

Estudio y modelización acústica del habla espontánea en diálogos hombre-máquina y entre personas

Memoria presentada para optar al grado de doctor por:
Luis Javier Rodríguez Fuentes

Dirigida por:
Dra. María Inés Torres Barañano

Leioa, mayo de 2004

Departamento de Electricidad y Electrónica
Facultad de Ciencia y Tecnología
Universidad del País Vasco

Índice General

Agradecimientos	XIX
1. Introducción.	1
1.1. Contexto y motivación del trabajo.	1
1.2. Objetivos y metodología.	2
1.3. Estructura de la memoria.	4
2. Esquema de anotación.	7
2.1. Inventario de fenómenos.	8
2.1.1. Disfluencias y fenómenos de habla espontánea.	8
2.1.2. Anotación: perspectiva histórica.	9
2.1.2.1. ATIS y Switchboard.	9
2.1.2.2. TRAINS y Map Task.	10
2.1.2.3. El proyecto MATE.	11
2.1.3. Nivel acústico.	11
2.1.3.1. Ruidos.	11
2.1.3.2. Pausas de silencio.	12
2.1.3.3. Pausas <i>habladas</i>	13
2.1.3.4. Alargamientos.	13
2.1.3.5. Solapamientos.	14
2.1.4. Nivel léxico.	14
2.1.4.1. Variantes de pronunciación.	15
2.1.4.2. Palabras cortadas.	15
2.1.4.3. Sonidos guturales de aceptación o negación.	15
2.1.4.4. Siglas.	16
2.1.4.5. Palabras extranjeras.	16
2.1.5. Nivel sintáctico.	16
2.1.5.1. Reformulaciones.	17
2.1.5.2. Frases abandonadas.	18
2.1.6. Nivel pragmático.	19
2.1.6.1. Marcadores de discurso.	19
2.2. El formato de anotación.	20
2.2.1. El problema de la sincronización.	21
2.2.2. ¿Uno o varios ficheros?	21
2.2.3. Anotación: estado del arte.	22
2.2.4. Un formato de anotación de fenómenos de habla espontánea.	23

2.3.	La herramienta de anotación.	30
2.3.1.	Formato de anotación simplificado.	32
2.3.2.	Guías de anotación.	32
2.3.2.1.	Guía para diálogos hombre-máquina.	34
2.3.2.2.	Guía para conversaciones naturales entre personas.	35
3.	Estudio de los fenómenos.	41
3.1.	INFOTREN: diálogos hombre-máquina.	42
3.2.	CORLEC: diálogos naturales entre personas.	44
3.2.1.	Definición de un corpus reducido: CORLEC-EHU.	45
3.2.2.	Filtrado de las transcripciones originales de CORLEC-EHU.	47
3.2.3.	El subcorpus CORLEC-EHU-1.	48
3.3.	Distribución de fenómenos.	49
3.3.1.	Distribución de fenómenos en INFOTREN.	49
3.3.1.1.	Ruidos.	51
3.3.1.2.	Disfluencias acústicas.	51
3.3.1.3.	Distorsiones léxicas.	53
3.3.1.4.	Disfluencias sintácticas.	54
3.3.1.5.	Marcadores de discurso.	55
3.3.2.	Distribución de fenómenos en CORLEC-EHU-1.	55
3.3.2.1.	Ruidos.	56
3.3.2.2.	Disfluencias acústicas.	56
3.3.2.3.	Distorsiones léxicas.	58
3.3.2.4.	Disfluencias sintácticas.	60
3.3.2.5.	Marcadores de discurso.	60
3.4.	Dependencia con respecto a la duración.	61
3.4.1.	Dependencia del número de fenómenos con respecto a la duración en INFOTREN.	61
3.4.2.	Dependencia del número de fenómenos con respecto a la duración en CORLEC-EHU-1.	63
3.4.3.	Comparación con trabajos previos realizados sobre bases de datos en inglés.	64
3.5.	Dependencia con respecto al hablante.	68
3.5.1.	Dependencia del número de fenómenos con respecto al hablante en INFOTREN.	68
3.5.2.	Dependencia del número de fenómenos con respecto al hablante en CORLEC-EHU-1.	70
3.6.	Detección de disfluencias sintácticas.	74
3.6.1.	Fenómenos acústicos y disfluencias sintácticas.	74
3.6.1.1.	Resultados sobre INFOTREN.	74
3.6.1.2.	Resultados sobre CORLEC-EHU-1.	75
3.6.2.	Distorsiones léxicas y disfluencias sintácticas.	76
3.6.2.1.	Resultados sobre INFOTREN.	76
3.6.2.2.	Resultados sobre CORLEC-EHU-1.	76
3.6.3.	Marcadores de discurso y disfluencias sintácticas.	77
3.6.3.1.	Resultados sobre INFOTREN.	77

3.6.3.2.	Resultados sobre CORLEC-EHU-1.	78
4.	El sistema de referencia.	81
4.1.	Descripción general	81
4.1.1.	Los parámetros acústicos.	81
4.1.2.	La cuantificación vectorial.	83
4.1.3.	El reconocedor.	84
4.1.3.1.	El conjunto de unidades subléxicas.	84
4.1.3.2.	Las transcripciones fonéticas.	84
4.1.3.3.	El léxico de la tarea.	86
4.1.3.4.	El modelo de lenguaje.	87
4.1.3.5.	Evaluación del sistema de reconocimiento integrado.	88
4.2.	Los modelos acústicos.	89
4.2.1.	Estimación de los modelos acústicos.	89
4.2.1.1.	El algoritmo de Baum-Welch.	90
4.2.1.2.	El algoritmo de entrenamiento por Viterbi.	91
4.2.1.3.	Entrenamiento híbrido.	92
4.2.1.4.	Fórmulas de reestimación.	93
4.2.2.	Inicialización de los modelos continuos y semicontinuos.	93
4.2.3.	Evaluación de los modelos acústicos.	93
4.3.	Decodificación acústico-fonética	94
4.3.1.	Bases de datos.	94
4.3.1.1.	Distribución de fonemas.	96
4.3.1.2.	Cuantificación vectorial.	98
4.3.1.3.	Matrices de covarianza.	99
4.3.2.	Resultados con MOM discretos.	101
4.3.2.1.	MOM discretos sobre habla leída.	102
4.3.2.2.	MOM discretos sobre habla espontánea.	103
4.3.2.3.	Algunas conclusiones sobre MOM discretos.	104
4.3.3.	Resultados con MOM continuos y semicontinuos.	104
4.3.3.1.	MOM continuos y semicontinuos sobre habla leída.	104
4.3.3.2.	MOM continuos sobre habla espontánea.	105
4.4.	Experimentos de reconocimiento.	106
5.	Modelización de los fenómenos.	115
5.1.	Ampliación del conjunto de unidades	115
5.1.1.	Un conjunto alternativo de unidades subléxicas.	118
5.2.	Resultados de DAF.	119
5.2.1.	Resultados de DAF sobre INFOTREN.	120
5.2.1.1.	Análisis de la matriz de confusión.	121
5.2.1.2.	Un ejemplo.	123
5.2.1.3.	Correlación entre la tasa de DAF y el número de fenómenos.	123
5.2.2.	Resultados de DAF sobre CORLEC-EHU-1.	125
5.2.2.1.	Análisis de la matriz de confusión.	125
5.2.2.2.	Un ejemplo.	128
5.2.2.3.	Correlación entre la tasa de DAF y la densidad de fenómenos.	130
5.2.3.	Resultados de DAF con unidades contextuales.	130

5.3.	Resultados de reconocimiento.	134
5.3.1.	Resultados de reconocimiento con unidades contextuales.	139
6.	Modelización del hablante.	143
6.1.	Alternativas de modelización.	143
6.1.1.	Normalización del hablante.	144
6.1.1.1.	Normalización de la longitud del tracto vocal.	144
6.1.2.	Adaptación al hablante.	145
6.1.2.1.	Adaptación MAP.	145
6.1.2.2.	Transformación de los parámetros de los modelos.	146
6.1.2.3.	<i>Clustering</i> de hablantes.	147
6.2.	<i>Clustering</i> de hablantes.	148
6.2.1.	Un modelo acústico simplificado para grupos de hablantes.	148
6.2.2.	Medida de disimilitud entre un hablante y un grupo de hablantes.	150
6.2.3.	Medida de disimilitud entre dos hablantes.	150
6.2.4.	El algoritmo de <i>clustering</i>	151
6.2.5.	Estimación de los MOM de grupos de hablantes.	153
6.2.6.	Alternativas de adaptación al hablante.	153
6.2.6.1.	Selección de MOM basada en las probabilidades de las decodificaciones.	153
6.2.6.2.	Selección de MOM basada en las probabilidades obtenidas de los modelos de histogramas.	155
6.2.6.3.	Selección de MOM basada en las probabilidades de un subconjunto de decodificaciones.	155
6.2.6.4.	Combinación de MOM basada en las probabilidades obtenidas de los modelos de histogramas.	155
6.2.7.	Resultados del <i>clustering</i>	156
6.2.7.1.	<i>Clustering</i> de hablantes en SENGLAR16.	156
6.2.7.2.	<i>Clustering</i> de hablantes en INFOTREN.	160
6.2.7.3.	<i>Clustering</i> de hablantes en CORLEC-EHU-1.	161
6.2.7.4.	Evolución de la distancia promedio de un hablante a su grupo.	165
6.3.	Modelos de género.	166
6.4.	Resultados de DAF.	169
6.4.1.	Ajuste de τ en la estimación MAP.	169
6.4.2.	Alternativas de adaptación.	170
6.4.3.	Resultados de DAF con modelos de género.	174
6.4.4.	Resultados de DAF con unidades subléxicas contextuales.	175
6.5.	Resultados de reconocimiento.	178
7.	Conclusiones y trabajo futuro.	183
7.1.	Conclusiones.	183
7.1.1.	Publicaciones relacionadas con esta tesis.	186
7.2.	Trabajo futuro.	187
A.	DTD del formato de anotación	189

B. Ejemplos de anotación.	193
B.1. Diálogo hombre-máquina.	193
B.1.1. Versión en formato simplificado.	193
B.1.2. Versión en formato XML.	194
B.2. Conversación entre personas.	201
B.2.1. Versión en formato simplificado.	201
B.2.2. Versión en formato XML.	204
C. Fórmulas de reestimación	215
C.1. Notación.	215
C.2. Algoritmo de Baum-Welch restringido.	216
C.3. Algoritmo de Baum-Welch general.	218
C.4. Algoritmo de entrenamiento por Viterbi.	221
D. Resultados con la parametrización del ETSI.	223
Bibliografía	225

Índice de Figuras

2.1. Cortado de señales en presencia de solapamientos.	36
3.1. Histograma de los diálogos con respecto al número de turnos en INFOTREN.	44
3.2. Histograma de los turnos de usuario con respecto a la duración de los mismos en INFOTREN.	44
3.3. Histograma de los diálogos con respecto al número de turnos útiles en CORLEC-EHU-1.	50
3.4. Histograma de los turnos útiles con respecto a la duración de los mismos en CORLEC-EHU-1.	50
3.5. Número efectivo de palabras (PE) en los turnos de INFOTREN con respecto a la duración de los mismos (t, en segundos). Un análisis por regresión lineal resulta en el siguiente ajuste: $PE = 1.831 * t + 0.9594$ (two tailed F-test, $F(1,1655)=6616.07$, $p=0.0000$).	64
3.6. Número de disfluencias acústicas y sintácticas (D), normalizado por el número efectivo de palabras (PE), en función del número efectivo de palabras en cada turno de INFOTREN. Análisis por regresión lineal: $D/PE = 0.001509 * PE + 0.156375$ (two tailed F-test, $F(1,1655)=5.3904$, $p=0.0204$).	66
3.7. Número de disfluencias acústicas y sintácticas (D), normalizado por el número efectivo de palabras (PE), en función del número efectivo de palabras en cada turno de CORLEC-EHU-1. Análisis por regresión lineal: $D/PE = -0.0003222 * PE + 0.154957$ (two tailed F-test, $F(1,2807)=3.6188$, $p=0.0572$).	66
3.8. Logaritmo de la probabilidad de que un turno sea fluido en función de la longitud del turno, en el caso de INFOTREN. Ajuste por regresión lineal: $\log Prob_{fluent}(n) = -0,1235 * n - 0,0730$ (two tailed F-test, $F(1,23)=113.51$, $p=0.0000$).	67
3.9. Logaritmo de la probabilidad de que un turno sea fluido en función de la longitud del turno, en el caso de CORLEC-EHU-1. Ajuste por regresión lineal: $\log Prob_{fluent}(n) = -0,1058 * n + 0,1432$ (two tailed F-test, $F(1,23)=127.7135$, $p=0.0000$).	67
3.10. Histograma de los hablantes de INFOTREN con respecto al número de disfluencias acústicas.	69
3.11. Histograma de los hablantes de INFOTREN con respecto al número de disfluencias sintácticas.	69
3.12. Histograma de los hablantes de CORLEC-EHU-1 con respecto al número de disfluencias acústicas.	72
3.13. Histograma de los hablantes de CORLEC-EHU-1 con respecto al número de disfluencias sintácticas.	72

