vocabulary speech recognition". *Procc of ICASSP, 94.* Vol II. pp 815-818.

Ortmanns, S. Ney, H. Aubert, X. (1997a). "A word grapth algoritm for large vocabulary continuos Speech Recognition". *Computer Speech and Language.* Vol II. pp 43-72

Ortmanns, S., Eiden, A., Ney, H. and Coenen, N. (1997b). "Look-ahead techniques for fast beam search". *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1783-1786.

Ousterhout, J.K. (1994) "Tcl and Tk toolkit". *Addison-Wesley Reading*, Massachussets

Paeseler, A. Ney, H. (1989). "Continuous Speech recognition using a stochastic language model". *Procc of ICASSP.* pp 719-722.

Partalo, M. Sijercic, Z. (1989). "Comparison of several speech signal feature parameters for automatic speech recognition". *Speech Communications.* Vol 8, n 4. pp 347-353.

Paul, D. B. (1991) "Algorithms for an Optimal A^* Search and Linearizing the search in the Stack Decoder". *Procc. ICASSP-91.* pp 693-696

Paul, D.B (1992) "An Efficient A* Stack Decoder Algorithm for Continuous Speech Recognition with a Stochastic Language Model". *Procc. ICASSP-92,* Vol. I. pp 25-28.

Peñagarikano, M. Bordel, G. Varona, A, López de Ipiña, K. (1999). "Using non-word lexical units in automatic speech understanding". *Procc of ICASSP*. Vol II. pp 621-625.

Picone, J. W. (1993). "Signal Modeling Techniques in Speech Recognition". *Procc of IEEE*. 81-9. pp 1215-1247. September.

Placeway. R. Schwartz, R. Fungs, P. (1993) "The estimation of powerful language models from small and large corpora". *In proceedings of the IEEE ICASSP*. Vol 2. pp 33-36 Minneapolis

Prat, F. Aibar, P. Marzal, A. Vidal, E. (1994). "El problema de la evaluación de un sistema de reconocimiento automático del habla mediante un único valor numérico". *Informe técnico. DSIC II/15/94 DSIC-UPV*

Quilis, A., (1981) "Fonética acústica de la Lengua Española". *Biblioteca Románica Hispánica. Editorial Gredos*. Madrid.

Rabiner, L.R. Schfer, R.W. (1978) "Digital Processing of Speech Signals". *Prentice-Hall Signal Processing Series Alan V. Oppenheim*. Series Editor

Rabiner, L. R., (1989) "A Tutorial on Hidden Markov Models and Selectec Aplications in Speech Recognition". *Procc IEEE*. Vol. 77. No, 2. February 1989

Rabiner, L. R.Juang, B.H. (1989) "Fundamental of speech recognition." New Jersey, *PTR Prentice-Hall*. 1993

Rabiner, L. R. Juang, B.H. (1993). "Fundamentals of Speech Recognition". *Prentice-Hall*.

Riccardi, G. Bocchieri, E. Pieraccini, R. (1995) "Non deterministic stochastic language models for speech recognition". *Procc IEEE ICASSP*. Vol I, pp 237-240. Detroit Michigan.

Riccardi, G. Pieraccini, P. Bocchieri, E. (1996). "Stochastic automata for language modeling". *Computer Speech and Language*. 10, pp 265-293.

Rodriguez, L.J. (1994) "Estudio Comparativo de Varias Representaciones Paramétricas para Reconocimiento Automático del Habla". *Informe interno de Investigación DEE-I/2/94*. Universidad del País Vasco.

Rodriguez, L.J. Torres, I. (1995). "A comparative study of several parametric representations for speech recognition". *VI Spanish Symposium on Pattern Recognition and Image Analysis*. pp 286-292. Cordoba.

Rodriguez, L. J. Torres, I. Alcaide, J. M. Varona, A. López de Ipiña, K. Peñagarikano, M. Bordel, G. (1999a) "An integrated System for Spanish CSR Tasks". *Procc of EUROSPEECH, 99*. vol 2. pp 951-954

Rodriguez, L.J. Torres, I Alcaide, J.M. Varona, A. López de Ipiña, K. Peñagaricano, M. Bordel, G. (1999b) "A new integrated system for the continuous speech recognition of Spanish". *Vol I Proceedings of the VIII Symposium Nacional de Reconocimiento de Formas y Análisis de Imágenes*. pp 93-100. Bilbao

Rosenfeld, R (1996). "A maximum Entropy Approach to Adaptive Statistical language modeling". Computer Speech and Language, Vol 10, No, 3, pp.187-228.

Rosenfeld, R. (1994) "Adaptive statistical language modeling: a maximum entropy approach". *PhD Thesis, Carneige Mellon University*, Boston, MA.

Rosenfeld, R. Agarwal. R, Byrne, B. Iyer, R, Liberman, M. Shriberg, L. Unverferth, Y. Wergyri.D. Vidal.E. (1995) "End-of-workshop report of the LMSI project team at the 1995". *Hopkins workshop on language Modeling* . Valtimore,

Rubio, A. J., Segura, J. C., Garcia, p. Diaz, J.E. (1994); "Automatic Speech Recognition at UGR" *Proceedings in Attificial Inteligence CRIMFORWISS workshop on Progress and Prospects of Speech Researh and Technology.pp 158-165*

Rubio, J.A. Diaz- Verdejo J.E. Garcia, P. Segura, J. C. (1997) "On the influence of frame-asynchronous grammar scoring in a CSR systems". *In Procc. ICASSP´97* Vol2, pp 895-899.

Rulot, H. (1992) "Un algoritmo de Inferencia Gramatical mediante Corrección de Errores". *Tesis doctoral*. Universidad Politécnica de Valencia.

Samuelson, C. Reichl, W (1999). "A class-Based language model for large vocabulary speech recognition extracted from part of speech statistics". *Procc of ICASSP´99*. pp 537-540.

Sánchez, J. A, Benedí J. M. (1997). "Consistency of stochastic context-free grammars from the k-best derivarions based on growth transformation*"*. *IEEE. Trans.Pattern Analysis and Machine Intelligence,* 19(9):1052-1055.

Sánchez, J. A, Benedí J. M. (1999). "Learning of Stochastic Context-Free Grammars by means of Estimation Algorithms". *Procc of EUROSPEECH*, *99*. Vol 4, pp 1799-1802.

Sánchez, J.A. Casacuberta, F. Aibar, P. Llorents, D. Castro, M.J. (1994) "Fast Phoneme loo-ahead in the ATROS system". *Vol I Proceedings of the VIII Symposium Nacional de Reconocimiento de Formas y Análisis de Imágenes*. pp 77-84. Bilbao

Sanchís, E. (1994) "Modelización estructural de unidades subléxicas en Castellano mediante una técnica de Inferencia Gramatical basada en el Análisis Sintáctico corrector de Errores". *Tesis doctoral*. Universidad de Valencia.

Schmid, L.A. (1994). "Parsing word graphs using a linguistic grammar and a statistical language model". *Procc of ICASSP, 94*. Vol II. pp 41-44.

Schwartz, R. Austin. S. (1991) "A comparison of several approximation algorithms for finding multiple N-best sentence hypotesis". *Procc. IEEE ICASSP-91*, Toronto, Canada pp.701-704

Schwartz, R. Chow, Y. L. (1990). "Efficient and exact procedure for finding the N most likely sentence hypotheses". *Procc of ICASSP´90.* Albuquerque pp 81-84.

Segarra, E. (1993) "Una aproximación inductiva a la Comprensión del Discurso Continuo", *Tesís Doctoral DSIC*. Universidad Politécnica de Valencia

Shannon D. E. Weaver, W. (1964). "The mathematical theory of communication". *The University of Illinois. Press. Urbana. II.*

Steinbiss, V. Tran, B. Ney, H. (1994). "Improvement in Beam Search". *Procc of the ICSLP*. pp 2143-2146

Steinbiss. V. (1991) " A Search Organization for Large-Vocabulary Recognition Based on N-Best Decoding". *Procc. EUROSPEECH´91*. pp 1217-1220

Suhm. B. Waibel, A. (1994). "Towards Better Language Models for Spontaneous Speech". *Procc of ICSLP. pp. 831-834.*

Suzuki, M. and Aso, H. (1999). "An automatic acquisition method of statistic finite-state automaton for sentences". *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*

Torres, I.(1992) "Selección de unidades sub-léxicas para la decodificación acústico-fonética del habla en castellano". *Informe de investigación DSIC*. Universidad Politécnica de Valencia.

Torres, I. Casacuberta. F. (1993) "Spanish Phone Recognition using semicontinuous Hidden Markov Models". *Procc Int. Conf. Acoust. Speech and Signal processing*. Vol II. pp 515-518.

Torres, I., Casacuberta, F., Varona, A. (1993). "Acoustic-Phonetic Decoding of Spanish Continous Speech with Hidden Markov Models ". *En NATO-ASI. " News Advanced and Trenck in Speech Recognition and Coding*. Granada (Spain). Lectures pp 43-46.

Torres, I., Varona, A., Casacuberta, F., (1994). "Automatic Segmentation and Phone Model Initialization in Continuous Speech Recognition". *Proceedings in Artificial Inteligence*. Ed. H. Niemann, De Mori, R. Hanrieder, G.pp. 286-289.

Torres, I., Casacuberta, F., Varona, A. (1995) "Acoustic-Phonetic Decoding of Spanish Continous Speech with Hidden Markov Models". *Speech Recognition and Coding. New Advances and Trends*, pp 84-87. Ed. A. Rubio & J.M. López. *NATO-ASI, Series, Springer-Verlag*.

Torres, I., Varona, A. (1999). "An efficient representation of K-TSS language models". *Memorias del Simposium Iberoamericano de Reconocimiento de patrones*. pp. 645-654.

Varona, A. (1993) "Segmentación y selección de unidades discriminantes discriminantes transitorias y estacionarias en Decodificación Acústico-Fonética de discurso continuo". *Tesina de Licenciatura*. Noviembre

Varona, A. (1995). "Construcción de sistemas de demostración de Reconocimiento Automático del Habla". *Memoria del Master de Electrónica y Automática del departamento de Electricidad y Electrónica*. UPV/EHU.

Varona, A. Torres, I. Casacuberta, F (1995) "Discriminative-transitional/steady units for Spanish Continuous Speech Recognition". *Proceedings del EUROSPEECH-95. 4th European Conference on Speech Comunication and Technology*. pp 1471-1474

Varona, A. Torres, I. (1999a) "Using Smoothed K-TSS Language Models in Continuous Speech Recognition". *Procc. IEEE Int. Conf. Acoust, Speech, Signal Processing*. Vol II pp. 729-732

Varona, A. Torres, I. (1999b) "Evaluating pruned K-TSS language models: perplexity and word recognition rates". *Proceedings of the VIII Symposium Nacional de Reconocimiento de Formas y Análisis de Imágenes*. Vol I pp 61-68. Bilbao. Seleccionado para ser publicado en el libro Pattern Recognition and Applicactions by *IOS Press*

Vidal, E. Casacuberta, F. García, P. (1995). "Grammatical Inference and Automatic Speech Recognition". *Speech Recognition and Coding. New Advances and Trends*, pp 174-191. Ed. A. Rubio & J.M. López. NATO-ASI, Series, Springer-Verlag.

Wright, J. H., Jones, G. J. F., Lloyd-Thomas, H. (1993). "A consolidated language model for speech recognition". *Procc of EUROSPEECH*, vol 2. Berlin. Germany. pp 977-980.

Wright, J.H., Jones, G.F., Wrigley, E.N. (1992) "Hybrid grammar-bigram speech recognition systems with first-order dependence model". *Procc of ICASSP*, pp 169-172.

Wright, J.M. Jones, G. J. F. Lloyd-Thomas, H. (1994) "A robust language model incorporating a substring parser and extended N-grams". *Procc of ICASSP´94*. Vol Y. pp 361-364.

Young, s. Odell, J. Ollason, D. Valtchev, V. Woodland, P (1997). "The HTK book" (*for HTK version 2.1)*.

Zalcstein, Y. (1972). "Locally Testable Languages". *JCSS6*, pp 151-167

Zhao, R. Kenny, P. Labute, P. O´Shaughnessy, D. (1993). "Issues in large scale statistical language modeling". *Procc of EUROSPEECH´93*. pp 965-968.

Zhou, Q. and Chou, W. (1997). "An Approach to Continuous Speech Recognition Based on Layered Self-Adjusting Decoding Graph", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol 3, pp. 1779-1782.

Apéndice A

Bases de datos

En este apéndice se presentan las bases de datos utilizadas a lo largo de la memoria. En la figura A.1. se puede ver un resumen de las mismas. En primer lugar se presenta la base de datos de entrenamiento acústico conocida como SENGLAR. A partir de esta base de datos acústica se obtienen los modelos acústicos (en este caso modelos ocultos de Markov) que se van a evaluar directamente en la etapa de Decodificación Acústico Fonética (DAF) (ver cap. 4).

Pero el objetivo del trabajo supone la evaluación de un sistema completo de RAH de discurso continuo. Para ello los modelos acústicos se deben integrar junto con los modelos de lenguaje. Estos modelos de lenguaje imponen restricciones sintáctico y/o semánticas de la combinación de palabras y se obtienen a partir de grandes bases de datos de texto para una tarea en concreto. En este trabajo se van a evaluar dos tareas distintas, una muy sencilla MLA de sólo 29 palabras y otra de vocabulario intermedio BDGEO (1208 palabras).

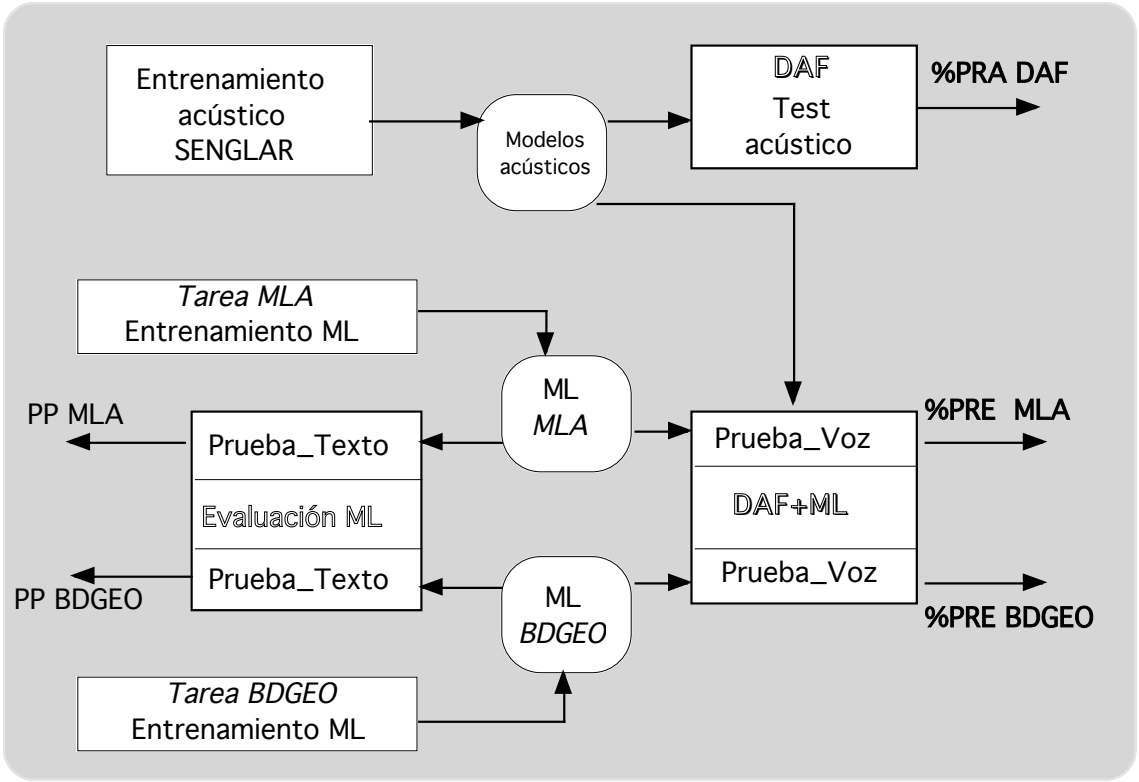


Figura A.1.- Resumen de las bases de datos de texto y voz utilizadas a lo largo de la memoria.

Para la construcción de los ML se cuenta con dos bases de datos de entrenamiento de texto, una para cada una de las tareas bajo estudio. La evaluación de los ML se puede realizar de forma ajena al sistema de reconocimiento (DAF+ML) utilizando como medida la perplejidad (PP) (ver capítulo 2), para ello es necesario contar con un conjunto de frases de texto que no se hayan utilizado en entrenamiento a las que llamaremos Prueba_Texto y por supuesto también habrá un conjunto distinto para cada una de las tareas que pretendemos evaluar. Pero la mejor evaluación de un ML se obtiene de la evaluación de las tasas de error (%PRE) obtenidas cuando se integra junto con los modelos acústicos en el sistema de RAH (ver capítulo 2 y 5). Para esta evaluación es necesario contar con un conjunto de frases de voz previamente grabadas y distintas de las frases de entrenamiento a las que llamaremos (Prueba_Voz) para cada una de las tareas que se pretende evaluar.

A continuación, se van a ir presentando cada una de las bases de datos mencionadas anteriormente. También se lista los conjuntos de Prueba_Voz para cada una de las tareas a evaluar. Finalmente se muestra un estudio de la distribución de fonemas de cada una de las bases de datos de voz que se presentan en este apéndice.

A.1.- Corpus asociado al modelado acústico.

Como se ha dicho anteriormente en este punto se van a presentar los corpus de voz utilizados para el entrenamiento y posterior evaluación de los modelos acústicos. La elección de un conjunto de entrenamiento y evaluación de los modelos acústicos se llevo a dentro del proyecto SENGLAR (TIC95-0884-C04-03) y se propuso el uso de una base de datos elaborada a partir de tres bases de datos acústicas previamente existentes:

EUROM1: es una base de datos fonética en Castellano grabada por la Universidad Politécnica de Cataluña durante el proyecto ESPRIT SAM-A (Moreno, 1993) en verano de 1993. La base de datos completa fue grabada por un total de 59 locutores y se grabaron aproximadamente 1900 frases.

PROBA: es una base de datos fonética en Castellano grabada por la Universidad Politécnica de Cataluña que corresponde al corpus fonético de Albaycin (Llisterri, 1991). Son 20 locutores que grabaron 25 frases cada uno.

FRASES: (Castro, 99). Frases adquiridas por el Grupo de Reconocimiento del Habla de la Universidad Politécnica de Valencia. Las frases han sido adquiridas por 10 locutores (5 masculinos y 5 femeninos). Cada uno de ellos grabó 120 frases fonéticamente balanceadas y 50 frases extraídas de las obras de la literatura contemporánea.

A.1.1.- Corpus de entrenamiento de los modelos acústicos

Para el entrenamiento de los modelos acústicos es necesario el uso de un conjunto de frases fonéticamente balanceadas según la distribución de aparición de los fonemas en la lengua

castellana. El conjunto final de frases que constituyen el corpus de entrenamiento de los modelos acústicos está formado por un subconjunto de cada uno de los corpus que se han presentado previamente, de la siguiente forma:

EUROM1: (43 locutores, 186 enunciados y 842 elocuciones)

PROBA: (7 locutores, 174 enunciados y 174 elocuciones)

FRASES: (7 locutores, 120 enunciados y 513 elocuciones)

locutores l2, l5, l6 y l9 (frases de f1 a f60)

(+ f97, f101, f105, f106, f113)*

locutores l1, l3, l8 y l9 (frases de f61 a f120)

(+ f6, f23, f33, f40, f48, f56)*

locutor 19: frases de f1 a f120

TOTAL: 57 locutores, 480 enunciados, 1529 elocuciones[6], y 60451 fonemas.

A continuación se van a presentar algunos ejemplos del tipo de frases que componen la base de datos:

EUROM1:

- Mama me lo estoy pasando estupendamente en Formentera.

- Los días de fiesta se sentaba y entretenía a los niños del pueblo explicándoles sus aventuras y los peligros que había corrido en altamar.

- Me tocó ser un joven agresivo y director de empresa y no fui capaz de meterme en el personaje.

PROBA:

- Tener un hijo se convirtió para él en una obsesión.

- La decoración de la casa no avanzaba tan rápido como quería.

- Enseguida dejaba caer el libro y se quedaba dormido.

FRASES:

- Tengo una ganga de coche.

- Era de color azul en mis sueños.

- Hay que urdir un plan convincente.

De las frases destinadas al entrenamiento de los modelos acústicos, se seleccionaron aproximadamente un 10% de para ser segmentadas manualmente. De trabajos previos estaban ya segmentadas varias frases del corpus Frases (77 frases) y del corpus de EUROM1 (44 frases). Además se seleccionaron un subconjunto adicional de frases de Proba (41 enunciados)

[6] Las frases marcadas con * se añadieron al corpus de entrenamiento por ser frases que han sido previamente segmentadas a mano.

para ser segmentados de forma que las frecuencias de aparición de las unidades fueran similares a las del entrenamiento. En resumen, hay 162 enunciados segmentados a mano en total. Las 77 frases del corpus Frases, se segmentaron en la Universidad Politécnica de Valencia. El resto de frases fueron segmentadas por Inés Torres de la Universidad del País Vasco/ Euskal Herriko Unibertsitatea.

A.1.2.- Corpus de evaluación de los modelos acústicos

A continuación se presentan el conjunto de frases que se han utilizado para la evaluación de los modelos acústicos y que se han denominado (test_acústico). La evaluación se lleva a cabo sobre el conjunto de frases que se muestra a continuación y que es independiente de vocabulario e independiente de locutor:

EUROM1: (17 locutores, 61 enunciados y 225 elocuciones)

PROBA: (13 locutores, 325 enunciados y elocuciones)

FRASES: (3 locutores, 50 enunciados y 150 elocuciones)

TOTAL: 33 locutores, 436 enunciados y 700 elocuciones.

A.2.- Corpus asociados al modelado de lenguaje

A continuación se presentan dos bases de datos correspondientes a las dos tareas de diferente dificultad que se han evaluado a lo largo de la memoria: una muy sencilla MLA de sólo 29 palabras y otra de vocabulario intermedio BDGEO (1208 palabras).

Como ya se ha explicado, los ML tienen que recoger en su estructura las restricciones sintácticas y/o semánticas del lenguaje de una determinada tarea. Estos modelos de lenguaje se obtienen a partir de grandes bases de datos de texto que sirven para el entrenamiento de los mismos.

Una vez que se obtienen los ML, la evaluación de los mismos puede llevarse a cabo de forma independiente al sistema de RAH utilizando como medida la perplejidad (ver capitulo 2) para lo que es necesario un nuevo conjunto de frases de la tarea. Este conjunto está formado por frases de texto y recibe el nombre de Prueba_Texto. Pero la mejor evaluación resulta de la integración directa de los ML en el sistema de RAH y de la obtención de las tasas de error sobre un conjunto de frases que han debido ser previamente grabadas y que reciben el nombre de Prueba_Voz.

A continuación de presentan estos tres conjuntos, de entrenamiento y evaluación en voz y texto para las dos tareas bajo estudio. Se presenta también una relación del conjunto de frases que pertenecen al conjunto Prueba_Voz de cada una de las dos tareas.