4.1.	El sistema de reconocimiento.	82
4.2.	Topología de los MOM de unidades subléxicas.	90
4.3.	Matrices de correlación de los coeficientes cepstrales, calculadas sobre los conjuntos de entrenamiento de SENGLAR16 (a), SENGLAR08 (b), INFOTREN (c) y la partición 2 de CORLEC-EHU-1 (d).	101
5.1.	Segmentaciones forzada y reconocida obtenidas con el conjunto ampliado (las dos primeras, empezando por arriba), el conjunto alternativo (la tercera y cuarta) y el conjunto básico de unidades (las dos últimas), para la frase " <i>quisiera sabEr WA M</i> ".	124
5.2.	Segmentaciones forzada y reconocida obtenidas con el conjunto ampliado (las dos primeras, empezando por arriba), el conjunto alternativo (la tercera y cuarta) y el conjunto básico de unidades (las dos últimas), para la frase " <i>G y vAn WM observando</i> ".	129
6.1.	Se elige como grupo a dividir aquél, g^* , cuya distorsión interna sea máxima. Los dos grupos en los que se divide g^* se inicializan con los modelos del representante $r(g^*)$ y del hablante más próximo a éste, $r'(g^*)$	152
6.2.	Algoritmo de <i>clustering</i>	154
6.3.	Probabilidades normalizadas (ω_g) de las intervenciones de un hablante del grupo 1 de SENGLAR16, obtenidas mediante los modelos de histogramas. El modelo del grupo 1 proporciona las probabilidades más altas en todos los casos.	158
6.4.	Porcentaje de intervenciones del corpus de entrenamiento de SENGLAR16 asignadas al grupo correcto, en función de la longitud del segmento de señal sobre el que se aplican los modelos de histogramas.	158
6.5.	Probabilidades normalizadas (ω_g) de las intervenciones de un hablante del grupo 6 de INFOTREN, obtenidas mediante los modelos de histogramas. Los modelos de los grupos 5 y 6 proporcionan las probabilidades más altas en todos los casos.	161
6.6.	Porcentaje de intervenciones del corpus de entrenamiento de INFOTREN asignadas al grupo correcto, en función de la longitud del segmento de señal sobre el que se aplican los modelos de histogramas.	162
6.7.	Probabilidades normalizadas (ω_g) de las intervenciones de un hablante del grupo 6 de la primera partición de CORLEC-EHU-1, obtenidas mediante los modelos de histogramas. El modelo del grupo 6 proporciona las probabilidades más altas en todos los casos.	165
6.8.	Porcentaje de intervenciones asignadas al grupo correcto en los conjuntos de entrenamiento correspondientes a las tres particiones de CORLEC-EHU-1, en función de la longitud del segmento de señal sobre el que se aplican los modelos de histogramas.	166
6.9.	Evolución de la distancia promedio de un hablante a su grupo $R(n)$ para las tres bases de datos consideradas en este capítulo.	167
6.10.	Evolución de la distancia promedio de un hablante a su grupo $R(n)$, normalizada por el valor inicial $R(1)$, para las tres bases de datos consideradas en este capítulo.	167
6.11.	Tasas de DAF obtenidas sobre el corpus de test de SENGLAR16, utilizando modelos de grupos de hablantes con el criterio MAP, para distintos valores del factor de adaptación τ : (a) MOM discretos, (b) MOM continuos de 32 gaussianas.	169
6.12.	Tasas de DAF obtenidas sobre los corpus de test de INFOTREN (a) y las tres particiones de CORLEC-EHU-1 (b), utilizando modelos de grupos de hablantes (MOM continuos de 32 gaussianas) con el criterio MAP, para distintos valores del factor de adaptación τ	170

6.13. Probabilidades normalizadas (ω_g) de las intervenciones de una hablante del corpus de test de SENGLAR16, obtenidas mediante los modelos de histogramas. El modelo del grupo 4 proporciona las probabilidades más altas en todos los casos.	171
6.14. Probabilidades normalizadas (ω_g) de las intervenciones de una hablante del corpus de test de INFOTREN, obtenidas mediante los modelos de histogramas. Los modelos de los grupos 2 y 8 proporcionan las probabilidades más altas en la mayor parte de los casos.	172
6.15. Probabilidades normalizadas (ω_g) de las intervenciones de un hablante del corpus de test de la partición 3 de CORLEC-EHU-1, obtenidas mediante los modelos de histogramas. El modelo del grupo 1 proporciona las probabilidades más altas en casi todos los casos.	172

Índice de Tablas

2.1. Equivalencias entre los elementos y atributos XML y las marcas en formato simplificado.	33
2.2. Marcas especiales.	33
3.1. Principales características de la base de datos de diálogos hombre-máquina INFOTREN.	45
3.2. Estadísticas de los diferentes bloques que conforman la base de datos CORLEC: número de palabras, tamaño del léxico y promedio de muestras por palabra (M/P). El tamaño del vocabulario es de 39785 palabras, sobre un total de 941386.	46
3.3. Estadísticas de fenómenos de habla espontánea presentes en las anotaciones originales de CORLEC, desglosadas por bloques: número absoluto de fenómenos (NFHE), porcentaje que este número representa sobre el total (%) y <i>ratio</i> de fenómenos por palabra (FHE/P).	47
3.4. Principales características de la base de datos de diálogos naturales entre personas CORLEC-EHU-1.	49
3.5. Estadísticas de los fenómenos de habla espontánea anotados en INFOTREN: número absoluto de apariciones (#FHE), número promedio de fenómenos por cada 100 palabras efectivas (%FHE/PE), porcentaje de fenómenos con respecto a la categoría a la que pertenecen (%FHE/C) y porcentaje que representa cada categoría con respecto al total (%C/Total).	52
3.6. Cuentas absolutas y relativas de los alargamientos anotados en INFOTREN.	53
3.7. Lista de los 20 alargamientos más frecuentes en INFOTREN.	54
3.8. Número de disfluencias sintácticas simples y compuestas en INFOTREN.	55
3.9. Los 10 marcadores de discurso más frecuentes en INFOTREN representan un 65 % del total.	55
3.10. Estadísticas de los fenómenos de habla espontánea anotados en CORLEC-EHU-1: número absoluto de apariciones (#FHE), número promedio de fenómenos por cada 100 palabras efectivas (%FHE/PE), porcentaje de fenómenos con respecto a la categoría a la que pertenecen (%FHE/C) y porcentaje que representa cada categoría con respecto al total (%C/Total).	57
3.11. Cuentas absolutas y relativas de los alargamientos anotados en CORLEC-EHU-1.	59
3.12. Lista de los 20 alargamientos más frecuentes en el subcorpus CORLEC-EHU-1. Los sonidos alargados aparecen subrayados.	59
3.13. Número de disfluencias sintácticas simples y compuestas en CORLEC-EHU-1.	60

3.14. Marcadores de discurso en CORLEC-EHU-1 desglosados por categorías: número de expresiones distintas, número total de instancias, número de instancias correspondientes a las 5 expresiones más frecuentes, y porcentaje sobre el total que representan éstas últimas.	61
3.15. Media, desviación típica y valor máximo del número de fenómenos de habla espontánea en los 1657 turnos de usuario de INFOTREN.	62
3.16. Coeficientes de correlación de Pearson para el número de fenómenos en un turno con respecto a la duración de dicho turno, calculados sobre los 1657 turnos de INFOTREN (two-tailed t-test, $p=0.0000$). R: ruidos, A: disfluencias acústicas, L: distorsiones léxicas, S: disfluencias sintácticas, D: marcadores de discurso, t: duración del turno.	62
3.17. Media, desviación típica y valor máximo del número de fenómenos de habla espontánea en los 2873 turnos de CORLEC-EHU-1.	63
3.18. Coeficientes de correlación de Pearson para el número de fenómenos en un turno con respecto al número efectivo de palabras de dicho turno, calculados sobre los 2873 turnos de CORLEC-EHU-1. R: ruidos, A: disfluencias acústicas, L: distorsiones léxicas, S: disfluencias sintácticas, D: marcadores de discurso, PE: número efectivo de palabras.	63
3.19. Media, desviación típica y valor máximo del número de fenómenos para los 75 hablantes de INFOTREN, en las 5 categorías habituales.	68
3.20. Coeficientes de correlación entre el número de fenómenos y el tiempo de elocución, calculados sobre el conjunto de hablantes de INFOTREN (two tailed t-test, $p=0.0000$). R: ruidos, A: disfluencias acústicas, L: distorsiones léxicas, S: disfluencias sintácticas, D: marcadores de discurso, t: tiempo total de elocución de un hablante.	70
3.21. Tiempo de elocución (t) y número de fenómenos de habla espontánea para 6 hablantes especialmente seleccionados de INFOTREN. Se muestran asimismo el número de fenómenos por unidad de tiempo para cada hablante, así como la media y la desviación típica de las cuentas de fenómenos, para una mejor evaluación de los datos.	71
3.22. Media, desviación típica y valor máximo del número de fenómenos en las 5 categorías habituales, para los 98 hablantes de CORLEC-EHU-1 que suman más de 120 palabras efectivas.	71
3.23. Coeficientes de correlación entre el número de fenómenos y el número efectivo de palabras de cada hablante, calculados sobre el conjunto de 98 hablantes estadísticamente útiles de CORLEC-EHU-1 (two tailed t-test, $p=0.0000$). R: ruidos, A: disfluencias acústicas, L: distorsiones léxicas, S: disfluencias sintácticas, D: marcadores de discurso, PE: número efectivo de palabras.	73
3.24. Número efectivo de palabras (PE) y número de fenómenos de habla espontánea para 6 hablantes especialmente seleccionados de CORLEC-EHU-1. Se muestran asimismo el número de fenómenos total (FHE) y el número de fenómenos por palabra efectiva (FHE/PE) para cada hablante, así como la media y la desviación típica de cada columna, para una mejor evaluación de los datos.	73
3.25. Distribución de fenómenos acústicos dentro de las disfluencias sintácticas en INFOTREN. R: ruidos, S: pausas de silencio, H: pausas habladas, A: alargamientos.	75

3.26. Distribución de fenómenos acústicos dentro de las disfluencias sintácticas en CORLEC-EHU-1. R: ruidos, S: pausas de silencio, H: pausas habladas, A: alargamientos.	76
3.27. Distorsiones léxicas: palabras mal pronunciadas (M) y palabras cortadas (C), dentro de disfluencias sintácticas en INFOTREN.	77
3.28. Distorsiones léxicas: palabras mal pronunciadas (M), palabras cortadas (C) y afirmaciones guturales (G) dentro de disfluencias sintácticas en CORLEC-EHU-1.	77
3.29. Número de marcadores de discurso (MD) que aparecen dentro y fuera de disfluencias sintácticas (DS) en INFOTREN. Se indica también el porcentaje que representan esos números dentro de cada categoría.	78
3.30. Distribución de los marcadores de discurso que aparecen dentro de disfluencias sintácticas en CORLEC-EHU-1.	79
4.1. Características del procesamiento de la señal de voz a 16 y 8 kHz. Se muestran también las características de un procesamiento alternativo a 8 kHz propuesto por el <i>European Telecommunications Standards Institute</i> (ETSI) para entornos ruidosos.	83
4.2. Conjunto de unidades subléxicas: codificación SAMPA, codificación interna del GRFyTH, descripción del sonido y ejemplo.	85
4.3. Bases de datos de habla leída.	95
4.4. Bases de datos de habla espontánea.	95
4.5. Distribución de fonemas en los corpus de entrenamiento.	97
4.6. Distorsión final de los diccionarios CV correspondientes a las bases de datos adquiridas a 8 kHz. Entre paréntesis se indica el número de iteraciones empleadas para construirlos.	98
4.7. Distorsión final de los diccionarios CV correspondientes a las bases de datos adquiridas a 16 kHz. Entre paréntesis se indica el número de iteraciones empleadas para construirlos.	98
4.8. Valor medio, desviación típica, máximo y mínimo del valor absoluto de los elementos fuera de la diagonal en las matrices de correlación obtenidas para los coeficientes cepstrales y sus derivadas primera y segunda.	100
4.9. Coeficiente de correlación entre energía y derivada de la energía.	100
4.10. Resultados de DAF sobre SENGLAR16 y SENGLAR08, utilizando MOM discretos obtenidos mediante los algoritmos de Baum-Welch general, Viterbi e híbrido.	102
4.11. <i>Calidad</i> de las segmentaciones automáticas obtenidas a partir de MOM discretos en SENGLAR16 y SENGLAR08.	103
4.12. Tasas óptimas de DAF obtenidas sobre INFOTREN y las tres particiones de CORLEC-EHU-1, utilizando MOM discretos y aplicando los algoritmos Baum-Welch general, Viterbi e híbrido.	103
4.13. Tasas de DAF óptimas obtenidas sobre SENGLAR16, utilizando MOM semicontinuos y continuos de 8, 16, 32 y 64 gaussianas.	105
4.14. Tasas de DAF óptimas obtenidas sobre SENGLAR08, utilizando MOM semicontinuos y continuos de 8, 16, 32 y 64 gaussianas.	105
4.15. Tasas de DAF óptimas obtenidas sobre INFOTREN y CORLEC-EHU-1, utilizando MOM continuos de 8, 16, 32 y 64 gaussianas, y aplicando el algoritmo de entrenamiento de Viterbi.	106

4.16.	Resumen de las tasas de DAF obtenidas sobre habla leída y habla espontánea, utilizando MOM discretos, semicontinuos y continuos de 8, 16, 32 y 64 gaussianas. En el caso de los MOM discretos, se han escogido las tasas obtenidas de las estimaciones por BWg. En el resto de casos, las tasas corresponden a las estimaciones por Viterbi. En el caso de CORLEC-EHU-1, se muestra la tasa promedio sobre las tres particiones.	107
4.17.	Tasas de DAF sobre BDGEO-16 y BDGEO-8, utilizando MOM discretos y MOM continuos de 32 gaussianas.	107
4.18.	Características de los corpus de texto utilizados para estimar y evaluar los modelos de lenguaje en BDGEO-16, BDGEO-8, INFOTREN y CORLEC-EHU-1.	108
4.19.	Perplejidades de los modelos de lenguaje ($k = 2$, $k = 3$ y $k = 4$), calculadas sobre los corpus de test de BDGEO, INFOTREN y CORLEC-EHU-1.	110
4.20.	Número de transiciones de que constan los modelos de lenguaje definidos para BDGEO, INFOTREN y CORLEC-EHU-1 ($k = 2$, $k = 3$ y $k = 4$).	111
4.21.	Error de reconocimiento (%WER) sobre BDGEO-8, para $k = 2$, $k = 3$ y $k = 4$, utilizando MOM discretos y MOM continuos de 32 gaussianas. Entre paréntesis se muestra el valor óptimo de α	111
4.22.	Error de reconocimiento (%WER) sobre las bases de datos de habla leída y habla espontánea, utilizando MOM discretos y MOM continuos de 32 gaussianas, y aplicando modelos de lenguaje k-ESS con $k = 3$. Entre paréntesis se muestra el valor óptimo de α	111
5.1.	Ampliación del conjunto de unidades subléxicas para las bases de datos de habla espontánea: codificación interna del GRFyTH y descripción del sonido.	116
5.2.	Número de muestras de las unidades acústicas adicionales en los corpus de entrenamiento y test de INFOTREN, y en los tres bloques de CORLEC-EHU-1. Entre paréntesis se indica el porcentaje sobre el total de unidades.	118
5.3.	Número de muestras de la <i>unidad basura</i> en los corpus de entrenamiento y test de INFOTREN, y en los tres bloques de CORLEC-EHU-1. Entre paréntesis se indica el porcentaje sobre el total de unidades.	119
5.4.	Tasas de DAF sobre INFOTREN, utilizando el conjunto de unidades subléxicas ampliado, el conjunto alternativo formado por los fonemas, el silencio y la unidad basura, y, como referencia, el conjunto básico de unidades subléxicas, con MOM discretos y MOM continuos de 8, 16, 32 y 64 gaussianas. Se indica también qué unidades intervienen en los alineamientos.	120
5.5.	Resultados de DAF sobre INFOTREN, utilizando el conjunto ampliado de unidades subléxicas y MOM continuos de 32 gaussianas, desglosados por unidades: número de aciertos (A), sustituciones (S), borrados (B) e inserciones (I), tasa de DAF y las 3 sustituciones más frecuentes —entre paréntesis se indica el porcentaje sobre el total de sustituciones.	122
5.6.	Correlación entre la tasa de DAF y la densidad de fenómenos en cada muestra del corpus de test de INFOTREN, utilizando los conjuntos ampliado, alternativo y básico de unidades con MOM continuos de 32 gaussianas. Se indica también la significancia estadística del resultado, con respecto a la hipótesis nula.	125
5.7.	Tasas de DAF sobre las tres particiones de CORLEC-EHU-1, utilizando el conjunto de unidades subléxicas ampliado y el conjunto alternativo formado por los fonemas, el silencio y la unidad basura, con MOM discretos y MOM continuos de 8, 16, 32 y 64 gaussianas. Se indica también qué unidades intervienen en los alineamientos.	126