A.2.1.- Tarea MLA

Se trata de una tarea de vocabulario limitado (29 palabras) compuesta por un conjunto de frases presentadas por primera vez en inglés ("*Miniature Language Acquisition Task*", (MLA)) (Feldmann et al, 1990).

Estas frases describen la inclusión o eliminado (se añade, se elimina) de figuras geométricas (triángulo, circulo, cuadrado) en un espacio plano. Este posicionamiento (izquierda, derecha, arriba, abajo) se lleva a cabo teniendo en cuenta las posiciones relativas de otros objetos anteriormente citados y que se identifican mediante su tonalidad (claro, oscuro) y/o tamaño (grande, pequeño, mediano).

Las sentencias utilizada forman parte de la base de datos generada por el grupo de Reconocimiento de Formas e Inteligencia Artificial de la Universidad Politécnica de Valencia, mediante un modelo de lenguaje de tipo 2 (de contexto libre) construido manualmente.

A continuación se presentan el número de frases y palabras contenidas en los conjuntos de entrenamiento (texto) y evaluación (texto y voz) de la tarea MLA (29 palabras). Para el conjunto Prueba_Voz se presenta también el número de fonemas que lo compone, ya que se utiliza en algunos experimentos de decodificación acústico fonética (ver capítulo 4).

	frases	palabras	fonemas
Entrenamiento (texto)	9150	156152	
evaluación (Prueba_Texto)	500	8397	
evaluación (Pueba_Voz)	1600	25664	108576

Las frases correspondientes al conjunto de evaluación Prueba_Voz fueron grabadas en el Departamento del Electricidad y Electrónica de la Facultad de Ciencias (UPV/EHU) por 16 locutores lo que hace un total de: 16 locutores, 100 enunciados y 1600 elocuciones.

En la figura A.2 se presenta un estudio de la distribución de las frases de los distintos subconjuntos de la base de datos de la tarea MLA según su longitud en palabras. La frase de longitud mínima tiene 5 palabras y la de longitud máxima 28. La media se sitúa alrededor de 16 palabras por frase.

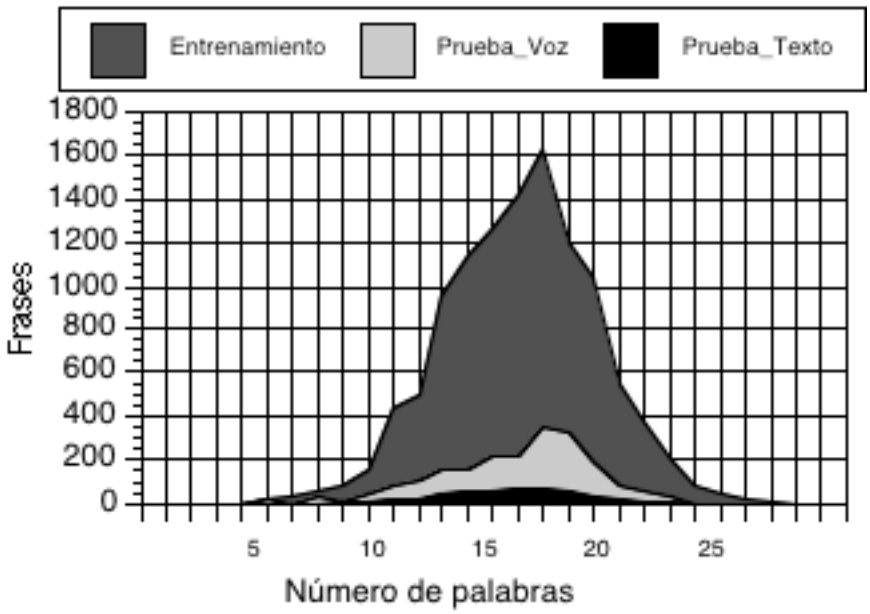


Figura A.2.- Datos correspondientes a la longitud en palabras de la base datos correspondiente a la tarea MLA.

Transcripción del conjunto de frases: Prueba_Voz:

1 un circulo y un circulo claro tocan a un circulo oscuro y un circulo grande

2 se elimina el circulo mediano y claro que esta encima del circulo claro

3 un cuadrado esta a la izquierda de un cuadrado mediano y claro

4 se elimina el circulo peque~no y claro que esta debajo del triangulo y del triangulo grande y claro

5 se elimina el circulo grande que esta debajo del triangulo peque~no y claro y del circulo grande

6 un circulo peque~no y un circulo grande tocan a un circulo mediano

7 se a~nade un circulo oscuro debajo del circulo oscuro y del circulo grande y claro

8 un triangulo claro y un circulo tocan a un circulo mediano y claro

9 se elimina el triangulo peque~no que esta muy por debajo del cuadrado grande

10 se a~nade un circulo grande y oscuro muy a la izquierda del triangulo mediano

11 un circulo y un circulo peque~no y claro tocan a un circulo grande y claro y un cuadrado claro

12 un cuadrado grande esta muy a la izquierda de un circulo claro y un cuadrado peque~no y claro

13 un circulo grande y un circulo claro estan muy a la derecha de un cuadrado grande y oscuro

14 se a~nade un cuadrado mediano y oscuro muy a la izquierda del circulo mediano y oscuro y del circulo peque~no y claro

15 un circulo y un triangulo grande estan a la derecha de un triangulo mediano y un cuadrado oscuro

16 se elimina el triangulo oscuro que esta a la izquierda del triangulo y del circulo grande y claro

17 se elimina el circulo que esta muy por debajo del circulo claro y del cuadrado mediano y claro

18 se a~nade un cuadrado muy a la izquierda del cuadrado mediano y del triangulo peque~no y oscuro

19 se a~nade un cuadrado peque~no y oscuro debajo del circulo grande

20 se elimina el cuadrado oscuro que esta a la izquierda del circulo y del triangulo oscuro

21 un cuadrado y un cuadrado grande tocan a un circulo claro

22 un circulo peque~no esta muy por encima de un circulo y un circulo grande

23 un cuadrado mediano y un triangulo tocan a un circulo y un circulo mediano

24 un circulo peque~no y un circulo claro estan muy por encima de un cuadrado mediano

25 un circulo claro y un circulo claro estan muy por debajo de un circulo claro

26 se a~nade un triangulo mediano y claro muy por debajo del cuadrado mediano y del circulo claro

27 un circulo oscuro y un cuadrado oscuro tocan a un triangulo oscuro y un triangulo

28 un circulo peque~no y oscuro y un circulo mediano y oscuro tocan a un circulo y un cuadrado

29 se a~nade un cuadrado claro debajo del triangulo grande y claro

30 un circulo grande y claro y un cuadrado tocan a un circulo oscuro y un cuadrado

31 un circulo y un cuadrado estan a la derecha de un triangulo mediano y claro y un triangulo

32 se elimina el circulo claro que esta encima del triangulo grande y claro y del circulo grande y claro

33 se elimina el cuadrado que esta muy a la derecha del cuadrado mediano y oscuro y del cuadrado grande y oscuro

34 se elimina el circulo mediano y oscuro que esta encima del triangulo oscuro y del circulo mediano

35 un triangulo claro esta debajo de un triangulo

36 se elimina el cuadrado grande y oscuro que esta debajo del circulo oscuro

37 un circulo mediano y claro y un triangulo estan a la derecha de un circulo y un circulo oscuro

38 se elimina el cuadrado mediano y claro que esta muy a la derecha del triangulo mediano y oscuro

39 un circulo oscuro y un cuadrado estan muy a la derecha de un circulo mediano y un circulo

40 se elimina el circulo que esta muy por encima del triangulo y del triangulo oscuro

41 se elimina el cuadrado grande y oscuro que esta muy a la izquierda del cuadrado claro y del circulo claro

42 se elimina el circulo claro que esta muy por encima del triangulo grande y claro y del circulo mediano y oscuro

43 un circulo oscuro toca a un circulo grande y claro y un cuadrado grande

44 se a~nade un triangulo muy por debajo del triangulo y del circulo oscuro

45 un circulo grande y claro esta a la izquierda de un circulo claro y un circulo oscuro

46 se elimina el circulo grande y oscuro que esta muy por encima del circulo mediano y claro y del circulo mediano

47 se a~nade un triangulo grande y oscuro muy por encima del triangulo y del circulo peque~no y oscuro

48 se elimina el cuadrado peque~no y claro que esta muy a la izquierda del cuadrado peque~no y del circulo grande y claro

49 un cuadrado claro toca a un circulo grande y un cuadrado grande y claro

50 se elimina el cuadrado peque~no y claro que esta muy a la izquierda del triangulo mediano

51 un circulo y un cuadrado estan a la derecha de un triangulo grande y oscuro y un triangulo

52 un circulo y un circulo mediano y claro tocan a un triangulo peque~no y un circulo oscuro

53 un cuadrado oscuro y un cuadrado mediano y oscuro estan a la izquierda de un circulo grande y oscuro

54 un circulo grande y oscuro esta encima de un circulo peque~no y un cuadrado oscuro

55 se elimina el cuadrado mediano que esta a la derecha del triangulo oscuro y del circulo oscuro

56 se elimina el cuadrado que esta encima del circulo peque~no y claro y del cuadrado oscuro

57 un circulo esta debajo de un triangulo mediano y oscuro

58 se elimina el circulo grande que esta muy por encima del circulo mediano y claro y del circulo peque~no

59 se a~nade un circulo a la derecha del cuadrado mediano

60 un cuadrado grande y un circulo peque~no estan muy por encima de un cuadrado mediano

61 un cuadrado grande y oscuro y un cuadrado mediano y claro estan encima de un circulo y un triangulo grande

62 un circulo mediano y claro y un circulo claro estan muy a la izquierda de un circulo grande

63 un triangulo oscuro esta a la izquierda de un circulo grande y oscuro

64 se elimina el triangulo claro que esta encima del circulo peque~no y oscuro y del circulo

65 un triangulo peque~no y claro y un circulo peque~no y oscuro estan muy a la derecha de un triangulo peque~no

66 se a~nade un cuadrado grande muy a la derecha del circulo grande y del cuadrado mediano y oscuro

67 se elimina el cuadrado mediano que esta debajo del cuadrado mediano y claro y del cuadrado claro

68 un triangulo peque~no y un triangulo peque~no y oscuro estan a la derecha de un circulo oscuro

69 se a~nade un triangulo a la derecha del circulo peque~no y claro y del triangulo grande

70 se a~nade un triangulo mediano y oscuro muy por debajo del cuadrado mediano y oscuro

71 se a~nade un triangulo a la derecha del triangulo peque~no y oscuro y del circulo peque~no y oscuro

72 se elimina el cuadrado peque~no que esta debajo del triangulo oscuro y del triangulo mediano y claro

73 se elimina el triangulo peque~no y oscuro que esta debajo del circulo y del cuadrado peque~no y oscuro

74 se elimina el triangulo que esta muy a la derecha del circulo grande y del circulo mediano

75 se a~nade un triangulo peque~no debajo del cuadrado

76 un circulo esta muy a la izquierda de un triangulo mediano y claro y un circulo grande y oscuro

77 se elimina el circulo peque~no que esta muy por encima del triangulo grande y del cuadrado mediano y oscuro

78 un triangulo mediano y claro y un triangulo estan muy por encima de un cuadrado

79 se elimina el circulo oscuro que esta muy por debajo del triangulo claro y del circulo grande

80 un cuadrado claro y un triangulo mediano y oscuro estan debajo de un cuadrado grande y claro

81 un triangulo oscuro esta encima de un circulo claro y un circulo

82 un triangulo oscuro esta encima de un cuadrado y un circulo

83 un circulo peque~no y claro esta muy a la derecha de un circulo y un triangulo mediano

84 se elimina el circulo mediano y claro que esta debajo del triangulo peque~no y del circulo claro

85 se elimina el triangulo grande que esta muy por debajo del cuadrado claro y del circulo grande y claro

86 se elimina el triangulo grande y oscuro que esta muy a la derecha del triangulo claro

87 un triangulo peque~no y oscuro y un cuadrado oscuro estan muy por debajo de un circulo y un triangulo claro

88 un circulo y un circulo grande estan debajo de un cuadrado peque~no

89 un cuadrado claro esta debajo de un circulo y un circulo mediano

90 se elimina el triangulo mediano y oscuro que esta debajo del triangulo peque~no y del circulo peque~no y claro

91 un circulo y un cuadrado peque~no

92 un circulo esta a la derecha de un circulo peque~no y un circulo claro

93 se a~nade un cuadrado peque~no y oscuro muy a la izquierda del cuadrado claro y del triangulo oscuro

94 se elimina el circulo peque~no que esta muy a la derecha del cuadrado grande y claro

95 un circulo y un circulo tocan a un circulo y un cuadrado mediano

96 un circulo peque~no y un cuadrado tocan a un circulo mediano y oscuro y un circulo claro

97 un triangulo grande y oscuro y un circulo claro tocan a un cuadrado y un circulo claro

98 se elimina el circulo mediano que esta a la izquierda del triangulo oscuro y del cuadrado mediano

99 se elimina el triangulo grande que esta muy por encima del circulo claro y del circulo

100 se elimina el triangulo grande que esta muy por debajo del cuadrado mediano y del cuadrado peque~no y oscuro

A.2.2.- Tarea BDGEO

Se trata de una tarea que describe el acceso a una base de datos geográfica en Castellano. El corpus BDGEO está formado por un conjunto de preguntas a una base de datos geográfica de España. Este corpus forma parte de la base de datos par investigación en Reconocimiento Automático del Habla "Albaycin", desarrollada dentro de un proyecto conjunto por diversos grupos de investigación españoles (Díaz et al, 1993) (Moreno, 1993). La tarea BDGEO está especialmente diseñada par realizar ensayos de sistemas completos de reconocimiento, presentando una perplejidad razonablemente baja.

A continuación se presentan el número de frases y palabras contenidas en los conjuntos de entrenamiento (texto) y evaluación (texto y voz) de la tarea MLA (29 palabras). Para el conjunto Prueba_Voz se presenta también el número de fonemas que lo compone, ya que se utiliza en algunos experimentos de decodificación acústico fonética (ver capítulo 4).

	frases	palabras	fonemas
Entrenamiento (texto)	8262	81620	
evaluación (Prueba_Texto)	1147	11823	
evaluación (Pueba_Voz)	600	5655	25098

Las frases correspondientes al conjunto de evaluación Prueba_Voz fueron grabadas en el Departamento de teoría de la señal y comunicaciones de la Universidad Politécnica de Cataluña (12 locutores, 200 enunciados y 600 elocuciones).

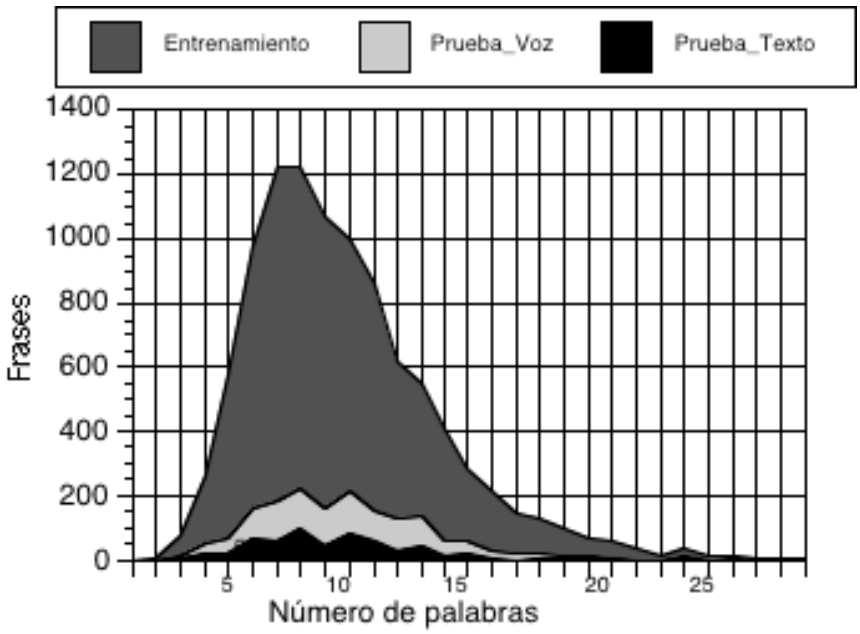


Figura A.3.- Datos correspondientes a la longitud en palabras de la base datos correspondiente a la tarea BDGEO.

En la figura A.3 se presenta un estudio de la distribución de las frases de los distintos subconjuntos de la base de datos de la tarea BDGEO según su longitud en palabras. La frase de longitud mínima tiene 3 palabras y la de longitud máxima 32. La media se sitúa alrededor de 9 palabras por frase.

A continuación se presenta el conjunto de frases que forman parte del conjunto Prueba_Voz.

Transcripción del conjunto de frases: Prueba_Voz:

1 a que mar va a parar el rio espa~nol de mayor longitud

2 caudal de los rios con mas de cien kilometros de longitud

3 caudal de los rios que pasan por castilla y leon

4 caudal del rio que pasa por la comunidad de valencia

5 como se llama el mar que ba~na valencia

6 comunidad autonoma mas grande

7 comunidades autonomas mas grandes que catalu~na

8 comunidades con mas de cinco rios

9 comunidades por las que pasa el rio ebro

10 comunidades que ba~na el mar mediterraneo

11 cual es el caudal de todos los rios de la comunidad valenciana

12 cual es el caudal del ebro

13 cual es el caudal del rio mas largo que pasa por andalucia

14 cual es el caudal maximo de los rios espa~noles

15 cual es el caudal y longitud del tajo

16 cual es el mar en el que desembocan mayor numero de rios con una longitud mayor de doscientos kilometros

17 cual es el mar que rodea las canarias

18 cual es el nombre del rio mas largo de la peninsula

19 cual es el rio de mayor longitud que desemboca en el mar cantabrico

20 cual es el rio mas caudaloso que pasa por extremadura

21 cual es el rio mas largo que atraviesa por lo menos dos comunidades

22 cual es la comunidad autonoma de mayor extension por la que pasa el rio ebro

23 cual es la extension de la comunidad autonoma en la que nace el rio ebro

24 cual es la longitud de todos los rios

25 cuales son las comunidades autonomas con una extension superior a veinte mil kilometros cuadrados

26 cuales son las comunidades autonomas por las que pasan mas rios

27 cuales son las comunidades que atraviesa el tajo

28 cuales son las comunidades que lindan con el mar

29 cuales son los rios catalanes mas largos que cien kilometros

30 cuales son los rios cuya longitud es superior a cien kilometros

31 cuales son los rios que desembocan en el cantabrico

32 cuales son los rios que pasan por extremadura y otras dos comunidades autonomas

33 cuales son los rios que pasan por la comunidad de valencia

34 cuantas comunidades estan ba~nadas por dos mares

35 cuanto mide el tajo

36 cuantos mares reciben agua de un rio

37 cuantos metros cubicos por segundo lleva el turia

38 cuantos rios con caudal mayor de ochocientos metros cubicos por segundo pasan por la comunidad valenciana

39 cuantos rios de castilla y leon tienen mas de cien kilometros

40 cuantos rios pasan por aragon y catalu~na

41 cuantos rios son mas largos de doscientos kilometros

42 de los rios del estado cuantos desembocan en el mediterraneo

43 deseo saber el caudal del rio mi~no

44 di el caudal del rio menos caudaloso

45 di el rio mas caudaloso que desemboca en el cantabrico

46 digame el nombre del rio mas largo

47 dime comunidades cuya superficie sea mayor a mil kilometros cuadrados

48 dime cual es la comunidad autonoma de menor extension

49 dime cuales son las comunidades autonomas

50 dime cuantos rios de la comunidad valenciana tienen mas de doscientos kilometros de longitud

51 dime donde desemboca el rio jucar

52 dime donde muere el rio ebro

53 dime donde nace el rio jucar

54 dime el caudal de los rios de catalu~na

55 dime el caudal de todos los rios que desembocan en el mar mediterraneo

56 dime el caudal del rio cuervo

57 dime el caudal del rio mas peque~no que pasa por la rioja

58 dime el caudal maximo de los rios

59 dime el mar donde desemboca el rio turia

60 dime el mar en que desemboca el mi~no

61 dime el nombre de las comunidades que linden con dos mares

62 dime el nombre de las tres comunidades autonomas mas grandes

63 dime el nombre de los mares que ba~nan la comunidad de andalucia

64 dime el nombre de los rios que desembocan en el oceano atlantico

65 dime el nombre de los rios que pasan por la comunidad de madrid

66 dime el nombre de los rios que tienen menos de cien kilometros

67 dime el nombre de todas las comunidades que tienen mar

68 dime el numero de rios que desembocan en el mediterraneo y que sean entre mil y doscientos kilometros de largo

69 dime el rio de mayor caudal que pase por la comunidad de valencia

70 dime el rio de menor longitud de catalu~na

71 dime en que comunidad autonoma nace el tajo

72 dime en que comunidad nace el rio turia

73 dime la comunidad en la que desemboca el rio turia

74 dime la extension de la comunidad asturiana

75 dime la extension de las comunidades por donde pasa el ebro

76 dime la longitud de los rios que pasan por la comunidad de madrid

77 dime la longitud del rio guadalquivir

78 dime la longitud del rio mas largo

79 dime las comunidades autonomas

80 dime las comunidades autonomas con extension superior a mil kilometros cuadrados

81 dime las comunidades que lindan con mas de un mar

82 dime lo grande que es el ebro

83 dime los mares

84 dime los mares que ba~nan andalucia

85 dime los rios con una longitud superior a quinientos kilometros

86 dime los rios de la comunidad autonoma gallega

87 dime los rios que desembocan en andalucia

88 dime los rios que desembocan en el atlantico

89 dime los rios que nacen en la comunidad foral de navarra

90 dime los rios que nacen y desembocan en la misma comunidad

91 dime los rios que pasan por la comunidad de madrid

92 dime los rios que tengan una longitud mayor que quinientos kilometros

93 dime que longitud tiene el rio jucar

94 dime que rio tiene el caudal mas grande

95 dime si por la comunidad de valencia pasa o no mas de un rio

96 dime todos los mares que ba~nan andalucia

97 dime todos los rios que desembocan en el mar cantabrico

98 donde desemboca el guadiana

99 donde nace el rio duero

100 donde nace el rio ebro

101 el rio ebro pasa por la comunidad autonoma de navarra

102 el rio mi~no por cuantas comunidades autonomas pasa

103 en que comunidad autonoma esta el rio mas caudaloso

104 en que comunidad autonoma hay mas rios

105 en que comunidad autonoma pasan nacen y desembocan mas rios

106 en que comunidad desemboca el rio ebro

107 en que comunidad nace y pasa el pisuerga

108 en que comunidad nacen mas rios

109 en que mar desemboca el rio mas caudaloso de la comunidad andaluza

110 en que mar desembocan mayor numero de rios

111 entre el rio ebro y el jucar cual de ellos es mas corto

112 enumera las comunidades autonomas por donde pasa el ebro

113 enumera los rios que tienen una longitud mayor de cien kilometros

114 enumerar los rios que atraviesan la comunidad autonoma de asturias

115 es el ebro mas caudaloso que el tajo

116 extension de la comunidad autonoma por la cual pasa el rio cuyo nombre es guadalquivir

117 extension del pais vasco

118 hay algun rio cuyo caudal sea mayor que cien metros cubicos por segundo

119 la extension de las comunidades autonomas que dan al mar atlantico

120 lista de las comunidades por las que pase algun rio de longitud mayor de mil kilometros