5.8. Tasas de DAF promedio para el conjunto ampliado, el conjunto alternativo y el conjunto básico de unidades subléxicas, filtrando todo salvo los fonemas, con MOM discretos y MOM continuos de 8, 16, 32 y 64 gaussianas.	126
5.9. Resultados de DAF sobre el bloque C2 de CORLEC-EHU-1, utilizando el conjunto ampliado de unidades subléxicas y MOM continuos de 32 gaussianas, desglosados por unidades: número de aciertos (A), sustituciones (S), borrados (B) e inserciones (I), tasa de DAF y las 3 sustituciones más frecuentes —entre paréntesis se indica el porcentaje sobre el total de sustituciones.	127
5.10. Correlación entre la tasa de DAF y la densidad de fenómenos en cada muestra de test de CORLEC-EHU-1, utilizando los conjuntos ampliado, alternativo y básico de unidades con MOM continuos de 32 gaussianas. Se indica también la significancia estadística del resultado, con respecto a la hipótesis nula.	130
5.11. Número de bifonemas a izquierda, bifonemas a derecha y trifonemas definidos en las bases de datos de habla leída y habla espontánea, aplicando un umbral de 300 muestras. Entre paréntesis se indica el porcentaje de unidades del corpus de entrenamiento y del corpus de test cubiertas por cada conjunto. En la última columna se indica el número total de unidades contextuales por conteo, así como la cobertura que proporcionan. . .	132
5.12. Tasas de DAF para SENGLAR16, SENGLAR08, INFOTREN y la partición 2 de CORLEC-EHU-1 utilizando unidades contextuales, por separado y conjuntamente, con MOM discretos. En todos los casos el modelo de reconocimiento incluye las unidades incontextuales, para asegurar la cobertura.	133
5.13. Tasas de DAF para SENGLAR16, SENGLAR08, INFOTREN y la partición 2 de CORLEC-EHU-1 utilizando unidades contextuales, por separado y conjuntamente, con MOM continuos de 32 gaussianas. En todos los casos el modelo de reconocimiento incluye las unidades incontextuales, para asegurar la cobertura.	133
5.14. Características de los corpus de texto —fenómenos de habla espontánea incluidos— utilizados para estimar y evaluar los modelos de lenguaje de INFOTREN y CORLEC-EHU-1.	136
5.15. Número de transiciones de los modelos de lenguaje de INFOTREN y CORLEC-EHU-1 ($k = 2$, $k = 3$ y $k = 4$), estimados a partir de transcripciones enriquecidas con pseudo-palabras que representan fenómenos de habla espontánea.	136
5.16. Perplejidades de los modelos de lenguaje de INFOTREN y CORLEC-EHU-1 ($k = 2$, $k = 3$ y $k = 4$), calculadas sobre los correspondientes corpus de test. Los modelos han sido estimados a partir de transcripciones enriquecidas con pseudo-palabras que representan fenómenos de habla espontánea.	136
5.17. Error de reconocimiento (%WER) sobre INFOTREN, utilizando MOM discretos y MOM continuos de 32 gaussianas, y modelos de lenguaje k-ESS con $k = 2$, $k = 3$ y $k = 4$ y los fenómenos de habla espontánea integrados. Entre paréntesis se muestran los valores óptimos de α	137
5.18. Error de reconocimiento (%WER) sobre INFOTREN, utilizando MOM discretos y MOM continuos de 32 gaussianas, y modelos de lenguaje k-ESS con $k = 3$, con los fenómenos integrados e insertados por <i>backoff</i> . Se incluye también, como referencia, el error obtenido sin modelar los FHE. Entre paréntesis se muestran los valores óptimos de α	138

5.19. Error de reconocimiento (%WER) sobre INFOTREN y CORLEC-EHU-1 con MOM de unidades contextuales y modelos de lenguaje k-ESS con $k = 3$ y los fenómenos de habla espontánea integrados. Con objeto de comparar la aportación de las unidades contextuales, se presentan también resultados de reconocimiento sobre BDGEO-8 y BDGEO-16, utilizando modelos de lenguaje k-ESS con $k = 3$, así como los resultados obtenidos con MOM de unidades incontextuales. Entre paréntesis se muestran los valores óptimos de α	140
6.1. Reparto de hablantes al aplicar el algoritmo de <i>clustering</i> sobre SENGLAR16. Se indica el número de hombres y mujeres asignados a cada grupo, así como la sub-base de datos de la que provienen.	157
6.2. Porcentaje de intervenciones asignadas correctamente (%IAC) mediante los modelos de histogramas, para el corpus de entrenamiento de SENGLAR16.	157
6.3. Porcentaje de intervenciones asignadas al grupo correcto, desglosado por grupos, en función de la longitud del segmento de señal sobre el que se aplican los modelos de histogramas, para el corpus de entrenamiento de SENGLAR16.	159
6.4. Reparto de hablantes al aplicar el algoritmo de <i>clustering</i> sobre INFOTREN.	160
6.5. Porcentaje de intervenciones asignadas correctamente (%IAC) mediante los modelos de histogramas, para el corpus de entrenamiento de INFOTREN.	160
6.6. Reparto de hablantes al aplicar el algoritmo de <i>clustering</i> sobre las tres particiones de CORLEC-EHU-1.	163
6.7. Porcentaje de intervenciones asignadas correctamente (%IAC) mediante los modelos de histogramas, para los conjuntos de entrenamiento de las tres particiones de CORLEC-EHU-1.	164
6.8. Porcentaje de intervenciones cuyo género se ha determinado correctamente mediante los modelos de histogramas.	168
6.9. Tasas de DAF obtenidas mediante distintas alternativas de adaptación al hablante. Entre paréntesis se indica el coste computacional, expresado como el número promedio de pasadas de reconocimiento que es necesario realizar para cada señal. También se muestran las tasas de referencia, sin adaptación al hablante.	170
6.10. Porcentaje de intervenciones del corpus de test para las que el grupo que da <i>Max-Prob</i> : (a) coincide con el que da <i>Max-HProb</i> , y (b) está incluido en el subconjunto <i>Haz-HProb</i>	173
6.11. Porcentaje de intervenciones del corpus de test para las que el grupo que da la tasa de DAF más alta: (a) coincide con el grupo que da <i>MaxProb</i> , y (b) coincide con el grupo que da <i>MaxHProb</i>	173
6.12. Tasa máxima de DAF alcanzable con los grupos de hablantes definidos sobre SENGLAR16, INFOTREN y las tres particiones de CORLEC-EHU-1.	174
6.13. Tasas de DAF obtenidas mediante modelos de género: (a) sobre el subcorpus de mujeres con los modelos de mujeres, (b) sobre el subcorpus de hombres con los modelos de hombres, y (c) sobre el corpus de test completo con los modelos (hombres/mujeres) que proporcionen la decodificación más probable. En la última columna se indica el porcentaje de intervenciones cuyo género se ha identificado correctamente en el caso (c).	175

6.14. Resumen de las tasas de DAF obtenidas mediante modelos de grupos de hablantes (Haz-HProb) y de género (Max-Prob). Entre paréntesis se indica el número promedio de pasadas de reconocimiento que es necesario realizar en cada caso. También se muestran las tasas de referencia, sin adaptación al hablante, obtenidas con MOM de unidades incontextuales.	176
6.15. Tasas de DAF obtenidas con modelos de grupos de hablantes y unidades subléxicas contextuales. Se han aplicado dos criterios de selección de grupo: <i>MaxProb</i> y <i>MaxH-Prob</i> . En este último caso, entre paréntesis se indica el porcentaje de coincidencias con el criterio <i>MaxProb</i>	176
6.16. Tasas de DAF obtenidas mediante modelos de género (MOM continuos de 32 gaussianas) y unidades subléxicas contextuales: (a) sobre el subcorpus de test de mujeres con los modelos de mujeres, (b) sobre el subcorpus de test de hombres con los modelos de hombres, y (c) sobre el corpus de test completo con los modelos (hombres/mujeres) que proporcionen la decodificación más probable. Junto a esta última tasa, entre paréntesis, se indica el porcentaje de intervenciones cuyo género se ha identificado correctamente.	177
6.17. Tasas de DAF obtenidas sobre los corpus de test de SENGLAR16, INFOTREN y la partición 2 de CORLEC-EHU-1, utilizando modelos de género y modelos de grupos de hablantes (MOM continuos de 32 gaussianas), con unidades subléxicas incontextuales (ICXT) y contextuales (CXT, bifenemas a derecha). También se muestran las tasas de DAF obtenidas sin adaptación al hablante.	178
6.18. Error de reconocimiento (%WER) sobre el corpus de test de BDGEO-16, utilizando los modelos de género de unidades incontextuales obtenidos sobre SENGLAR16 y el modelo de lenguaje de BDGEO-16 ($k = 3$, $\alpha = 5$ para MOM discretos y $\alpha = 9$ para MOM continuos).	179
6.19. Resultados de reconocimiento (%WER) sobre los corpus de test de BDGEO-16, INFOTREN y la partición 2 de CORLEC-EHU-1, utilizando modelos de género y modelos de grupos de hablantes (MOM continuos de 32 gaussianas), con unidades subléxicas incontextuales (ICXT) y contextuales (CXT). Entre paréntesis se indica el valor óptimo de α en cada caso. También se muestra el error de reconocimiento obtenido sin adaptación al hablante.	179
D.1. Tasas de DAF sobre INFOTREN, utilizando el estándar de parametrización del ETSI para entornos ruidosos, con MOM discretos y continuos del conjunto ampliado de unidades incontextuales. Se muestran únicamente las tasas obtenidas habiendo filtrado todas las unidades salvo los fonemas.	223

Agradecimientos

En primer lugar, quiero mostrar mi agradecimiento a Inés Torres, quien impulsó desde el principio la idea de estudiar y reconocer el habla espontánea. También le debo a ella la oportunidad de formarme como investigador, dentro del grupo que ella misma creó hace ya muchos años, y que tras varios cambios de nombre, ahora se denomina *Grupo de Reconocimiento de Formas y Tecnologías del Habla* (GRFyTH). Sin su trabajo, sus indicaciones y sugerencias esta tesis no habría podido realizarse.

Gracias también a todos los que han trabajado en la creación de los recursos lingüísticos y en la realización de los experimentos. El primer lugar de esa lista lo ocupa, sin duda, Amparo Varona, que ha soportado innumerables discursos y discursitos sobre los más diversos temas, antes, cuando compartíamos despacho, y después, cada uno en el suyo, y sin cuya dedicación no habría podido llevar a cabo los experimentos de reconocimiento. Tras ella, en esa lista, Raquel Justo, quien ha echado una mano muy importante en la anotación de fenómenos sintácticos y pragmáticos. Y, por supuesto, el resto del GRFyTH, en particular Víctor Gujarrubia y Alicia Pérez, que no han dudado en ayudar siempre que se lo he pedido.

Los trabajos de esta tesis se han realizado en el marco de dos proyectos financiados por la CICyT y el MCyT, en colaboración con investigadores de otras universidades españolas, que han aportado también su granito de arena, principalmente con su trabajo, pero también cediendo software, respondiendo a dudas y participando en discusiones más o menos polémicas en reuniones y congresos. Entre ellos quiero destacar —en orden alfabético— a José Miguel Benedí, Antonio Bonafonte, Francisco Casacuberta, Javier Hernando, Eduardo Lleida, José Mariño, Moisés Pastor y Joan Andreu Sánchez.

Vaya también mi agradecimiento a los compañeros del Departamento de Electricidad y Electrónica, colegas, amigos o como deba llamarse a quienes —sí, por circunstancias de la vida— comparten conmigo tantas horas de charla, en cafeterías y pasillos. Gracias a todos —no quiero destacar a ninguno— por la insistencia y por los ánimos, por la ironía y la confianza. El trabajo no habría sido lo mismo sin vosotros.

Y por último, gracias a Cristina, que ha pasado en casa más horas de las debidas, cuidando ella sola de las niñas, ocupándose de la logística y dando el apoyo moral necesario. Como suele decirse, ella ha sido el pan y el agua, el sol y el aire fresco de uno de esos días extraños en que hay menos coches, menos gente y más silencio, y puede uno leer un libro o comer un helado como si fuera la primera vez.

*Words like violence
Break the silence
Come crashing in
Into my little world
Painful to me
Pierce right through me
Can't you understand
Oh my little girl*

*All I ever wanted
All I ever needed
Is here in my arms
Words are very unnecessary
They can only do harm*

*Vows are spoken
To be broken
Feelings are intense
Words are trivial
Pleasures remain
So does the pain
Words are meaningless
And forgettable*

*All I ever wanted
All I ever needed
Is here in my arms
Words are very unnecessary
They can only do harm*

Enjoy the silence

Depeche Mode
Violator (1990)

Capítulo 1

Introducción.