121 listado de todos los rios con una longitud menor que la del jucar

122 longitud de los rios que desembocan en el mar cantabrico

123 longitud del rio ebro

124 longitud del rio que pasa por la comunidad valenciana

125 lugar donde desemboca el jucar

126 mar en el que desembocan mas rios

127 mares en los que desembocan cinco o mas rios de longitud superior a cien kilometros

128 mares que ba~nan la comunidad gallega

129 me podria decir cual es la comunidad donde esta el nacimiento del guadiana

130 nombra los rios que pasan por las comunidades autonomas que no dan al mar

131 nombrame los rios que pasan exactamente por tres comunidades autonomas

132 nombre de la comunidad autonoma en la que desemboquen mayor numero de rios

133 nombre de las comunidades con extension mayor que la comunidad valenciana

134 nombre de las tres comunidades de menor extension

135 nombre de los mares que estan en la comunidad valenciana

136 nombre de los rios cuya longitud no supere los mil kilometros y no sea menor de cien kilometros

137 nombre de los rios cuyo caudal es superior a ochocientos metros cubicos por segundo

138 nombre de los rios que desembocan en cada mar

139 nombre de los rios que nacen en la rioja y pasan por aquellas comunidades por las que solo pasa ese rio

140 nombre de los rios que pasen por castilla y leon desembocan en el atlantico y su caudal sea menor que el del rio tajo

141 nombre de todos los mares que ba~nan andalucia

142 nombre del mar en el que desemboca un rio que nace en aragon

143 nombres de comunidades autonomas cuya extension se encuentra entre mil y dos mil kilometros cuadrados

144 numero de mares del estado espa~nol

145 numero de rios que nacen y desembocan en la comunidad valenciana

146 obtener las comunidades autonomas por donde pasa el ebro

147 pasa algun rio por mas de cuatro comunidades

148 pasa el rio duero por la comunidad de madrid

149 por cuantas comunidades pasa el ebro

150 por donde pasa el rio con mas caudal

151 por donde pasa el rio duero

152 por que comunidad pasan mas rios

153 por que mar esta ba~nada asturias

154 que caudal tiene el ebro

155 que caudal tiene el mi~no

156 que comunidad autonoma es menos extensa

157 que comunidad ba~nada por el mediterraneo es la mas extensa

158 que comunidades no son ba~nadas por algun mar

159 que comunidades son ba~nadas por el tajo

160 que comunidades tienen una extension mayor de mil kilometros cuadrados

161 que extension tiene el pais vasco

162 que longitud tiene el rio mas largo

163 que mar ba~na asturias

164 que mar ba~na las costas de la comunidad de madrid

165 que mar ba~na las costas del pais vasco

166 que mar esta junto a la comunidad valenciana

167 que rio cruza menos comunidades

168 que rio desemboca en el mar mediterraneo y pasa por murcia

169 que rio es mas largo el tajo o el ebro

170 que rio tiene mas caudal el tajo o el ebro

171 que rios desembocan en el mar menor

172 que rios extreme~nos tienen una longitud superior a los doscientos kilometros

173 que rios hay en asturias

174 que rios nacen en cantabria

175 que rios pasan por asturias y no nacen alli

176 que rios poseen un caudal superior a ochocientos metros cubicos por segundo

177 que rios tienen mas caudal que el rio duero

178 que rios tienen una longitud comprendida entre quinientos y mil kilometros

179 quiero saber los nombres de los rios mas largos de doscientos kilometros

180 quisiera conocer cuantos rios tienen un caudal de mas de doscientos metros cubicos por segundo y son de menos de mil kilometros de largo

181 quisiera saber en que mar desemboca el segura

182 quisiera saber que comunidades autonomas no tienen salida al mar

183 rio mas corto que desemboca en el cantabrico

184 rio mas largo que nazca en extremadura

185 rios con caudal superior al del rio guadalquivir

186 rios cuya longitud sea mayor de mil kilometros

187 rios de cantabria de mas de cien kilometros de longitud

188 rios de la comunidad autonoma gallega

189 rios que atraviesen mas de tres autonomias

190 rios que desembocan en el cantabrico con una longitud mayor a cien kilometros

191 rios que desemboquen en el cantabrico

192 rios que mueren en el cantabrico

193 rios que nacen en la comunidad de madrid

194 rios que nacen en una comunidad ba~nada por el mar y desembocan en otra comunidad

195 rios que pasan por la comunidad autonoma de valencia

196 rios que tengan un caudal superior a ochocientos metros cubicos por segundo

197 seguro que el segura pasa por la comunidad de valencia

198 tiene alguna comunidad mas extension que la comunidad andaluza

199 tienen la misma longitud y el mismo caudal el rio guadiana y el rio guadalquivir

200 todos los rios

A.3.- Distribución de fonemas de los corpus de voz

En este punto se presenta (ver tabla A.1) la distribución de fonemas de todas las bases de datos de voz que han sido presentadas en este apéndice.

Tabla A.1.- Distribución en fonemas de los conjuntos de voz presentados en el apéndice A.

fonemas	entrenamiento	%	test_acústico	%	MLA	%	BDGEO	%
@	405	**0.67**	239	**0.71**	0	**0.00**	456	**1.81**
B	1566	**2.69**	906	**2.71**	432	**0.39**	459	**1.82**
D	2563	**4.24**	1532	**4.58**	7616	**7.00**	1752	**6.98**
G	710	**1.17**	378	**1.13**	2160	**1.98**	171	**0.68**
J	153	**0.25**	97	**0.29**	1008	**0.92**	87	**0.34**
L	435	**0.72**	197	**0.59**	1696	**1.56**	33	**0.13**
S	3695	**6.11**	1400	**4.18**	3200	**2.94**	1200	**4.78**
T	959	**1.59**	576	**1.72**	2880	**2.64**	222	**0.88**
a	7859	**13.00**	4392	**13.13**	11968	**11.00**	3015	**12.01**
c	250	**0.41**	92	**0.27**	320	**0.29**	12	**0.05**
e	7779	**12.87**	4356	**13.02**	9680	**8.89**	2994	**11.93**
f	454	**0.75**	174	**0.52**	0	**0.00**	6	**0.02**
i	3733	**6.17**	2296	**6.87**	8464	**7.77**	1833	**7.30**
k	2152	**3.56**	1174	**3.51**	8000	**7.34**	1431	**5.70**
l	2969	**4.91**	1520	**4.54**	8064	**7.40**	1443	**5.75**
m	1965	**3.25**	1029	**3.08**	2464	**2.26**	1416	**5.64**
n	3834	**6.34**	2166	**6.47**	7072	**6.50**	1737	**6.92**
o	5449	**9.01**	2989	**8.94**	10784	**9.90**	2664	**10.61**
p	1514	**2.50**	902	**2.70**	1088	**1.10**	363	**1.45**
r	3412	**5.64**	1951	**5.83**	9248	**8.49**	1056	**4.21**
s	3822	**6.32**	2304	**6.89**	3184	**2.92**	1923	**7.66**
t	2481	**4.10**	1398	**4.18**	2560	**2.35**	798	**3.18**
u	1977	**3.27**	1102	**3.29**	9456	**8.68**	1083	**4.31**
x	315	**0.52**	264	**0.78**	432	**0.62**	144	**0.57**
total	60451	**100**	33434	**100**	108576	**100**	25098	**100**

El conjunto de entrenamiento presenta una distribución balanceada de los fonemas en función de la frecuencia de ocurrencia de los mismos en la lengua castellana. Esta distribución se mantiene para el conjunto de evaluación acústico (test_acústico), algo que no ocurre para los otros dos conjuntos Prueba_Voz de las tareas MLA y BDGEO.

De hecho para la tarea MLA hay dos fonemas que ni siquiera aparecen (@ y f) y algunos como por ejemplo la “r” y la “u” aparecen en una proporción mucho mayor. La tarea BDGEO resulta más balanceada aunque tampoco puede ser catalogada como una base de datos de evaluación de DAF. Esto tiene que tenerse en cuenta en los experimentos de DAF que se muestran en el capítulo 4 para las tres bases de datos y por tanto habrá que considerar los experimentos sobre el test acústico como los más adecuados.

Apéndice B

Representación de la estructura del modelo integrado

En primer lugar, en el subapéndice B1 se presenta la estructura del modelo que integra K modelos *k*-EE suavizados (*k*=1...K) mediante una red finita. Esta red representa una estructura inicial en forma de árbol construido a partir de un conjunto de muestras de entrenamiento. La naturaleza de la estructura de tipo árbol permite desarrollar tanto la estructura del modelo suavizado como la distribución de la probabilidad en una estructura muy simple (ver capítulo 3). En el subápendice B2 se presenta el procedimiento para crear la estructura del modelo integrado y suavizado como un array.

B.1.- La estructura del modelo K-EE suavizado como una red finita

En primer lugar, los datos que se obtienen a partir de las muestras de entrenamiento se guardan en una estructura de tipo árbol con K-1 niveles. Cada nodo en cada nivel *k*, con *k* = 1...K-1, se asocia a un estado del modelo integrado y suavizado etiquetado por una secuencia de palabras de longitud *k*, w_{i-k}^{i-1}, que han sido observadas en el conjunto de muestras de entrenamiento (eventos vistos).

El número de hijos de cada nodo $q \equiv w_{i-k}^{i-1}$ es igual al número de palabras que aparecen después de la subsecuencia w_{i-k}^{i-1} en las sentencias de entrenamiento, es decir, $|\Sigma_q|$. La raíz del árbol se asocia al estado *nil* representando una secuencia de palabras vacía. El nivel uno representa el modelo 1-EE y está formado por $|\Sigma|$ nodos correspondientes a cada palabra del vocabulario. Como ya se ha dicho con anterioridad el autómata tiene un único estado inicial y final $q_0 \equiv q_f$, etiquetado como \$. Este hecho permite analizar consecutivamente conjuntos de sentencias forzadas a empezar y terminar con el mismo símbolo \$. Este estado es diferente del nodo raíz de la estructura arborescente, etiquetado como *nil*, puesto que la probabilidad $P(w_i/nil)$ es la probabilidad estimada $P(w_i)$, mientras que $P(w_i/\$)$ es la probabilidad asociada a w_i siendo w_i la palabra inicial de la sentencia. Como consecuencia, cada vez que se evalúa una nueva sentencia la probabilidad $P(\$/nil)$ debiera ser considerada, siendo su valor estimado dependiente del tamaño del corpus de entrenamiento en términos del número de sentencias. Además este valor tendría una fuerte incidencia en el valor estimado de $P(w_i)$ $\forall w_i \in \Sigma$. Pero, el enlace se

puede eliminar si cada sentencia se evalúa individualmente. De esa manera, la estructura en árbol original se transforma en un bi-trie donde *nil* y $ son los nodos raíz pero solo $ representa la estado inicial del autómata.

A continuación se ilustra este procedimiento mediante un ejemplo. En la figura B.2 se muestra la estructura bi-trie para K=4 con el conjunto de entrenamiento R^+ presentado en la figura B.1. El corpus de entrenamiento corresponde a un poema del famoso poeta español Miguel Hernández.

llego con tres heridas
la del amor
la de la muerte
la de la vida
con tres heridas viene
la de la vida
la del amor
la de la muerte
con tres heridas yo
la de la vida
la de la muerte
la del amor

Miguel Hernández

$$R^+ = \left\{ \begin{array}{l} \$llego\ con\ tres\ heridas\ \$\ la\ del\ amor \\ \$\ la\ de\ la\ muerte\ \$\ la\ de\ la\ vida \\ \$\ con\ tres\ heridas\ viene\ \$\ la\ de\ la\ vida \\ \$\ la\ del\ amor\ \$\ la\ de\ la\ muerte \\ \$\ con\ tres\ heridas\ yo\ \$\ la\ de\ la\ vida \\ \$\ la\ de\ la\ muerte\ \$\ la\ del\ amor\ \$ \end{array} \right\}$$

$$\Sigma = \left\{ \begin{array}{l} la,\ de,\ del,\ amor,\ muerte,\ vida, \\ llego,\ con,\ tres,\ heridas,\ viene,\ yo \end{array} \right\}$$

Figura B.1.- Corpus de entrenamiento seleccionado para realizar la construcción de los autómatas *k*-EE.