1.1. Contexto y motivación del trabajo.

Esta tesis se enmarca dentro de dos proyectos, uno ya finalizado, denominado BASURDE (Bonafonte et al., 2000), y otro en curso, denominado DIHANA (DIHANA, 2003-2005), que es continuación del primero. Ambos tratan de desarrollar sistemas de diálogo hombre-máquina para acceder a información en un dominio semántico restringido pero permitiendo a los usuarios interactuar espontáneamente. En el primer caso la interfaz está basada en la red telefónica fija y se ubica en un entorno de oficina, mientras que en el segundo caso se consideran dos interfaces adicionales, una de ellas basada en telefonía móvil y otra ubicada en un coche, además de introducir una serie de mejoras producto de la experiencia anterior. El reconocedor es sólo una parte de estos sistemas, pero de su rendimiento depende que las intervenciones de los usuarios sean comprendidas y por tanto, que el sistema reaccione adecuadamente. Para esta tesis se cuenta con una base de datos de diálogos hombre-máquina adquirida en el primero de los proyectos, alrededor de la cual gira una parte importante de los análisis y experimentos realizados.

La lengua hablada espontánea no es un reflejo de la lengua escrita. Los recursos con que cuenta la lengua escrita (signos de puntuación, ortografía, tiempo de elaboración y, por tanto, de corrección del discurso, empleo de las palabras más adecuadas, etc.) hacen que los textos lleguen al lector sin ambigüedades, con la longitud y precisión necesarias. Por otra parte, cuando un locutor lee un texto —dando lugar a lo que se conoce como *habla leída*—, está reproduciendo esa situación, ya que pronuncia con cuidado y a una velocidad uniforme, entonando de acuerdo a los signos de puntuación. En cambio, la lengua hablada espontánea —o simplemente, *habla espontánea*— dispone de otros recursos, relacionados con la inmediatez de la comunicación y con la necesidad de elaborar y corregir el discurso en tiempo real. La calidad del resultado dependerá de la habilidad y del dominio de la lengua del hablante, pero en ese camino las prioridades no serán tanto la precisión y la propiedad como la comunicación en sí misma. Por esta razón la lengua hablada espontánea no es *gramatical* en el sentido en que lo es la lengua escrita. Aparecen en ella *fenómenos* de todo tipo, como pausas, vacilaciones, sonidos (ruidos) extra-lingüísticos, palabras cortadas o pronunciadas de forma incompleta o poco ortodoxa, repeticiones, correcciones, irregularidades sintácticas de toda índole, utilización recurrente de ciertos giros —palabras o grupos de palabras—, etc. Algunos de estos *fenómenos* pueden ser modelados a nivel acústico, o integrados en el modelo de lenguaje del reconocedor, de modo que éste podría asumirlos sin problemas. Otros, por el contrario, son difíciles de detectar —por ejemplo, las palabras cortadas—, o pueden conducir a errores graves de com-

preensión si se integran sin más en el modelo de lenguaje —como es el caso de las correcciones o reformulaciones.

En aplicaciones de diálogo que operan con habla espontánea, como la descrita en el primer párrafo de este apartado, es imprescindible estudiar la fenomenología asociada a esta modalidad del habla y plantear estrategias a distintos niveles: acústico, léxico y sintáctico, con objeto no sólo de corregir los errores de reconocimiento que, sin duda, aparecen cuando se utiliza un reconocedor originalmente diseñado para operar con habla leída, sino también para establecer relaciones entre las palabras reconocidas que reflejen la intención del hablante y permitan comprenderle.

1.2. Objetivos y metodología.

Desde un punto de vista tecnológico, el objetivo principal de este trabajo es mejorar las prestaciones de los sistemas de reconocimiento del habla en lo que se refiere al uso del habla espontánea en las interacciones. Desde un punto de vista más teórico, se plantea otro objetivo, que es caracterizar el habla espontánea frente al habla leída, y distinguir, si fuera posible, entre distintos tipos de habla espontánea.

Para alcanzar estos objetivos, en primer lugar es necesario crear recursos lingüísticos que permitan estudiar y modelar el habla espontánea en castellano —recursos que no existían en el momento de comenzar los trabajos de esta tesis. La adquisición de una base de datos de diálogos hombre-máquina en el proyecto BASURDE cumple en parte con ese requisito. Sin embargo, en las anotaciones de esta base de datos no quedan reflejados adecuadamente los fenómenos de habla espontánea, por lo que se hace necesario definir un esquema de anotación y llevar a cabo las anotaciones de acuerdo a dicho esquema.

Lógicamente, antes de definir o elegir un esquema de anotación, es necesario recabar información sobre los esquemas existentes, y revisar trabajos previos relacionados con la definición y categorización de disfluencias y otros fenómenos de habla espontánea. De este modo, con un conocimiento suficiente del tema, será posible diseñar un esquema de anotación propio, o adaptar un esquema existente, que permita abarcar todos los fenómenos de una forma coherente, y que pueda crecer en la medida que se planteen nuevas necesidades. Esta revisión y puesta al día constituye en sí misma un objetivo del trabajo, ya que partimos de un desconocimiento absoluto del tema.

Durante el estudio del habla espontánea en diálogos hombre-máquina, surge la cuestión de si en conversaciones naturales entre personas se producen los mismos fenómenos, y en qué medida. Así pues, se plantea también como objetivo comparar el habla espontánea que se produce en una tarea específica, como es el acceso a información a través de un terminal automatizado, con el habla espontánea que se produce en circunstancias normales entre personas. Una parte de ese estudio debe consistir en analizar el tipo de fenómenos que aparecen en ambas situaciones, así como sus frecuencias, relacionándolas, por ejemplo, con la duración de las intervenciones y con la identidad del hablante. Otra parte, no menos importante, debe consistir en comprobar cómo se comporta el reconocedor en uno y otro caso. Para todo ello es necesario disponer de una base de datos de diálogos naturales entre personas, y aplicar sobre ella los mismos procedimientos de análisis y modelización desarrollados para la base de datos de diálogos hombre-máquina. La tarea de creación de esta base de datos —mediante el reciclaje, transformación y ampliación de una base de datos existente—, aunque poco *agradecida* desde el punto de vista de la investigación, ha llevado una parte muy importante del tiempo de ejecución de esta tesis.

En cuarto lugar, se plantea comparar el rendimiento del reconocedor sobre habla espontánea con su rendimiento sobre habla leída, y estudiar las posibles fuentes de error que se añaden cuando se trata con habla espontánea. De ahí que, además de las dos bases de datos de habla espontánea, se hayan considerado otras dos bases de datos de habla leída, adquiridas a 16 kHz en entorno de laboratorio y a 8 kHz simulando una línea telefónica, respectivamente.

Finalmente, el tratamiento específico del habla espontánea admite muchas perspectivas. Se puede plantear desde el punto de vista de la prosodia, camino prácticamente inexplorado en lo que respecta al habla leída; también desde el punto de vista de la modelización acústica de los sonidos de la lengua, o desde el punto de vista de la modelización léxica, ya que la variabilidad de sonidos y pronunciaciones es muchísimo más alta en habla espontánea que en habla leída; desde el punto de vista de la modelización sintáctica, puesto que las construcciones de la lengua hablada responden a una lógica interna que difiere de la gramática de la lengua escrita; o incluso desde una perspectiva semántico-pragmática. Con respecto a este último plano de análisis, nótese que, a diferencia de la lengua escrita, en la que la corrección gramatical, la propiedad y la ausencia de ambigüedad son los principios básicos, en la lengua hablada es esencial el aspecto comunicativo-interactivo, es decir, hacerse entender. En esta tarea el hablante aplica el principio básico del mínimo esfuerzo, aportando la cantidad de información mínima —en todos los sentidos: acústico, léxico, sintáctico y semántico— para que su audiencia, con la competencia lingüística que se le atribuye, comprenda el mensaje. Aunque algunos autores han tratado de hacer un estudio o incluso una modelización conjunta de varios de estos planos, resulta difícil atacar todos ellos en una primera aproximación como ésta, tentativa y pionera en castellano. Por ello, se ha restringido el estudio y la modelización al plano acústico.

En la práctica, las condiciones ambientales y de canal constituyen una de las principales fuentes de variabilidad acústica en el habla espontánea, y producen una degradación considerable del rendimiento de los sistemas de reconocimiento. Esto requeriría aplicar estrategias de compensación o eliminación del ruido, y quizá extraer parámetros acústicos más robustos. Sin embargo, el ruido como tal no es una circunstancia que dependa de la modalidad del habla —el habla espontánea podría no contener ruido—, sino del equipo y el entorno de grabación, por lo que su estudio y modelización no serán acometidos en este trabajo.

Así pues, el esfuerzo se ha concentrado en mejorar los modelos acústicos en aquellos aspectos que se han considerado específicos del habla espontánea. Se han seguido dos líneas de trabajo. Por un lado, se ha ampliado el conjunto de unidades subléxicas con fenómenos acústicos que no aparecen en habla leída: sonidos fáticos de duda o de mantenimiento del turno, alargamientos de sonidos, asentimientos guturales, ruidos producidos por el locutor y ruidos externos. La frecuencia de estos fenómenos en habla espontánea —no sólo en conversaciones entre personas, sino también en diálogos hombre-máquina— es lo bastante alta como para que su inclusión en el inventario de modelos acústicos sea fundamental, pues evitará que sean confundidos con palabras y facilitará, por tanto, el reconocimiento de éstas. Por otra parte, se ha modelado la variabilidad acústica asociada al hablante. Puesto que no es lo mismo *leer* que *hablar*, la variabilidad asociada al hablante está muy limitada en las bases de datos de habla leída: los locutores que participan en grabaciones de habla leída pronuncian con cuidado, a una velocidad constante, inhibiendo aspectos personales en la realización de los sonidos y las palabras. Es en el habla espontánea donde las características propias de los hablantes se manifiestan plenamente. Esto ha de tenerse en cuenta tanto en la estimación de los modelos acústicos como en el uso de éstos durante el reconocimiento, mediante estrategias de normalización del hablante o de adaptación al hablante. En este trabajo se ha optado por una estrategia de adaptación basa-

da en la definición automática de grupos de hablantes y en la selección del grupo más adecuado durante el reconocimiento.

Tanto en la definición de unidades subléxicas específicas para los fenómenos de habla espontánea de tipo acústico como en la estrategia de adaptación al hablante, se ha seguido una misma metodología: estimar los modelos acústicos correspondientes y comprobar su rendimiento en *decodificación acústico-fonética* (DAF) y en reconocimiento. La experimentación ha sido mucho más exhaustiva en DAF, debido al elevado coste de los experimentos de reconocimiento, que se han planteado como un *banco de pruebas* para los modelos acústicos. De ahí que no se hayan realizado los ajustes necesarios para optimizar el rendimiento del reconocedor —ajustes que habrían requerido un barrido experimental exhaustivo y, por tanto, muy costoso. Por el contrario, se ha elegido y mantenido una misma configuración básica del sistema que ha permitido medir y comparar el rendimiento de las distintas alternativas de modelización. Siempre que ha sido posible, se han realizado dos tipos de análisis: habla espontánea frente a habla leída, y habla espontánea en diálogos hombre-máquina frente a habla espontánea en conversaciones entre personas.

1.3. Estructura de la memoria.

El resto de la memoria consta de 6 capítulos y 4 apéndices. Los capítulos del 2 al 6 describen, paso a paso, el trabajo realizado, mientras que el capítulo 7 enumera las aportaciones más destacables de esta tesis, las publicaciones relacionadas con cada uno de los capítulos, y los trabajos planificados para el futuro más inmediato. La intención del autor es que cada capítulo, que comienza con una breve descripción de su contenido, pueda leerse, en la medida de lo posible, independientemente del resto. Como consecuencia, ciertas ideas aparecen repetidas en varios de ellos. Aún así, es importante mantener la secuencia de lectura, ya que muchos términos y siglas se definen una sola vez, a medida que aparecen; además, en cada capítulo hay numerosas referencias a capítulos previos. Por otra parte, en lugar de escribir un capítulo específico en el que revisar y discutir los trabajos previos más destacables, dichos contenidos se han ido ubicando en los capítulos correspondientes (2 y 6, fundamentalmente). Por último, los apéndices, aunque no son imprescindibles, ayudan a comprender y valorar el trabajo realizado. A continuación se resumen los contenidos de los 5 capítulos centrales:

- ▷ El capítulo 2 presenta un esquema de anotación de fenómenos de habla espontánea en diálogos hombre-máquina y entre personas. En primer lugar se define el concepto de *fenómeno de habla espontánea*, frente al concepto más restringido de *disfluencia* que suele aparecer en la literatura. A continuación se repasan los trabajos de clasificación y anotación de este tipo de fenómenos realizados previamente en lengua inglesa. En tercer lugar se describen los fenómenos que forman parte del esquema de anotación, a nivel acústico, léxico, sintáctico y pragmático, y se discute el tratamiento que han recibido en la literatura, así como el tratamiento que recibirán en este trabajo. En cuarto lugar se enumeran las características que debe tener un formato de anotación lingüística y se repasan el estado del arte y los estándares en lo que se refiere a anotación de habla espontánea en diálogos. En quinto lugar, partiendo del inventario de fenómenos descrito previamente, se define el formato de anotación como una aplicación XML. Por último, tras una serie de consideraciones sobre la herramienta de anotación y sobre las distintas alternativas existentes, se describe el entorno de anotación utilizado en este trabajo, basado en herramientas de

libre distribución, en un formato simplificado y en una serie de recomendaciones sobre el cortado y anotación de las señales. Los apéndices A y B contienen, respectivamente, la especificación formal de la aplicación XML, y dos ejemplos de anotación de habla espontánea —en diálogos hombre-máquina y en conversaciones entre personas—, en formatos XML y simplificado.

- ▷ En el capítulo 3 se describen en primer lugar dos bases de datos de habla espontánea, una de ellas —la adquirida dentro del proyecto BASURDE— ligada a una tarea específica, y por tanto con un vocabulario y una gramática bastante restringidos, y la otra completamente genérica, no ligada a ningún dominio semántico en particular, por lo que tanto el vocabulario como el modelo de lenguaje son generales. A continuación se realiza un estudio pormenorizado de las estadísticas de aparición de los fenómenos de habla espontánea anotados en cada una de ellas, y se estudia en qué medida la cantidad de fenómenos depende de la duración de las intervenciones o de la identidad del hablante. Finalmente, se explora la posibilidad de determinar la presencia de reformulaciones en función de la presencia de fenómenos acústicos, léxicos y pragmáticos. Todos los estudios realizados incluyen un análisis comparativo entre habla espontánea en diálogos hombre-máquina y habla espontánea en conversaciones naturales entre personas.
- ▷ El capítulo 4 describe las características del sistema de reconocimiento del *Grupo de Reconocimiento de Formas y Tecnologías del Habla* (GRFyTH) de la UPV/EHU, tal como se aplica sobre habla leída. En esta descripción se incluyen los parámetros acústicos, la cuantificación vectorial, el conjunto de unidades subléxicas, la forma en que se han obtenido las transcripciones fonéticas, el modelo de lenguaje y, muy especialmente, los modelos acústicos —Modelos Ocultos de Markov (MOM) discretos, continuos y semicontinuos. No olvidemos que uno de los principales objetivos de esta tesis es precisamente evaluar el rendimiento de los modelos acústicos sobre habla espontánea, y mejorarlo en la medida de lo posible. Finalmente, se muestran los resultados de DAF y de reconocimiento sobre habla espontánea utilizando dicha configuración. También se presentan resultados sobre dos bases de datos de habla leída, con objeto de contrastarlos con los de habla espontánea y establecer posibles fuentes de error. En el apéndice C se presentan las fórmulas de reestimación utilizadas en este trabajo, correspondientes a los algoritmos clásicos, para MOM discretos, continuos y semicontinuos.
- ▷ En el capítulo 5 se propone una ampliación del conjunto de unidades subléxicas que cubre todos los fenómenos acústicos que aparecen en habla espontánea, y se presentan las estadísticas de aparición de estos fenómenos en las dos bases de datos de habla espontánea consideradas. También se presenta una aproximación intermedia, consistente en definir una única *unidad basura* para modelar todos los fenómenos de habla espontánea. Con objeto de detectar las fuentes error en DAF, se analizan con cierto detalle las matrices de confusión obtenidas en un par de experimentos. Asimismo, se muestran ejemplos concretos —segmentaciones obtenidas en DAF y secuencias de palabras obtenidas en reconocimiento— que revelan la precisión con que algunos de estos fenómenos son reconocidos, y, en contraste, las confusiones que se producen en su entorno cuando no se dispone de los modelos adecuados. Finalmente, se presentan varios conjuntos de unidades subléxicas contextuales definidos a partir del conjunto ampliado —sólo en los casos en que se dispone de un número suficiente de muestras—, así como los resultados de DAF y de reconocimiento obtenidos con ellos.

- ▷ Por último, en el capítulo 6, tras una revisión de las técnicas de adaptación al hablante y de normalización del hablante, se describe una técnica de *clustering* del conjunto de hablantes, de modo que para cada uno de los subconjuntos resultantes se estiman modelos específicos. Con ello se pretende eliminar la variabilidad debida al hablante, con objeto de que los modelos acústicos sólo den cuenta de las diferencias entre los distintos sonidos de la lengua. El algoritmo propuesto está basado en el clásico algoritmo de cuantificación vectorial LBG (Linde et-al., 1980). Como alternativa a esta metodología se explora también el uso de modelos de género. La utilidad práctica de estas aproximaciones depende de la existencia de un método rápido y eficaz para determinar el mejor conjunto de modelos para cada señal de entrada. El propio modelo utilizado para definir los grupos de hablantes proporciona una solución simple y elegante a ese problema, ya se trate de grupos obtenidos automáticamente o de grupos definidos *a priori*, si bien se proponen, además, algunas otras alternativas de selección/combinación de modelos. Se presentan los resultados del *clustering* así como los índices de acierto de la función de selección entre los hablantes del conjunto de entrenamiento, y algunos ejemplos que muestran la coherencia de esa misma función de selección entre los hablantes del conjunto de test. Finalmente, se presentan resultados de DAF y de reconocimiento, con modelos de grupos de hablantes y con modelos de género, sobre una base de datos de habla leída y sobre las dos bases de datos de habla espontánea consideradas en este trabajo.

Capítulo 2

Un esquema de anotación de fenómenos de habla espontánea.

En los últimos años se ha disparado el número de aplicaciones de la ingeniería lingüística, desde las conocidas máquinas de dictado hasta los sistemas de traducción simultánea, pasando por sistemas automáticos de atención al cliente basados en interfaces de habla natural, sistemas de indexado y resumen automático de noticias de radio y televisión, etc. La mayor parte de estos sistemas utilizan modelos probabilísticos entrenados a partir de grandes bases de datos de habla y texto, que incorporan varios niveles de anotación. El coste de creación y desarrollo de estas bases de datos es enorme, sobre todo en términos de esfuerzo humano, y requiere el diseño de herramientas y formatos adecuados, que aceleren el proceso de anotación y faciliten el intercambio de información.

El estudio y la modelización de los fenómenos específicos del habla espontánea requiere también en primer lugar la anotación de los mismos. Incluso en el contexto de una tarea de diálogo hablado hombre-máquina relativamente sencilla, con un vocabulario medio y una tipología de frases bastante cerrada, aparece un gran número de fenómenos que entorpecen seriamente el rendimiento del reconocedor. En realidad, se superponen fenómenos a distintos niveles: acústico, fonológico, prosódico, léxico, morfosintáctico, semántico, pragmático, etc. En la construcción de un reconocedor que trate con este tipo de habla se deben tener en cuenta al menos ciertos fenómenos, en concreto los que suceden a nivel acústico, léxico y sintáctico, con objeto de obtener modelos suficientemente precisos y robustos. Otros módulos del sistema de diálogo, como los de comprensión del habla, gestión del diálogo y generación de respuesta, podrían beneficiarse también de la información que aportan estos y otros fenómenos —a nivel prosódico y pragmático, principalmente—, para resolver ambigüedades, interpretar correctamente las intervenciones del usuario y actuar en consecuencia.

Por otra parte, fuera del contexto de una aplicación concreta, en diálogos naturales entre personas, es de esperar que estos fenómenos tengan una presencia aún más notable, hasta el punto de que podamos considerarlos como recursos del lenguaje hablado. Estudiar la frecuencia y la tipología de tales fenómenos en el habla espontánea genérica tiene interés sobre todo desde el punto de vista lingüístico pero también con una perspectiva práctica a más largo plazo, ya que las restricciones impuestas actualmente a los reconocedores se irán relajando a medida que el rendimiento de éstos mejore.

En cualquier caso, definir un esquema de anotación se plantea como una condición necesaria para el estudio empírico y la modelización de este tipo de fenómenos. En los siguientes

apartados se describen los elementos de que constará dicho esquema: los niveles de anotación, el inventario y la jerarquía de fenómenos, el formato y la herramienta de anotación.

2.1. Los niveles de anotación y el inventario de fenómenos.

En nuestra aproximación, los niveles de anotación se corresponden con los niveles de análisis del lenguaje. Dentro de cada nivel, dependiendo del grado de detalle deseado, será necesario distinguir varias categorías y subcategorías de fenómenos. Dos objetivos concretos guiarán la definición del inventario de fenómenos. Por un lado, obtener transcripciones tan fieles a la señal como sea posible, y por otro, proporcionar tanta información como sea posible sobre las llamadas auto-correcciones o reformulaciones, ya que una adecuada modelización de las mismas será fundamental para la comprensión e interpretación del habla espontánea. Asimismo, el inventario de marcas deberá ser completo y preciso, y a ser posible deberá evitar las redundancias, lo cual facilitará la tarea de anotación, ya que todo fenómeno podrá ser clasificado y siempre lo será de la misma forma.

El corpus de entrenamiento de los modelos acústicos y de lenguaje de un reconocedor de habla espontánea deberá contener al menos los parámetros acústicos de las señales, así como sus transcripciones ortográficas y fonéticas. En este trabajo ambos tipos de transcripciones se han enriquecido con fenómenos lingüísticos y extra-lingüísticos, organizados en cuatro niveles de anotación: acústico, léxico, sintáctico y pragmático. El esquema de anotación resultante considera únicamente los fenómenos observados en una aplicación de diálogo hombre-máquina y en conversaciones naturales entre personas. Algunas características se han añadido pensando en aplicaciones futuras. Tal es el caso de las marcas de sincronización, que podrían servir para definir segmentaciones, aunque, como veremos en el capítulo 3, las anotaciones iniciales no incluyen la definición de segmentos. Por otra parte, no todas las marcas tienen sentido en todos los casos. Por ejemplo, en interacciones hombre-máquina apenas aparecen solapamientos. Además, en la base de datos de diálogos hombre-máquina utilizada en este trabajo, las dos señales ocupan canales distintos, por lo que sería posible manejarlas y aplicarlas en modelización acústica aunque existiesen solapamientos.

Antes de describir en detalle cada uno de los niveles, definiremos algunos conceptos importantes y revisaremos los trabajos más representativos realizados hasta la fecha —todos en habla inglesa— en lo que se refiere a anotación de fenómenos de habla espontánea.

2.1.1. Disfluencias y fenómenos de habla espontánea.

En la literatura no se suele hablar de fenómenos de habla espontánea, sino de *disfluencias*. Sin embargo no está claro qué fenómenos pueden considerarse dentro de esta categoría. Dependiendo de los autores el término *disfluencia* se refiere a todos o sólo a algunos de los siguientes fenómenos: sonidos fáticos de duda o de mantenimiento del turno, palabras cortadas, repeticiones y reformulaciones.

El término *disfluencia* se aplica, en general, a cualquier ruptura de la continuidad acústica y/o sintáctica del habla, ruptura que se produce debido a problemas de planificación del discurso, acceso léxico deficiente, etc. (Levelt, 1989, capítulo 12). El estudio de las causas que provocan tales rupturas del discurso, así como su procesamiento interno a nivel cognitivo, son cuestiones clave desde el punto de vista de la psicolingüística. La manifestación externa de

las disfluencias, las complejas estructuras sintácticas que resultan, así como los modelos de interpretación semántica, son cuestiones que pertenecen al área del procesamiento del lenguaje natural.

Sin embargo, nuestro objetivo principal es incrementar la robustez de los sistemas de reconocimiento para tratar el habla espontánea. Desde este punto de vista, fenómenos acústicos como ruidos, alargamientos de sonidos y pausas, o pragmáticos como los marcadores de discurso, son tan importantes como aquellos otros considerados propiamente como disfluencias. Por ello hablaremos de *fenómenos de habla espontánea*, un concepto más amplio que el de disfluencia, ya que incluye cualquier característica de tipo acústico, léxico, sintáctico o incluso pragmático, que distinga el habla espontánea del habla leída. Todos los fenómenos de habla espontánea deben ser identificados, anotados y modelados para mejorar las prestaciones del reconocedor.

2.1.2. Anotación de fenómenos de habla espontánea: perspectiva histórica.

El esquema de anotación definido para este trabajo no surge de la nada, sino que resulta de la combinación, adaptación y/o simplificación de esquemas definidos anteriormente para bases de datos de habla espontánea en inglés (Shriberg, 1994; Heeman, 1997; Lickley, 1998), trabajos que repasamos a continuación en orden cronológico.

2.1.2.1. ATIS y Switchboard.

El desarrollo de sistemas de reconocimiento y comprensión del habla espontánea a gran escala puede decirse que comenzó hace poco más de diez años, con la grabación y anotación de dos grandes bases de datos en lengua inglesa: **ATIS** (MADCOW, 1992) —*Air Travel Information System*—, una tarea de diálogo hombre-máquina en la que un usuario trata de obtener información sobre horarios, precios y otras condiciones de vuelos entre ciudades de Estados Unidos, y **Switchboard** (Godfrey et al., 1992), una colección de conversaciones telefónicas entre dos personas sobre diversos temas. Puesto que el principal objetivo de estas bases de datos era obtener modelos acústicos más robustos para los sistemas de reconocimiento, las anotaciones incluyeron prioritariamente fenómenos acústico-prosódicos: ruidos, pausas, sonidos fáticos de duda o de mantenimiento del turno (que en inglés reciben el nombre de *filled pauses*), marcas de entonación, etc. Sin embargo, también se incluyeron ciertos fenómenos de tipo léxico: palabras cortadas y mal pronunciadas, y fenómenos de tipo sintáctico, en concreto palabras que han sido pronunciadas pero que se consideran borradas implícitamente.

Posteriormente, las transcripciones de ATIS y Switchboard fueron enriquecidas con más fenómenos, sobre todo de tipo sintáctico. En esta tarea fueron importantísimos los trabajos de Elisabeth Shriberg (Shriberg, 1994), cuyo esquema de anotación fue aplicado en ambos casos, con algunas diferencias entre uno y otro. El esquema que plantea Shriberg se refiere únicamente a cierto tipo de disfluencias, en concreto repeticiones, reformulaciones y reinicios de frase, todos ellos tratados bajo el mismo formalismo, que es independiente de las diversas teorías que explican tales fenómenos y no propone una interpretación semántica, ya que se basa únicamente en la manifestación externa de los mismos. Este esquema lo explicaremos con detalle en el apartado 2.1.5.

En el caso de ATIS, repeticiones, reformulaciones y reinicios de frase se anotaron con objeto

de entrenar y evaluar algoritmos que las detectaran y corrigieran automáticamente mediante técnicas de identificación de patrones, por lo que a cada palabra sustituida, borrada o añadida se le asignó un identificador y una marca que la relacionaba con otras palabras (Bear et al., 1993).

En cambio, la anotación de esos mismos fenómenos en Switchboard siguió estrictamente el esquema de Shriberg, sin identificar individualmente las palabras implicadas, ni relacionándolas entre sí (Meteer et al., 1995). El esquema de anotación de Switchboard incluye marcas para sonidos fáticos de duda o de mantenimiento del turno, *señales de edición*, marcadores de discurso, conjunciones coordinantes y aposiciones, y define un tipo especial de unidades, llamadas en inglés *slash units*, para aquellas partes de una intervención que tienen un significado completo en sí mismas: frases subordinadas, pequeñas frases cortas, expresiones, etc. y que irían separadas unas de otras mediante marcas de entonación.

2.1.2.2. TRAINS y Map Task.

Alrededor de 1995 aparecieron dos nuevas bases de datos de habla espontánea: el corpus **TRAINS** (Heeman and Allen, 1995b), desarrollado en la Universidad de Rochester, y el corpus **Map Task** (Anderson et al., 1991), desarrollado en la Universidad de Edimburgo. Ambas bases de datos consisten en diálogos entre personas orientados a resolver una tarea de forma conjunta y cooperativa. En el caso de TRAINS el objetivo era transportar mercancías a través de una red de ferrocarriles, y en el caso de Map Task se trataba de que un operador ayudase a un usuario a realizar un recorrido a través de un espacio geográfico ficticio utilizando mapas parcialmente distintos. Como resultado, la espontaneidad de TRAINS y Map Task es superior a la de ATIS, pero inferior a la de Switchboard.