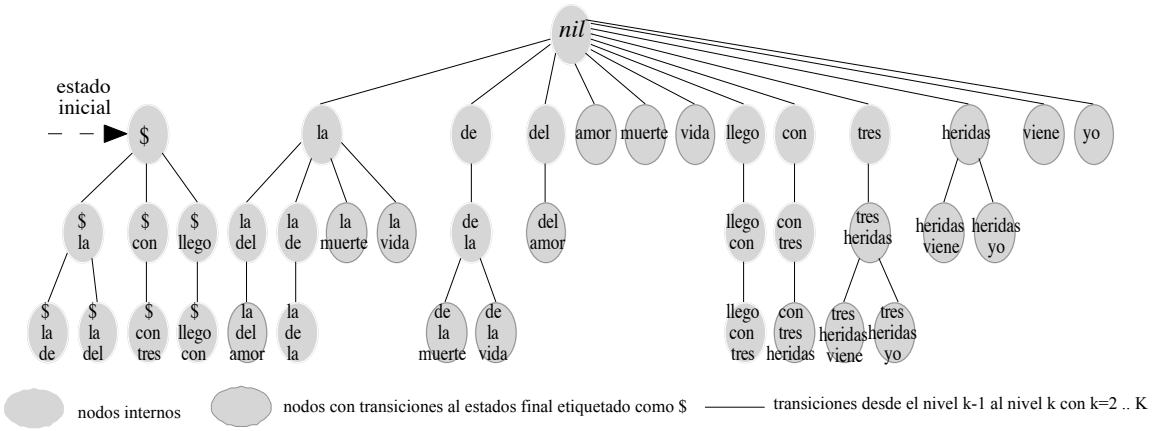


Figura B.2.- Bi-trie obtenido directamente del conjunto de muestras de entrenamiento R^+ de la figura B.1 después de realizar el primer paso de la transformación para un modelo K-EE con K=4.

Se puede ver que todos los estados del autómata están representados por sus nodos correspondientes en la estructura del bi-trie. Las transiciones del tipo (a) (ver ecuación 3.17) están representadas por enlaces entre la raíz *nil* y sus nodos hijos. Los enlaces conectando el estado $ con sus hijos representan transiciones del tipo (b) (ver ecuación 3.18). Las transiciones correspondientes a cadenas de palabras de longitud inferior a K (tipo (c), ver ecuación 3.19)

están representadas por $|\Sigma_q|$ enlaces conectando k=1...K-2 niveles con $|\Sigma_q|$ hijos. Los nodos del nivel K-1 representan los estados asociados a cadenas de palabras de longitud igual a K-1 etiquetadas como $w_{i-(K-1)}^{i-1}$. Las transiciones al estado final (tipo e, ver ecuación 3.21) existen para todos los nodos en el bi-trie cada vez que aparece el símbolo $ en el corpus de entrenamiento, aunque por motivos de claridad, no aparecen dibujadas en la figura B.2. Los nodos que representan estados con transiciones al estado etiquetado con $ se han marcado simplemente de forma especial.

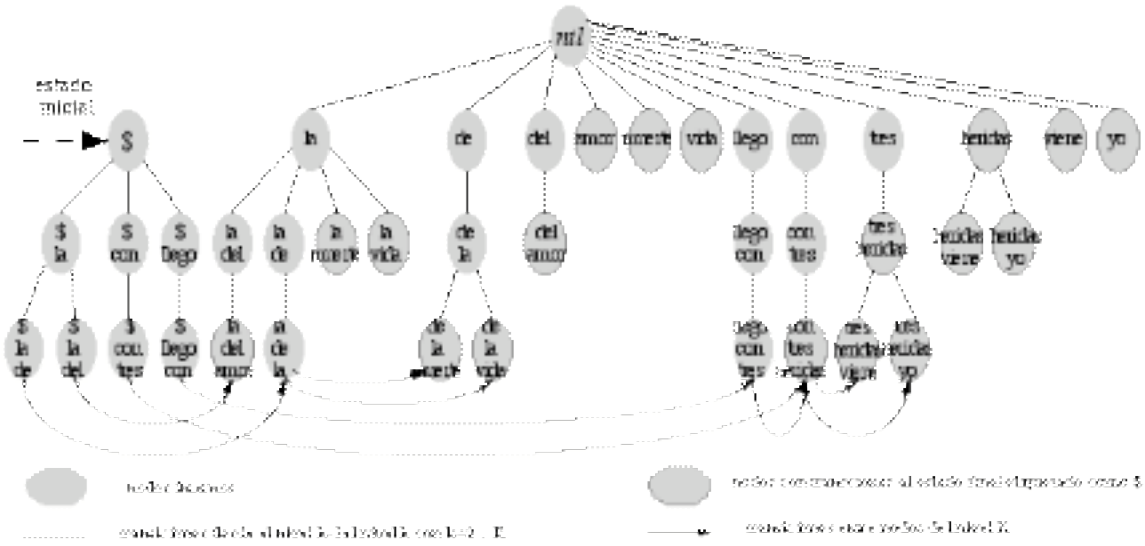


Figura B.3.- La red representando la estructura del bi-trie no suavizado (K=4) que se obtiene a partir de la figura B.2.

Para completar la representación del modelo, es necesario añadir las transiciones que corresponden a cadenas de palabras de longitud máxima K que conectan estados asociados a secuencias de palabras de longitud K-1 (tipo (d), ver ecuación 3.20). Esas transiciones están representadas por enlaces entre nodos de nivel K-1 etiquetados como $w_{i-(K-1)}^{i-1}$. La red finita que se obtiene incluyendo esas transiciones representa la estructura completa del modelo no suavizado (ver figura B.3) para el corpus de entrenamiento de la figura B.1.

Por último, se aplica la técnica de suavizado (ver ecuación 3.13) en cada nodo para obtener la red final que representa al modelo integrado y suavizado para el que se deben usar nuevos enlaces que representan transiciones de back-off (ver ecuaciones 3.23 y 3.24). Como ya se ha dicho con anterioridad, el estado de back-off b_q asociado a cada estado q se puede encontrar en el modelo de orden (K-1) (ver figura B-4 que coincide con la figura 3.6). Por eso, el nodo destino para este enlace debe ser buscado en el nivel inmediatamente superior del árbol.

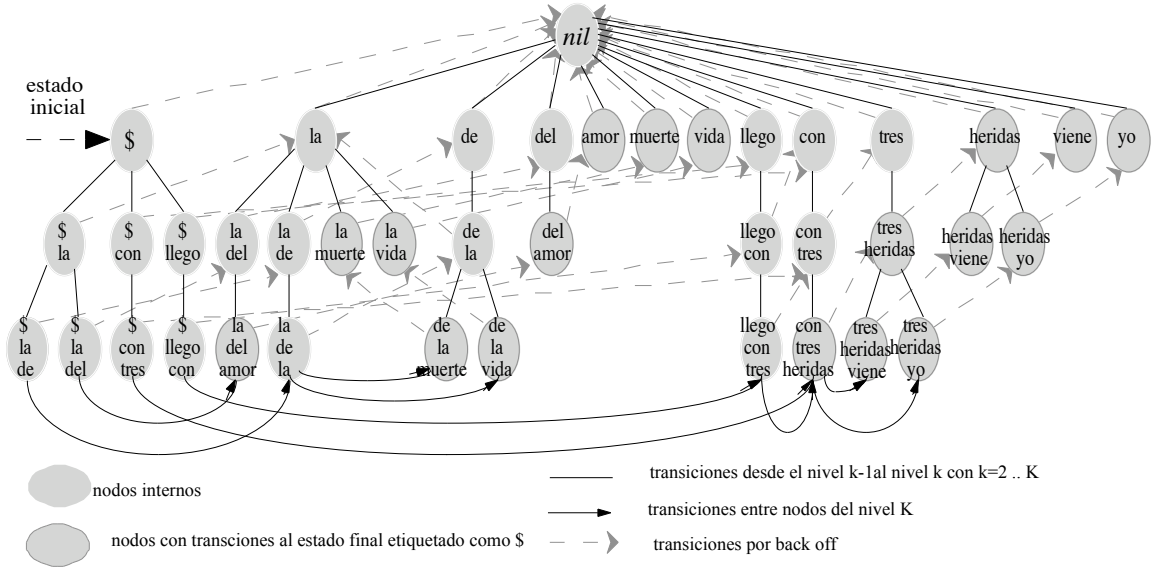


Figura B.4.- La red completa representando la estructura del modelo integrado y suavizado (K=4) obtenido a partir del corpus de entrenamiento de la figura B.1.

B.2.- Construcción de la estructura mediante un array

El procedimiento para crear la estructura tipo array a partir de las muestras de entrenamiento consta de tres pasos:

1. Se construye la estructura preliminar que representa a cada par (q,w) $\forall w_i \in \Sigma_q$ y $\forall q \in Q^K$. Se crea un nuevo estado cada vez que se encuentra un k-grama, k=1,..,K distinto a los previamente observados en el corpus de entrenamiento. Se incrementa en una unidad la cuenta correspondiente $N(w)$ y se reserva una posición adicional por estado para ser posteriormente asociada a la transición de back-off, etiquetada como U. La estructura de tipo array contiene los datos obtenidos a partir de las muestras de entrenamiento: $|\Sigma_q|$, $w_i \in \Sigma_q$ y $N(w)$ $\forall w_i \in \Sigma_q$.
2. Se completa la estructura rellenando el índice del estado destino asociado a cada transición en la tabla. El camino para encontrar el índice adecuado depende del tipo de transición:

- Las transiciones a partir del estado etiquetado como *nil* (tipo (a)): cada una de las $|\Sigma|$ transiciones etiquetadas por w_i i=1...$|\Sigma|$ encuentran el índice del estado destino en la primera posición del estado etiquetado por w_i en el nivel k=2.
- Las transiciones desde el estado inicial etiquetado por \$ (tipo (b)): cada una de las $|\Sigma_\$|$ transiciones etiquetadas por w_i i=1...$|\Sigma_\$|$ encuentran su índice destino en la primera posición asociada al estado etiquetado como $\$w_i$ en el nivel k=3.
- Las transiciones desde los estados asociados al nivel k a estados asociados a los niveles k+1, k=2,...,K-1 (tipo (c)): cada una de las $|\Sigma_q|$ transiciones etiquetadas por w_i i=1...$|\Sigma_q|$ asociadas a cada estado q etiquetado como w_{i-k}^{i-1}, $q \in Q^k$ k=2...K-1, encuentran el índice asociado a su destino en la primera posición asociada al estado etiquetado por w_{i-k}^{i} en el nivel k+1.

- Las transiciones entre estados asociados al nivel superior K (tipo (d)): cada una de las $|\Sigma_q|$ transiciones etiquetadas por w_i i=1...$|\Sigma_q|$ asociadas a cada estado q etiquetado por $w_{i-(K-1)}^{i-1}$ encuentran el índice asociado a su destino en la primera posición asociada al estado etiquetado por w_{i-K}^{i-1} en el mismo nivel K.
- Las transiciones al estado final etiquetado como $ (tipo (e)): cada transición etiquetada por $ asociada a cada estado $q \in Q^K$ encuentra el índice asociado a su destino en la primera posición asociada al estado etiquetado como $, es decir, la posición $|\Sigma|+1$ del array.
- Transiciones de Back-off: cada transición etiquetada por U asociado a cada estado $q \in Q^k$ con k=2...K, encuentra el índice asociado a cada destino en la primera posición asociado a su estado de back-off b_q, en el nivel k-1. Así, las transiciones etiquetadas por U asociadas al estado $q \equiv w_{i-(K-1)}^{i-1}$ del nivel K encuentran el índice de su destino en la primera posición asociada al estado $q \equiv w_{i-(K-1)+1}^{i-1}$ en el nivel K-1.