TRAINS. Las transcripciones iniciales de TRAINS contenían una descripción bastante fiel de las señales, con marcas de fenómenos acústicos y léxicos: ruidos, respiraciones, silencios, palabras cortadas o mal pronunciadas, números y horas tal cual habían sido pronunciados, sonidos fáticos de duda o de mantenimiento del turno, contracciones, etc. (Heeman and Allen, 1995a). Posteriormente, se añadieron marcas de bordes de entonación, auto-correcciones y categorías sintácticas (en inglés, *Part-Of-Speech tagging*).

El esquema de anotación de auto-correcciones, basado en el de ATIS, distingue entre varios tipos: reinicios de frase, reformulaciones (donde se incluyen las repeticiones) y reformulaciones reducidas (*abridged repairs* en inglés). Estas últimas corresponden en realidad a sonidos fáticos de duda o de mantenimiento del turno, que a juicio de los autores revelan la presencia de una reformulación que el hablante ha logrado reprimir, es decir, la reformulación habría sucedido en su cabeza pero no habría sido pronunciada. Esta interpretación se basa en una teoría sobre el procesamiento interno de las disfluencias, y como tal no estaba incluida en el esquema de Shriberg. De ahí que ni ATIS ni Switchboard incluyeran este tipo de reformulaciones. Tampoco incluye el esquema de Shriberg distinción alguna entre reinicios de frases y reformulaciones.

Map Task. Las convenciones ortográficas del corpus Map Task incluyen marcas para fenómenos acústicos extralingüísticos como pausas, sonidos fáticos de duda o de mantenimiento del turno, alargamientos de sonidos, gruñidos, etc. y también para ciertas *anomalías* léxicas, como palabras cortadas o mal pronunciadas, palabras deletreadas, etc. Cada conversación del corpus incluye también marcas TEI/SGML que permiten: (1) identificar conversación, hablante y turno; (2) sincronizar la transcripción ortográfica con la señal; (3) delimitar segmentos donde

las señales de dos hablantes se solapan; y (4) marcar ciertos eventos especiales, como palabras extranjeras, ruidos y segmentos difíciles de transcribir.

Las anotación de reformulaciones, realizada en una segunda fase a partir de las transcripciones ya sincronizadas (Lickley, 1998), aplica el esquema de Shriberg, distinguiendo entre cuatro tipos: repeticiones, sustituciones, inserciones y borrados, éstos últimos equivalentes a los reinicios de frases. Cada palabra que participa en una reformulación lleva una marca que indica si es repetida, sustituida, insertada o borrada, pero no se establecen relaciones entre palabras. Por otro lado, cada reformulación es anotada en una capa lógica distinta, lo cual permite definir anidamientos estrictos pero también solapamientos parciales entre reformulaciones.

2.1.2.3. El proyecto MATE.

Recientemente, en el marco de un proyecto europeo denominado **MATE** (*Multilevel Annotation Tools Engineering*), se ha desarrollado un meta-esquema de anotación multi-nivel para la creación de bases de datos de diálogos hablados, así como una interfaz gráfica y diversas herramientas de anotación y procesamiento (McKelvie et~al., 2001). Este meta-esquema va acompañado de propuestas de anotación en cada nivel, basadas normalmente en uno o varios esquemas existentes. En concreto, dentro del nivel morfosintáctico, se incluye un módulo de anotación de disfluencias, llamado *Edited Transcription Coding Module* —véase (Mengel et~al., 1999), páginas 48 a 64—, que incluye marcas no sólo para repeticiones, reinicios de frases y reformulaciones, sino también para formas irregulares, palabras omitidas, frases interrumpidas y frases que completan otras frases.

La novedad de este esquema con respecto a los revisados anteriormente es que distingue varios niveles de detalle en las anotaciones. Así, hay un *nivel obligatorio* que requiere marcar los segmentos de la transcripción que intervienen en las disfluencias, El *nivel recomendado* permite establecer la relación entre los segmentos definidos en el nivel obligatorio cuando éstos intervienen en una reformulación. Así pueden definirse repeticiones, sustituciones, reformulaciones con inserción de elementos y reformulaciones con borrado de elementos. Finalmente el *nivel opcional* permite identificar, dentro de una reformulación, los tres elementos que la forman (según la definición de Shriberg): el segmento a corregir (o repetir), las *señales de edición* (en caso de que existan) y el segmento que actúa como *corrección*.

2.1.3. Nivel acústico.

En este nivel se incluyen *aquellos fenómenos que afectan a los modelos acústicos del reconocedor*, como ruidos, pausas, sonidos fáticos de duda o de mantenimiento del turno, alargamientos de sonidos, etc. Para integrar estos fenómenos en el reconocedor será necesario definir nuevos modelos acústicos y/o refinar los existentes, tal como se describe en (Rose and Riccardi, 1999).

2.1.3.1. Ruidos.

Teniendo en cuenta la fuente, distinguimos dos tipos de ruidos: aquéllos producidos por el propio hablante, y aquéllos que provienen del exterior o que son debidos al canal de transmisión (micrófono, línea telefónica, etc.). Así, por ejemplo, si el hablante se coloca muy cerca del micrófono, incrementará drásticamente el número de saturaciones, produciendo una distorsión muy notable. Este tipo de ruidos o distorsiones se consideran producidos por el canal.

Ruidos externos o de canal. Suceden de forma aleatoria, no tienen ninguna relación con la señal de voz, simplemente se superponen y *molestan*. Por otra parte, ni su frecuencia es tan elevada ni su tipología tan definida como para distinguir subcategorías. Así pues, independientemente de la fuente, el *color* y la duración, consideraremos un solo tipo de ruidos externos. Por otra parte, aunque podemos encontrar ruidos externos aislados que no llegan a superponerse con la señal de voz, lo normal es que se prolonguen durante un tiempo y afecten a varias palabras e incluso a frases o turnos completos. En estos casos se marcará como ruidoso el segmento completo, que podría ser descartado para entrenar modelos acústicos.

Ruidos producidos por el hablante. Con este tipo de ruidos sucede justo lo contrario: son muy abundantes y responden a una tipología muy precisa, lo cual permite modelarlos con suficiente precisión. En concreto, distinguiremos tres tipos de ruidos producidos por el hablante, cada uno de ellos con características espectrales y de duración muy definidas: *aspiración/inspiración*, *chasquido de labios* y *tos/carraspeo*. Los ruidos producidos por el hablante no suceden por casualidad, sino que responden —salvo en el caso de las toses— a un patrón más o menos cíclico de *descansar para tomar aire y retomar el discurso*. Conviene distinguir estas *pausas técnicas* de las pausas de silencio, ya que, además de mostrar características espectrales y de duración distintas, su aparición responde también a causas distintas.

Por último, mientras que los ruidos externos pueden superponerse con segmentos de habla, tal superposición es físicamente imposible en el caso de ruidos producidos por el hablante. En todo caso, el hablante podría interrumpirse o cortar una palabra para tomar aire o toser, en cuyo caso podría producirse una reformulación, pero nunca una superposición de señales.

2.1.3.2. Pausas de silencio.

Las pausas de silencio desempeñan un papel clave en la producción y comprensión del habla espontánea (Levelt, 1989, páginas 32-37 y 385-387). En habla leída (y en general, en habla fluida, sea leída o espontánea) actúan como elementos *expresivos*, no como puntos de ruptura. Suelen marcar la separación entre dos frases o entre dos unidades semánticas independientes. Esta circunstancia hace de las pausas de silencio pistas especialmente útiles para segmentar automáticamente la señal en unidades más pequeñas, lingüísticamente coherentes, lo cual podría simplificar y acelerar el proceso de reconocimiento.

En habla espontánea, las pausas de silencio son mucho más frecuentes que en habla leída. A veces cumplen una función meramente expresiva, pero otras muchas ponen de manifiesto un momento de duda, un lapso que el hablante necesita para planificar el discurso. En estos casos, las pausas de silencio suelen ir acompañadas de sonidos fáticos (vocalizaciones o nasalizaciones) con las que el hablante marca su intención de continuar el discurso. En ocasiones, alrededor una pausa de silencio tiene lugar una repetición o una reformulación, de manera que la presencia de una pausa de silencio (de 200 milisegundos o más) podría utilizarse como pista para detectar reformulaciones.

Las pausas de silencio pueden desempeñar funciones comunicativas muy complejas. El hablante podría utilizarlas, por ejemplo, para explorar el nivel de comprensión de la audiencia, dando la oportunidad a un posible interlocutor de interrumpir su discurso e intervenir. Si la pausa se prolonga y nadie se decide a interrumpir, el hablante considera que puede continuar su discurso. Este tipo de interpretaciones podrían ser tenidas en cuenta por un modelo de comprensión o por un modelo de diálogo para mejorar la interacción con los usuarios.

2.1.3.3. Pausas *habladas*.

Tras una difícil exploración de alternativas, *pausa hablada* —traducción libre de *filled pause*— se nos antojó la forma más expresiva y compacta de definir el fenómeno en castellano (frente a su traducción literal: *pausa rellena*), puesto que deja claro que se trata de un fenómeno producido por el hablante y que su contenido es habla.

En apartados anteriores nos hemos referido en varias ocasiones a estos fenómenos como *sonidos fáticos de duda o de mantenimiento del turno*. Se trata de fenómenos extralingüísticos, realizados acústicamente como vocalizaciones o nasalizaciones, que aparecen por distintas razones, muy habitualmente como recursos para mantener el turno: el hablante no ha planificado aún lo que va a decir, o tiene problemas para encontrar las palabras adecuadas, y se toma tiempo para pensarlo, pero hace notar su intención de seguir hablando, ya que si simplemente callara, sus interlocutores podrían interpretar que ha terminado su turno y tomarían la iniciativa.

Las pausas habladas constituyen el fenómeno extralingüístico más frecuente en habla espontánea y suelen ir asociadas a otros fenómenos más complejos, como repeticiones o reformulaciones. En tales casos, el hablante detecta un error de adecuación o de articulación y marca el punto de retroceso y corrección con una pausa hablada. También, al igual que las pausas de silencio, dependiendo de la entonación y del contexto semántico, una pausa hablada puede ser la señal que indica a un interlocutor que puede intervenir.

La importancia de las pausas habladas en la producción y comprensión del habla espontánea ha sido puesta de manifiesto desde diversas perspectivas en numerosos trabajos (véanse, por ejemplo, (Shriberg and Stolcke, 1996; Shriberg, 1999)). Las pausas habladas aparecen con mayor probabilidad en los bordes de unidades lingüísticas (Shriberg, 1994), de modo que podrían ser utilizadas para segmentar el habla espontánea. De hecho, algunos autores distinguen entre pausas habladas de borde (iniciales y, en menor medida, finales) y pausas habladas intermedias, ya que se supone cumplen funciones distintas y deberían ser procesadas por separado (Stolcke and Shriberg, 1996; Heeman and Allen, 1999). Finalmente, como se indicaba en el apartado 2.1.2.2, pausas habladas que no van aparentemente unidas a una reformulación, se interpretan a veces como reformulaciones reducidas, es decir, signos externos de un cambio precoz en la planificación (Heeman and Allen, 1999).

En nuestra aproximación, las pausas habladas serán consideradas como fenómenos acústicos que a veces pueden formar parte de otros fenómenos más complejos. En general, serán tratadas como palabras especiales que cumplen una función comunicativa. No distinguiremos entre pausas habladas de borde y pausas habladas internas, y en ninguna circunstancia serán consideradas en sí mismas como un tipo especial de reformulación.

2.1.3.4. Alargamientos.

A diferencia de las pausas habladas, los alargamientos (sobre todo de vocales, pero también de ciertas consonantes) han recibido un tratamiento muy limitado en la literatura (Eklund, 2001), a pesar de que ambos tipos de fenómenos desempeñan funciones muy similares en el habla espontánea, como mantener el turno o marcar una reformulación. Además, como se verá en el capítulo 3, los alargamientos muestran una distribución aún más amplia que la de las pausas habladas.

Entre lingüistas se suele utilizar el sufijo ‘:’ para indicar el alargamiento de un sonido. Este es precisamente el símbolo utilizado en algunas bases de datos de habla espontánea, como ATIS, para señalar la presencia de alargamientos. En otras bases de datos, como Switchboard, no se

proporciona forma alguna de marcar alargamientos. Sin embargo, parece conveniente definir modelos acústicos específicos de alargamientos y reconocerlos como tales, ya que aportan información útil en la comprensión e interpretación del habla espontánea.

En este trabajo, los alargamientos serán tratados como fenómenos acústicos que bien podrían marcar una reformulación, una frontera lingüística o nada en absoluto. Es decir, tendrán un estatus equivalente al de las pausas habladas.

2.1.3.5. Solapamientos.

Llamamos solapamiento a la superposición del habla de dos o más personas. No se trata estrictamente de un fenómeno sino más bien de un *accidente* del habla espontánea, producido por un conflicto en la apropiación del turno. El hablante que en un momento dado ocupa el turno hace ver que llega al final, mediante un silencio, una pausa hablada o una marca de entonación (por ejemplo, una caída simultánea de la energía de la señal y de su frecuencia fundamental), y entonces varios interlocutores tratan de apropiarse del turno simultáneamente. También puede producirse porque un interlocutor interpreta erróneamente que el hablante en posesión del turno va a finalizar su intervención, por lo que comienza a hablar, pero el primero, que no había finalizado, continúa hablando y sus voces se solapan. Otras situaciones son posibles también. Por ejemplo, las expresiones de aceptación o de asentimiento, de tipo lingüístico o fático, que un interlocutor introduce sobre el discurso de un hablante para hacerle ver que le sigue o que está de acuerdo. O también, un interlocutor que, violando las reglas implícitas de cortesía dialéctica, interviene deliberadamente sobre el discurso de otro para rebatirle o cortarle, situación que cualquiera podrá identificar en los debates televisivos.

Los solapamientos son muy habituales en diálogos naturales entre personas, no así en tareas de diálogo hombre-máquina. La máquina siempre espera la finalización del turno del usuario —que detecta mediante un tiempo mínimo de silencio. En cuanto al usuario, su interacción con la máquina no es como la que tendría con una persona, sino mucho más limitada. Puede llegar a interrumpirla en algún momento, pero en general esperará también a que finalice sus intervenciones.

Desde el punto de vista acústico los solapamientos no suponen un problema, siempre que cada señal disponga su propio canal, desacoplado de los demás. Así es, por ejemplo, en el caso de conversaciones telefónicas en las que los dos canales están separados. No obstante, podría darse el caso de que todas las señales ocupasen el mismo canal. En tales casos, los segmentos solapados deberán descartarse para entrenar modelos acústicos, pero no para entrenar el modelo de lenguaje. Por ello, será necesario marcar dónde empiezan y dónde acaban, pero manteniendo la anotación de fenómenos de habla espontánea tanto dentro como fuera de los mismos.

2.1.4. Nivel léxico.

Definimos como fenómenos de tipo léxico aquéllos que afectan a la pronunciación de las palabras, es decir, pronunciaciones no canónicas o incompletas. Estos fenómenos también aparecen en habla leída y en habla planeada. Sin embargo, su frecuencia y el grado de distorsión que producen son mucho mayores en habla espontánea, de manera que se hace imprescindible anotarlos. También consideraremos como fenómenos léxicos los sonidos guturales utilizados para aceptar o rechazar una proposición, así como las siglas deletreadas total o parcialmente y las palabras de origen extranjero cuya ortografía no ha sido asimilada todavía. Aunque los dos

últimos no podemos considerarlos como fenómenos de habla espontánea, los incluimos en el inventario para evitar errores en el entrenamiento de los modelos acústicos.

2.1.4.1. Variantes de pronunciación.

Ciertas teorías sobre la producción del habla argumentan que los movimientos articulatorios siguen siempre un camino de mínimo esfuerzo, proceso en el que se tiene en cuenta no solamente la referencia de los sonidos a producir (en su forma *canónica*), sino también la competencia de la audiencia para entender realizaciones acústicas alejadas de la norma. Así, un conferenciante andaluz que se dirija a una audiencia de diversas regiones de España, tratará de hacer el esfuerzo de articular *correctamente*, de manera que todos le entiendan. Ese mismo hablante, en otro contexto, coloquial o familiar, con amigos o personas que hayan demostrado una competencia auditiva mayor (con respecto a su forma natural de hablar), relajará notablemente la articulación de los sonidos, ya que no necesita esforzarse tanto para hacerse entender.

El habla espontánea puede producirse en situaciones bastante formales, como en el caso de una tarea de diálogo hombre-máquina, en la que el hablante no atribuye una gran competencia auditiva a la máquina. Aún en tales situaciones, si el hablante observa que la máquina le entiende, inconscientemente le irá atribuyendo una mayor competencia auditiva, e irá paulatinamente relajando su pronunciación. Así pues, cualquiera que sea el contexto en el que se produzca el habla espontánea, será común encontrar variantes de pronunciación, es decir, desviaciones de la norma debidas a características dialectales, de velocidad del habla o propias del hablante. También encontraremos, aunque en menor medida, errores de pronunciación, debidos a un exceso de carga cognitiva que impide gestionar adecuadamente el proceso de articulación, o a una velocidad de articulación demasiado elevada, que la musculatura no es capaz de seguir. El hablante no suele considerar que se produzca un problema de comprensión debido a estos errores de pronunciación, por lo que no suelen conducir a reformulaciones.

La información obtenida de las variantes de pronunciación podría utilizarse para inducir reglas o modelos de pronunciación basados en las desviaciones observadas con respecto a la pronunciación canónica.

2.1.4.2. Palabras cortadas.

Las palabras cortadas surgen por la detección tardía de un error de planificación o de acceso léxico: el hablante no llega a completar la palabra porque se da cuenta —tarde— de que quiere decir otra cosa, o la misma cosa de otra manera. Estos errores son corregidos inmediatamente mediante una reformulación: o bien la misma palabra se repite —pronunciada correctamente—, o bien otra palabra distinta sustituye a la palabra cortada. En las anotaciones será necesario indicar la transcripción de la palabra cortada pero también la de la palabra completa, que normalmente puede deducirse del contexto. Esto permitirá mantener la calidad de los modelos acústicos, por un lado, y efectuar análisis lingüísticos completos a nivel sintáctico y semántico, por otro.

2.1.4.3. Sonidos guturales de aceptación o negación.

Se trata de afirmaciones o negaciones que no se expresan lingüísticamente mediante un *sí* o un *no*, sino mediante sonidos guturales, cortos, nasalizados, con una determinada entonación que les da un significado u otro. Se consideran fenómenos de tipo léxico porque funcionan

como palabras aunque no es posible asignarles una transcripción ortográfica. No obstante, en inglés es habitual encontrar convenciones ortográficas para este tipo de sonidos. Por ejemplo, en Switchboard se utilizan *uh-huh* para el sonido gutural de afirmación, y *huh-uh* para el de negación. En nuestra aproximación, se ha optado por anotarlos como fenómenos léxicos sin asignarles una transcripción. Serán las propias muestras de estos fenómenos las que se utilicen para generar los modelos acústicos, que a su vez funcionarán como modelos léxicos.

2.1.4.4. Siglas.

Las siglas o acrónimos no tienen una forma regular de pronunciación. Algunas veces se deletrean y otras se pronuncian como palabras normales. Por ejemplo, FBI se deletrea mientras que CIA se lee como una palabra. Ello no impide que los hablantes elijan una u otra forma de pronunciación según su costumbre.

La tendencia a minimizar el esfuerzo hace que muchas siglas se pronuncien como palabras, lo cual es mucho más eficiente —y más breve— que deletrearlas. Tal es el caso de CAMP-SA, REPSOL, etc. que en nuestra aproximación no se considerarán fenómenos léxicos, ya que se pronuncian según el patrón. Por otro lado, no siempre es posible pronunciar las siglas como palabras, debido a combinaciones impronunciables de consonantes. Estos casos suelen ser tratados por los hablantes de forma mixta, con una parte que se deletrea y otra que se lee, como en CSIC, leído *ze-sik* en lugar de *ze-ese-i-ze*.

Precisamente, cuando unas siglas se deletrean, total o parcialmente, tiene lugar una pronunciación atípica que debe ser anotada como fenómeno léxico. No se trata de un fenómeno de habla espontánea, ya que en habla leída se produciría igualmente, pero resulta necesario marcarlo para generar transcripciones fonéticas adecuadas y entrenar debidamente los modelos acústicos.

2.1.4.5. Palabras extranjeras.

Nos referimos a palabras tomadas de otra lengua cuya pronunciación no se ha castellanizado, es decir, palabras que se escriben y pronuncian en su lengua original, como *software*, *light*, etc. Suelen ser palabras en inglés, sobre todo nombres propios, términos técnicos y palabras en argot. Como en el caso de las siglas, será necesario anotar una versión castellanizada de su pronunciación (*sofwar*, *lait*, etc.) para generar transcripciones fonéticas adecuadas.

2.1.5. Nivel sintáctico.

En este nivel se incluyen, en principio, todo tipo de irregularidades en la estructura de frases y expresiones —con respecto a la gramática de la lengua—, que aparecen específicamente en el habla espontánea. En este trabajo nos concentraremos en lo que se conoce como auto-correcciones o *reformulaciones* (en inglés, *speech repairs* o también *self-repairs*). Otras irregularidades, como omisiones, faltas de concordancia, conjugaciones incorrectas, etc. se considerarán características intrínsecas de la *gramática natural* del habla, y como tales serán absorbidas y se integrarán en el modelo de lenguaje, a través de técnicas de aprendizaje automático y suavizado de parámetros, que ya han sido aplicadas con éxito previamente en tareas de reconocimiento de habla leída (Torres and Varona, 2001; Varona and Torres, 2003).

Las reformulaciones podrían ser absorbidas también en el modelo de lenguaje, como el resto de fenómenos. Sin embargo, ello provocaría errores de comprensión, ya que la secuencia

reconocida de palabras podría interpretarse de varias formas. Anotar las reformulaciones podría servir para entrenar modelos específicos, que insertados/combinados con el modelo de lenguaje general, permitirían bien desechar la información innecesaria, generando cadenas de palabras con una sola interpretación, bien reconocer las reformulaciones como tales dentro de la cadena y hacer llegar esa información a otros módulos (comprensión, diálogo, etc.), que se encargarían de resolver las posibles ambigüedades, o incluso de efectuar interpretaciones más complejas sobre el estado cognitivo del hablante y sus intenciones.

En habla espontánea las repeticiones suelen desempeñar un papel similar al de las pausas habladas —mantener el turno—, y otras veces tienen función enfatizadora: el hablante repite aquello que considera primordial. Por su parte, las auto-correcciones se utilizan para corregir errores sobre la marcha (Levelt, 1989, capítulo 12). Sin embargo, dado que la estructura externa de unas y otras es muy similar, siguiendo el criterio de Shriberg vamos a representar ambos fenómenos mediante una misma categoría: reformulaciones. Haremos una excepción con los inicios fallidos de frase, considerados en otros trabajos como un tipo especial de auto-corrección, para los que definiremos una categoría aparte. La diferencia fundamental entre los fenómenos caracterizados como reformulaciones y los inicios fallidos de frase, es que los primeros contienen al menos dos elementos: el segmento a corregir y el segmento corrector, mientras que los segundos carecen de segmento corrector.

2.1.5.1. Reformulaciones.

Tanto las repeticiones como las reformulaciones propiamente dichas muestran la misma estructura externa (Shriberg, 1994), formada por cuatro elementos:

- ▷ un segmento a corregir, modificar o repetir, que denominamos *reparandum*;
- ▷ el *punto de interrupción*, que marca el final del *reparandum* y que representa el momento en que el hablante detiene el flujo normal del habla;
- ▷ un segmento opcional, que recibe el nombre de *interregnum* o *señal de edición*, que puede incluir pausas de silencio, pausas habladas o incluso palabras o expresiones; y
- ▷ un segmento, que llamaremos *corrección*, con el material que sustituye al *reparandum*.

Aunque el punto de interrupción es el elemento más significativo del fenómeno, no es necesario marcarlo explícitamente, ya que se sobreentiende que sucede al final del *reparandum*. Por tanto, basta con marcar los dos segmentos obligatorios, *reparandum* y *corrección*, más un segmento opcional, la *señal de edición*, entre ambos, como se muestra en el siguiente ejemplo:

Ejemplo 1: Reformulación con sustitución de elementos.

Resido [R en Granada] [S perdón] [C en Málaga] y me gustaría saber si hay trenes para el siete de agosto del dos mil para Granada .

En la anotación de reformulaciones no nos importa la lógica interna del hablante, salvo en algún caso especial que detallaremos más adelante. Sólo nos interesa la estructura superficial del fenómeno. No obstante, distinguiremos entre cuatro tipos de reformulaciones: (1) repeticiones, (2) sustituciones, (3) reformulaciones con inserción de elementos, y (4) reformulaciones con borrado de elementos.

Las repeticiones se identifican de manera obvia. No sucede lo mismo con los otros tipos de reformulaciones. La pista más importante para identificar una reformulación es comprobar si el *reparandum* y su *corrección* desempeñan la misma función sintáctica o son semánticamente intercambiables. En general, la parte corregida es fácil de identificar. Sin embargo, sobre el segmento corrector puede haber más de una interpretación. Considerese el siguiente ejemplo:

Ejemplo 2: Reformulación con inserción de elementos.

Quisiera saber [R horarios de] [S <pausa hablada>] [C horarios y precios de] trenes para ir de Madrid a Valencia el próximo diecinueve de junio .

En este caso el hablante, que inicialmente pedía sólo *horarios*, después ha decidido pedir *horarios y precios* de trenes para un viaje de Madrid a Valencia. En la definición del segmento corrector no se escogerán sintagmas completos, sino aquellos segmentos *mínimos* que sustituyan al *reparandum*. En el ejemplo anterior, podríamos haber elegido como *corrección* el segmento *horarios y precios de trenes*, o incluso quizá un segmento más largo, pero entonces la correspondencia entre *reparandum* y *corrección* no sería exacta.

En otras ocasiones, cabe la duda de si un segmento que añade contenidos más específicos a un segmento anterior puede constituir una reformulación. Considerese el siguiente ejemplo:

Ejemplo 3:

*En clase turista <pausa hablada>¿ qué precio es ? O sea , clase **que no sea primera** , normal .*

En este caso el segmento *que no sea primera* podría marcarse como *reparandum* y la palabra que le sigue —*normal*— como *corrección*, pero no estaríamos interpretando correctamente el fenómeno. El hablante no corrige sino que añade precisión, especifica en mayor medida lo que quiere decir. Por tanto, no podemos considerar este caso como una auto-corrección. En general, dependerá de la entonación y del contexto sintáctico y/o semántico el que consideremos un cierto segmento como una corrección o como una parte sintácticamente correcta que detalla o matiza el significado de contenidos previos.

2.1.5.2. Frases abandonadas.

La alternativa a definir una categoría especial para los inicios fallidos de frase era considerar un caso degenerado de reformulaciones en el que sólo apareciera el *reparandum*. Esta aproximación tenía la ventaja de representar todos los fenómenos sintácticos mediante un único formalismo, pero los fenómenos representados no mostraban todos ellos la misma estructura. Se ha considerado más adecuado separar los casos degenerados, y mantener la estructura básica de las reformulaciones.

Llamamos *frase abandonada* a un inicio de frase fallido, o a un fragmento en el interior de una frase que no ha sido continuado, como en los siguientes ejemplos:

Ejemplo 4: Inicio de frase fallido que incluye una palabra cortada y una repetición.

no . es que aún no he acabado . [A [R te-] [C tenía] más] <pausa hablada>quería hacerle otra pregunta ¿ es posible ?

Ejemplo 5: Fragmento abandonado.

pues sí . quería saber [A a cuánto] <pausa hablada>[R el] [C el] precio del billete de ida y vuelta en coche cama .

En el primer caso, el segmento *te- tenía más* se interrumpe, y después de una pausa, se inicia una frase completamente distinta. En el segundo caso no es un inicio de frase lo que se abandona sino un fragmento que aparece en el interior: la pregunta en estilo indirecto *a cuánto*. A continuación, tras un momento de duda marcado por una pausa y una repetición, el segmento introductorio *quería saber* es continuado por un fragmento totalmente nuevo: *el precio del billete de ida y vuelta en coche cama*. Casos como éste último pueden verse como borrados de palabras —en inglés, *verbal deletions*. De hecho, si eliminamos los fragmentos abandonados, las frases que resultan son sintácticamente correctas.

Las frases abandonadas suelen ir seguidas de alargamientos, pausas, repeticiones, determinadas expresiones de relleno, etc. cuya función es mantener el turno —ganar tiempo— mientras el hablante compone una alternativa. En otros trabajos la frase abandonada era marcada como *reparandum*, las pausas y expresiones de relleno constituían la *señal de edición*, y la *corrección* se dejaba vacía (Shriberg, 1994; Heeman, 1997). En nuestra aproximación marcaremos sólo el primer elemento. Las pausas, expresiones de relleno y otros fenómenos —a excepción de los alargamientos finales, que van indisolublemente unidos a una palabra—, aunque anotados como tales, no se incluirán en la frase abandonada ni se agruparán en una *señal de edición*.