3. Por último, se calcula la probabilidad asociada a cada transición. Primero, se obtiene la probabilidad $P(w/q)$ según la ecuación 3.13 para $w_i \in \Sigma_q$, después se obtiene $P(b_q /q)$ usando la ecuación 3.15 y se inserta en cada posición del array etiquetada por U.

Apéndice C

Resultados de decodificación dependientes del conjunto de frases de evaluación

A lo largo de la memoria se ha observado la fuerte dependencia existente entre la distribución de probabilidad proporcionada por las técnicas de suavizado y el valor del parámetro de ponderación de las probabilidades del ML α, necesario para alcanzar la eficiencia máxima del sistema. Se ha visto también que los valores de la perplejidad, cuando las diferencias no son realmente grandes, no predicen con una absoluta fiabilidad el comportamiento del ML dentro del sistema completo de RAH.

En general, se obtienen mejores resultados de perplejidad (más bajos) para las distribuciones de probabilidad que son el resultado de un mayor suavizado (para una distribución de probabilidad poco dispersa, en el rango de valores permitidos de la probabilidad). Sin embargo, con estas distribuciones se obtiene un peor comportamiento del porcentaje de error (%PRE) cuando el factor de ponderación de las probabilidades del ML se mantiene igual a la unidad y es necesario un valor de α bastante alto para alcanzar el %PRE óptimo que en general es bastante bajo.

El objetivo de este apéndice es observar el comportamiento de la perplejidad (PP) y de las tasas de error (%PRE) cuando se separan las frases de reconocimiento de la base de datos BDGEO (600 frases, ver apéndice A) en dos conjuntos: bdgeo_464 contiene aquellas 464 frases que proporcionan mejores resultados de perplejidad (más bajos), y bdgeo_136 las 136 restantes, es decir, aquellas que presentan unos valores de perplejidad más altos. No se trata de hacer un estudio exhaustivo, lo cual implicaría utilizar un método de "*leaving one out*", sino que lo que se pretende es una ligera aproximación al problema.

En la figura C.1 se muestran los resultados de perplejidad obtenidos con los modelos integrados de orden K (K modelos *k*-EE con K=2...6) suavizados con el descuento de Witten-Bell sobre los tres conjuntos de frases mencionados: bdgeo_600 representa al conjunto total de 600 frases que contiene 5655 palabras, bdgeo_464 es el conjunto de 464 frases que contienen 4254 palabras y bdgeo_136 es el conjunto de 136 frases que contiene 1401 palabras.

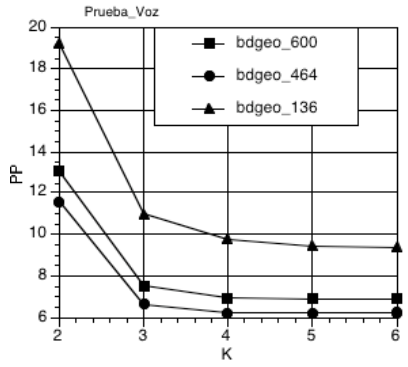


Figura C.1.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) tras la separación del conjunto de evaluación (bdgeo_600) en dos, en función de los resultados de perplejidad. bdgeo_464 contiene las frases con mejores resultados de perplejidad y bdgeo_136 contiene las frases que dan los peores resultados de perplejidad.

La mejora en los valores de la perplejidad observada para todos los valores de K con el conjunto bdgeo_464 es mucho menos significativa que el empeoramiento observado para el conjunto bdgeo_136, con respecto a los valores obtenidos con el conjunto total bdgeo_600. Para valores altos de K se observan unos valores constantes de la perplejidad para los conjuntos bdgeo_600 y bdgeo_464, sin embargo para el conjunto bdgeo_136 los valores de la perplejidad siguen disminuyendo ligeramente con K.

En la figura C.2 se pueden observar los resultados de decodificación obtenidos cuando estos mismos modelos fueron integrados en el sistema de decodificación utilizando la regla de Bayes directamente. En la figura C.2a se muestra el %PRE obtenido frente a K. En la figura C.2b se puede ver también el Número Medio de Nodos Activos (NMNA) por frame para los distintos modelos integrados de orden K (K=2...6), y en la figura C.2c se muestra el %PRE frente al NMNA (base logarítmica)

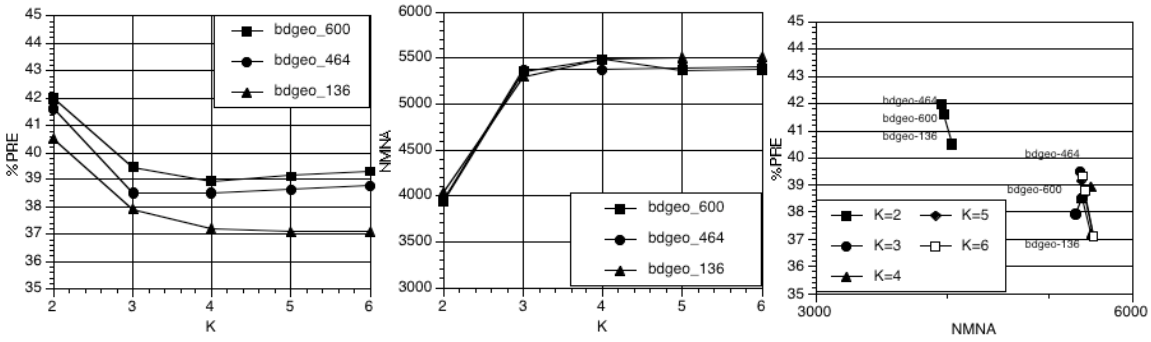


Figura C.2.- Resultados de error obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) y los diferentes conjuntos bdgeo_600, bdgeo_464 y bdgeo_136 (tarea BDGEO).

Los mejores (más bajos) porcentajes de error se obtienen para el conjunto bdgeo_136 manteniéndose prácticamente constantes para valores altos de K. Sin embargo con los conjuntos bdgeo_600 y bdgeo_464 la tendencia es hacia una subida ligera del error a partir de K=4. El NMNA prácticamente se mantiene constante independientemente del conjunto que se esté evaluando.

El suavizado sobre los modelos integrados de orden K se ha realizado utilizando el descuento de Witten-Bell. Como se ha visto en el capítulo 5 se trata de un descuento que proporciona unos modelos bastante suavizados ya que otorga bastante probabilidad a los eventos no vistos en entrenamiento. Según los resultados observados, cuando los modelos se integran directamente en el sistema de decodificación, las frases con situaciones menos vistas en entrenamiento (con peores resultados de perplejidad) son las que mejor se reconocen a pesar de que el algoritmo de decodificación utiliza la técnica de búsqueda en haz para desactivar los caminos menos probables (ver capítulo 4). Esto significa que las situaciones no vistas en entrenamiento están teniendo probabilidades muy altas, puede que incluso mayores que las obtenidas para alguna de las situaciones vistas.

En la figura C.3, se presentan los resultados de error obtenidos para cada uno de los modelos integrados de orden K con K=2...6, cuando se integran en el sistema de decodificación pero aplicando un factor de ponderación sobre las probabilidades del ML (ver capítulo 4). Como en capítulos anteriores, los puntos situados el margen inferior izquierdo de cada figura son los que representan la mayor eficacia del sistema, es decir el %PRE más bajo junto con un menor número medio de nodos activos (NMNA).

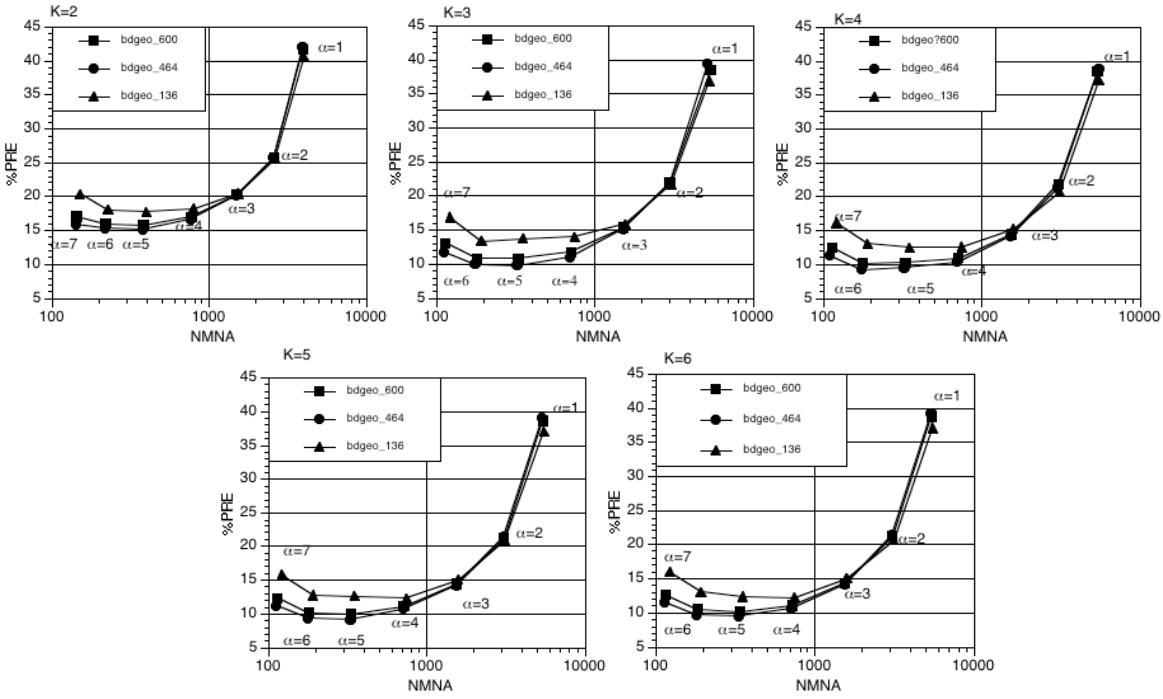


Figura C.3.- Resultados de error obtenido sobre los modelos integrados de orden K (K=2, 3, 4, 5 y 6) cuando se aplica el descuento de Witten-Bell sobre los tres conjuntos de frases: bdgeo_600, bdgeo_464 y bdgeo_136 y diferentes valores del parámetro de escalado de las probabilidades del ML (α).

Para valores bajos (inferiores al óptimo) del parámetro de ponderación de las probabilidades del ML, se obtienen prácticamente los mismos resultados sobre los tres conjuntos evaluados. Con valores de $\alpha>4$ los resultados para el conjunto bdgeo_136 son algo inferiores (porcentajes de error más altos). Parece que la dispersión de las probabilidades del ML que provoca el parámetro α está haciendo que dejen de activarse cada vez más caminos,

alguno de los cuales corresponde a las situaciones no vistas en entrenamiento que están representadas en las frases del conjunto bdgeo_136.

C.1.- Evaluación de los distintos métodos de descuento sobre los ML

En este punto vamos a observar que es lo que ocurre cuando se aplican distintos descuentos sobre los modelos integrados de orden K con K=2...6. En concreto se van a evaluar los principales descuentos presentados en el capítulo 5, descuento Sencillo (dS), descuento Absoluto (dAb b=0.4 y b=0.01), descuento Acotado (dAc d=0.7 y d=0.99) y descuento Lineal (dL l=0.1 y l=0.01).

En la figura C.4 se muestran los resultados de perplejidad obtenidos para los distintos descuentos y en la figura C.5 en detalle para aquellos descuentos que presentan resultados más parecidos. Las diferencias en los resultados observados son más notables para el conjunto bdgeo_136, pero básicamente las relaciones se mantienen. Quizá la única diferencia es que para el conjunto bdgeo_464 los mejores resultados se consiguen con el descuento Absoluto b=0.4, mientras que para el conjunto bdgeo_136 los mejores resultados se obtienen con el descuento de Witten-Bell, si bien, las diferencias no son en absoluto significativas.

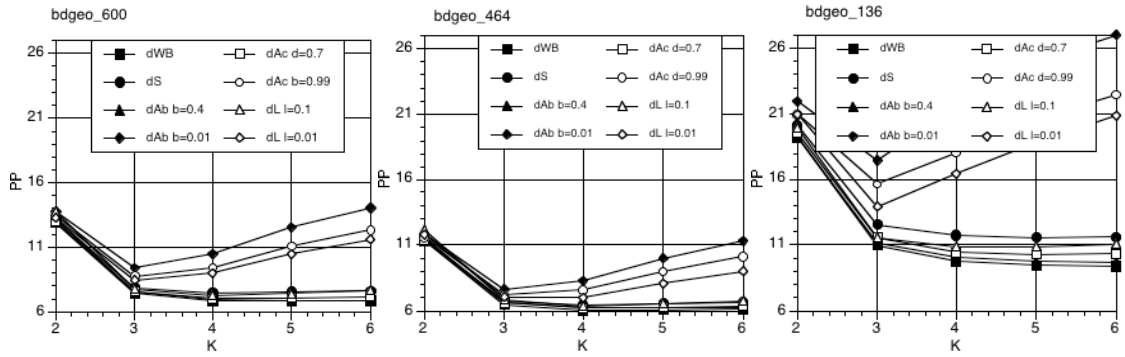


Figura C.4.- Resultados de perplejidad obtenidos para los modelos integrados de orden K (K=2... 6) con los conjuntos de 600, 464 y 136 frases respectivamente.

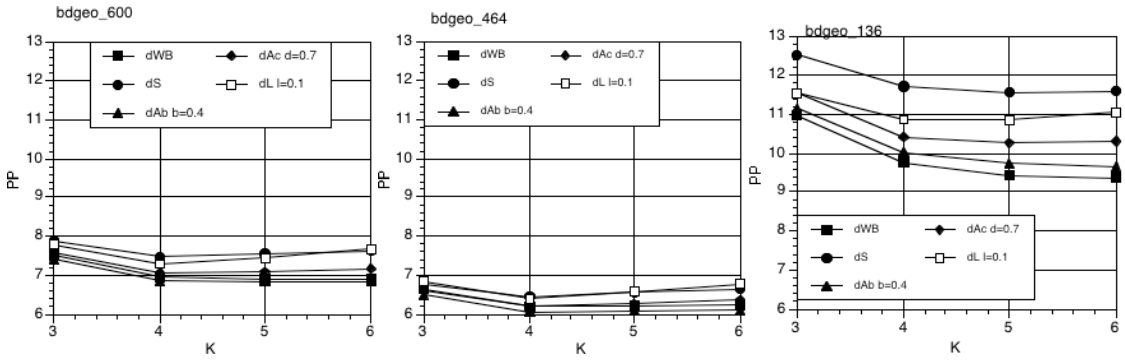


Figura C.5.- Resultados de perplejidad en detalle obtenidos para los modelos integrados de orden K presentados en la figura C.4.

A continuación se presentan los resultados de decodificación obtenidos cuando los modelos integrados de orden K suavizados con los distintos descuentos se integraron en el sistema de decodificación utilizando directamente la regla de Bayes. En la figura C.6 se pueden ver los resultados de error y en la figura C.7 se muestra el Número Medio de Nodos Activos (NMNA) obtenido en función del valor de K.

Se observa que los modelos integrados de orden K (K=2...6) con los descuentos que proporcionan una distribución de probabilidad más suavizada: descuento de Witten-Bell, descuento Absoluto b=0.4, descuento Lineal l=0.1, descuento Acotado d=0.7 logran prácticamente las mismas tasas de error para los tres conjuntos (algo mejores para el conjunto bdgeo_136 con los descuentos de Witten-Bell y Lineal l=0.1). Sin embargo, cuando se utilizan descuentos que proporcionan modelos menos suavizados (descuento Lineal l=0.01, descuento Acotado d=0.99 y descuento Absoluto b=0.01) los peores resultados e obtienen para el conjunto bdgeo_136. Los resultados que se obtienen en NMNA prácticamente se mantienen para todos los descuentos.

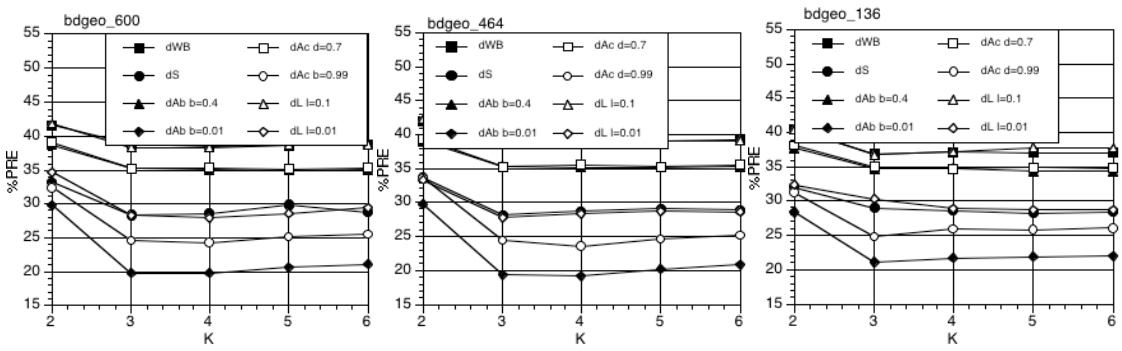


Figura C.6.- Resultados de error obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) con los conjuntos de 600, 464 y 136 frases respectivamente.

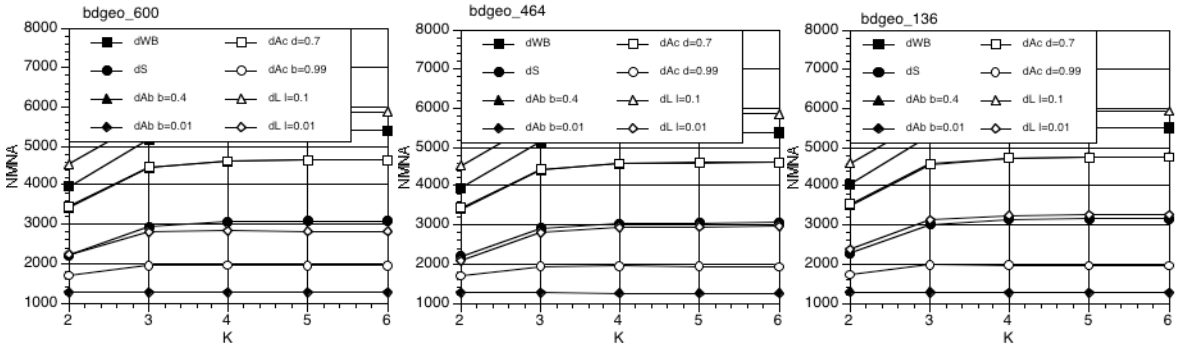


Figura C.7.- Resultados del NMNA obtenidos para los modelos integrados de orden K (K=2, 3, 4, 5 y 6) con los conjuntos de 600, 464 y 136 frases respectivamente.

En la figura C.8 se pueden ver los resultados de decodificación obtenidos en función del parámetro de ponderación de las probabilidades del ML α para el modelo K-EE con K=4. Para el resto de los modelos con distintos valores K, los resultados obtenidos son muy parecidos por lo que las conclusiones que se derivan son prácticamente las mismas. En la figura C.9 se pueden ver los mismos resultados obtenidos alrededor del valor óptimo del porcentaje de error.

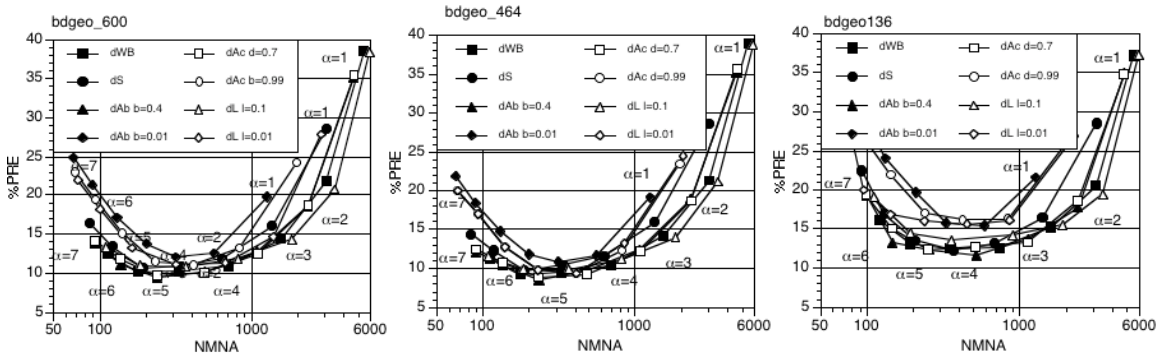


Figura C.8.- Resultados de error obtenidos para el modelos integrados de orden K con K=4 con los conjuntos de 600, 464 y 136 frases respectivamente.

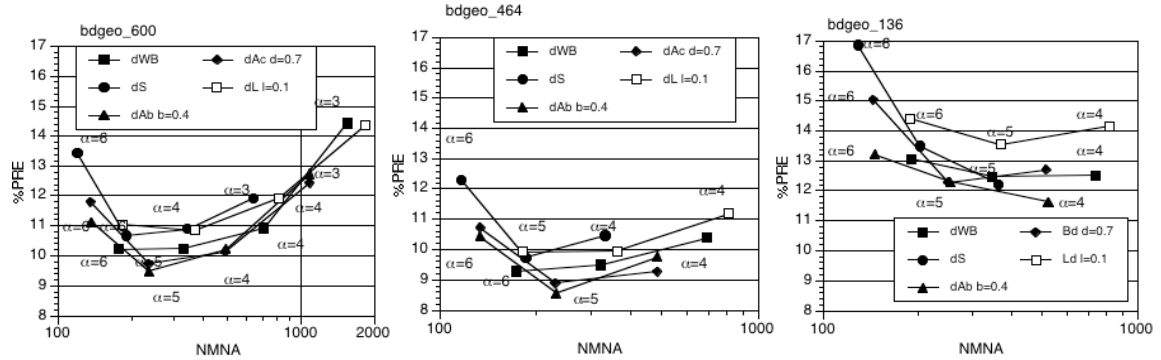


Figura C.9-. Resultados de error (en detalle) obtenidos para el modelo K-EE con K=4 con los conjuntos de 600, 464 y 136 frases respectivamente.

Cuando se evalúa el conjunto bdgeo_136 se observa que los modelos integrados de orden K poco suavizados con descuentos que proporcionaban baja probabilidad para los eventos no vistos en entrenamiento (descuento Absoluto b=0.01, descuento Acotado d=0.99 y descuento Lineal l=0.01) resultan con valores bastante más altos del %PRE. En el resto de los descuentos evaluados, el comportamiento de los distintos modelos integrados de orden K es muy similar independientemente de que las frases sean más favorables (bdgeo_464) o desfavorables (bdgeo_136) con la lógica diferencia del valor promedio sobre el que se sitúan las tasas de error.

Se puede concluir que el trabajar con frases "fáciles" (bdgeo_464) o "difíciles" (bdgeo_136) pueden influir notablemente en la evaluación de los descuentos aplicados sobre los ML usando como medida la perplejidad. Sin embargo estas diferencias no se manifiestan en decodificación donde lo más importante es trabajar con el valor del parámetro de ponderación de las probabilidades del ML más adecuado en cada caso, reafirmando las conclusiones obtenidas en los capítulos 5 y 6.