2.1.6. Nivel pragmático.

En este nivel se consideran todos aquellos fenómenos relacionados con el acto comunicativo y con las intenciones o fines del hablante. Se trata, por tanto, de fenómenos que no aportan significado, sino eficacia en la gestión de la comunicación. Serían al habla espontánea lo que los signos de puntuación al habla escrita, ya que su objetivo sería ayudar en la comprensión, enfatizar significados, establecer estructuras discursivas y relacionarlas entre sí.

La función pragmática o interactiva es llevada a cabo fundamentalmente mediante patrones de entonación, y en gran medida también por algunos de los fenómenos que ya hemos descrito a otros niveles: pausas de silencio, pausas habladas, repeticiones, etc. Pero también participan en la estrategia conversacional unos fenómenos de tipo lingüístico que carecen de auténtico significado, los así llamados **marcadores de discurso**.

2.1.6.1. Marcadores de discurso.

Se trata de palabras o expresiones muy comunes que, como se ha dicho, no aportan contenido pero que desarrollan una función pragmática en la interacción hablada, como abrir una conversación (*hola, buenos días*) o cerrarla (*gracias, adiós, buenas tardes*), mantener el turno (*bueno, vamos a ver*), marcar una reformulación (*perdón, quiero decir*), etc. Algunas de estas expresiones actúan sólo como marcadores de discurso, pero en general expresiones y palabras que a veces actúan como marcadores de discurso, otras veces actúan como elementos léxicos normales. También es posible que la misma expresión pueda desempeñar distintas funciones como marcador de discurso. En este trabajo se han definido ocho posibles funciones: apertura, cierre, aceptación, rechazo, explicación/edición, solicitud, relleno y exclamación/énfasis. Se trata de una lista no exhaustiva, que simplemente proporciona la cobertura necesaria para marcar todas las instancias halladas en las bases de datos que hemos manejado.

Los marcadores de discurso no deben ser considerados como disfluencias, no rompen el flujo del habla, sino más bien todo lo contrario: están ahí para mantenerlo. Se incluyen en este inventario porque pueden ayudar en la segmentación automática de turnos, en la detección de

reformulaciones, en la determinación de la estructura de un diálogo, etc. Varios trabajos han demostrado ya su utilidad en la modelización del habla espontánea (Heeman and Allen, 1999).

2.2. El formato de anotación.

La documentación del CES (*Corpus Encoding Standard*) (Ide and Priest-Dorman, 2000), trabajo que forma parte de la iniciativa EAGLES (*Expert Advisory Group on Language Engineering Standards*) (Calzolari et-al., 1996), y que ha sido integrado posteriormente en el proyecto europeo MULTEXT (*Multilingual Text Tools and Corpora*) (Ide and Véronis, 1994b), incluye un informe titulado "*Background and Context for the Development of a Corpus Encoding Standard*" (Ide and Véronis, 1994a), en el que se establecen los criterios de diseño de un formato de anotación lingüística. Aunque el CES aborda una problemática muy amplia, las recomendaciones expresadas en dicho documento pueden aplicarse perfectamente a nuestro caso. Entre las características que se citan en dicho informe destacamos las siguientes:

- ▷ **completitud:** el formato debe proporcionar medios para anotar todos los fenómenos.
- ▷ **consistencia:** cada fenómeno sólo puede ser etiquetado de una forma, es decir, habrá un único y riguroso procedimiento de etiquetado. Los conceptos de etiqueta, atributo y valor deben estar basados en principios (jerarquías) muy claros y consistentes, con objeto de facilitar operaciones de validación y búsqueda.
- ▷ **recuperabilidad:** debe ser posible recuperar la información original, mediante un algoritmo sencillo. Por ejemplo, si partimos de una secuencia de caracteres, los siguientes principios podrían facilitar la recuperabilidad:
 - ningún carácter de la secuencia original será cambiado o borrado;
 - los caracteres originales aparecerán siempre como valores, nunca como etiquetas o atributos;
 - aparte de los caracteres originales, ninguna otra información aparecerá como valor; y
 - el orden original de los caracteres permanecerá inalterado.
- ▷ **validabilidad:** aquí nos referimos a la necesidad de comprobar que el documento generado encaja con la especificación del formato, es decir, que aparecen sólo marcas válidas, no falta ninguna, están anidadas correctamente, etc. Debe ser posible —y sencillo— efectuar esta comprobación. Para asegurar la validación de los documentos anotados, convendría:
 - restringir al máximo la estructura de los documentos, permitiendo sólo unas determinadas combinaciones de etiquetas, y
 - definir etiquetas suficientemente específicas, evitando en lo posible el uso de atributos.
- ▷ **compacidad:** el formato debe reducir al máximo el inventario y el tamaño de las marcas y de sus posibles atributos, así como establecer mecanismos de referencia que eviten repeticiones innecesarias. De esta forma, los documentos generados no crecerán desproporcionadamente con respecto a los originales.

- ▷ **legibilidad:** en general se dispondrá de algún software de visualización que evite la interacción directa con el documento anotado, pero en ocasiones podría ser necesario acceder directamente a dicho documento y convendría, por tanto, que fuera mínimamente legible.
- ▷ **capturabilidad:** el marcado de un documento se puede realizar de modo manual o de modo semi-automático. Obviamente es preferible automatizar este proceso al máximo, ya que el marcado manual resulta demasiado lento, teniendo en cuenta el enorme tamaño de las bases de datos. Un menor coste de marcado determina una mejor capturabilidad.
- ▷ **procesabilidad:** un formato de anotación puede encontrar dos tipos de uso: intercambio de información y procesamiento (captura, edición, búsqueda, análisis, etc.). Si se aplica al intercambio de información, deberá ser lo más general posible, independiente de la aplicación y del dominio, y deberá tener máxima capacidad expresiva, resultado de la suma de las capacidades expresivas todos los formatos que puede representar. Para ello será necesario definir complejos mecanismos de sustitución y minimización, así como características opcionales que ralentizarán los procesos de conversión. En cambio, si el formato encuentra aplicación en tareas de procesamiento —como es nuestro caso—, el criterio de diseño será reducir al máximo los costes de las mismas.
- ▷ **extensibilidad:** característica imprescindible, ya que permitirá definir nuevas marcas para fenómenos que inicialmente no estaban previstos.

2.2.1. El problema de la sincronización.

Además de las características citadas, cuando el formato de anotación trata con lenguaje hablado resulta imprescindible disponer de un mecanismo de sincronización entre las marcas correspondientes a los distintos niveles. Esta sincronización permitirá extraer correlaciones entre fenómenos, que a su vez ayudarán a resolver ambigüedades y a predecir sucesos en un determinado nivel. La sincronización entre fenómenos suele resolverse recurriendo a la señal, es decir, añadiendo a cada marca una referencia temporal. De este modo cada nivel de anotación es completamente independiente del resto de niveles, pero también pueden establecerse relaciones entre dos niveles cualesquiera, cruzando sus referencias a la señal.

La inserción automática de marcas temporales requiere de una herramienta que integre el audio, la forma de onda y la transcripción. Aún así el coste de marcado sigue siendo muy elevado, ya que la segmentación de la señal en fonemas, sílabas y/o palabras es un proceso totalmente manual y requiere escuchas cuidadosas. En la literatura se han probado con éxito estrategias mixtas, que parten de una segmentación manual *grosso modo* —por ejemplo, en palabras o pequeñas frases— y aplican después métodos automáticos de refinamiento.

2.2.2. ¿Uno o varios ficheros?

La cuestión del formato de anotación lleva aparejada la pregunta que encabeza este apartado. En la literatura podemos encontrar dos aproximaciones. La primera consiste en almacenar en ficheros distintos las anotaciones correspondientes a los distintos niveles. Esta aproximación aísla unos niveles de otros, de modo que para cada uno de ellos podemos tomar decisiones sobre cómo marcar, y definir herramientas específicas que automaticen el proceso de etiquetado. La única restricción sería la de mantener algún mecanismo de sincronización que permita entrenar y reconocer dependencias entre unos niveles y otros.

La segunda aproximación consiste en almacenar todas las informaciones, salvo típicamente la señal, en el mismo fichero. En este caso cada nivel mantiene un cierto grado de independencia, con su propio catálogo de marcas y su propia jerarquía, pero comparte formato con el resto de niveles y debe *entenderse* con ellos. Al mezclar unas marcas con otras, extraer los fenómenos correspondientes a un cierto nivel puede ser una tarea difícil. En este caso se puede optar por una sincronización *explícita* entre los niveles, basada en un nivel de referencia externo, o por una sincronización *implícita*, basada en el orden de marcado y el anidamiento de los fenómenos. Esta última opción, aunque simple, podría ser insuficiente en algunos casos, ya que pueden producirse anidamientos cruzados, es decir, solapamientos parciales entre fenómenos, siendo imposible establecer con precisión la relación entre ellos. Finalmente, compartir formato implica muy probablemente compartir también herramientas de etiquetado, lo cual significa un mejor aprovechamiento del software.

2.2.3. Estado del arte de la anotación lingüística.

El *Linguistic Data Consortium* (LDC, <http://www ldc.upenn.edu>), institución estadounidense dedicada a la creación y distribución de recursos lingüísticos y herramientas de anotación y procesamiento, y a la promoción de estándares en ese ámbito, proporciona desde una de sus secciones (*Linguistic Annotation*, <http://www ldc.upenn.edu/annotation>), un listado completo —actualizado periódicamente— de herramientas y formatos para crear y gestionar anotaciones lingüísticas, ya sea a partir de texto o de habla. En ese listado aparecen alrededor de veinte formatos —además de muchos otros enlaces a herramientas de anotación y procesado, así como a instituciones que desarrollan y ofrecen software—, lo que da una idea de la magnitud del problema y la diversidad de soluciones que la comunidad científica ha planteado (véase (Rodríguez, 1999)).

Podemos distinguir en ese listado dos tipos de formatos, según utilicen o no lenguajes de marcado basados en SGML (*Standard Generalized Markup Language*) (Sperberg-McQueen and Burnard, 1994) o XML (*eXtensible Markup Language*) (XML, 2000). Los formatos más antiguos no están basados en SGML/XML, son muy compactos y muestran una buena legibilidad, pero son poco o nada extensibles y necesitan herramientas específicas de búsqueda y validación. Entre éstos cabe destacar: (1) el formato **SAM** utilizado en los proyectos SpeechDat (Senia, 1997); (2) el formato **PARTITUR** (Schiel et al., 1998), extensión del formato SAM diseñada para el proyecto VerbMobil, financiado por el gobierno alemán (<http://verbmobil.dfki.de>); (3) el formato utilizado para generar las transcripciones de Switchboard, que es similar a los formatos utilizados para las bases de datos ATIS y Wall Street Journal (Paul and Baker, 1992); y (4) el formato de anotación de disfluencias (Meteer et al., 1995) definido también para **Switchboard**, dentro del proyecto *Penn Treebank* (Marcus et al., 1994), formato que comprende una buena parte de los fenómenos que pretendemos anotar en este trabajo.

Los formatos basados en SGML/XML requieren un mayor esfuerzo inicial de desarrollo, ya que deben acogerse a las especificaciones de estos metalenguajes. No muestran la legibilidad ni la compacidad de los formatos no SGML/XML, pero a cambio disponen de un gran volumen de software genérico que permite visualizar, transformar, validar y efectuar búsquedas —incluso a través de Internet— en los documentos generados. Además, tienen como principal característica su extensibilidad, que permite llevar a cabo refinamientos sucesivos de las anotaciones.

Entre los formatos SGML/XML, destacamos en primer lugar el ya mencionado *Corpus Encoding Standard*, una suerte de arquitectura/formato standard para bases de datos lingüísticas,

basada originariamente en SGML y conforme con la *Text Encoding Initiative* (TEI) (Ide, 1994; Burnard and Sperberg-McQueen, 1995). Actualmente se está elaborando una especificación alternativa, *XCES* (Ide et al., 2000), basada en XML.

En segundo lugar destacamos el *Universal Transcription Format* (UTF), basado también en SGML, definido por el grupo de procesamiento del habla del *National Institute of Standards and Technology* (NIST, 1998) de los Estados Unidos, y utilizado por el *Linguistic Data Consortium* para algunas de sus bases de datos.

En tercer lugar, el metaesquema y la herramienta de anotación desarrollados dentro del proyecto **MATE** se edifican sobre una arquitectura XML que permite integrar formatos existentes así como definir nuevos formatos en tareas de anotación multi-nivel de diálogos hablados. En la jerarquía de niveles definidos en dicha arquitectura, cada capa básica se anota en ficheros separados y se relaciona con otras capas a través de enlaces XML. Por ejemplo, dentro del nivel de anotación de disfluencias descrito en el apartado 2.1.2.3, las marcas correspondientes a los subniveles *recomendado* y *opcional* utilizan referencias a las marcas del subnivel *obligatorio*. A su vez, éstas hacen referencia a las marcas de palabras del *nivel ortográfico*. Finalmente, las palabras se transcriben fonéticamente mediante una capa específica que relaciona cada una de ellas con una secuencia de unidades fonéticas, que a su vez están sincronizadas con la señal de voz.

El proyecto más ambicioso en lo que se refiere a anotación lingüística es, hasta la fecha, el liderado por Steven Bird desde el *Linguistic Data Consortium* alrededor de lo que se conoce como *Annotation Graphs* (Bird and Liberman, 2001). Se trata de un marco formal para representar anotaciones lingüísticas que, siguiendo la línea del proyecto MATE, hace abstracción de esquemas de anotación, formatos e interfaces, proporcionando en cambio una capa lógica que establece las relaciones entre los distintos niveles de anotación. Dispone de un *toolkit* (Maeda et al., 2002) y diversas aplicaciones que permiten manejar distintos formatos preexistentes, así como generar transcripciones de conversaciones, árboles sintácticos, etc. (Bird et al., 2002). Aunque los *grafos de anotación* son un concepto independiente del formato de anotación, el intercambio de anotaciones se ha resuelto mediante una aplicación XML. En realidad, los *grafos de anotación* forman parte del proyecto ATLAS (*Architecture and Tools for Linguistic Analysis systems*, <http://www.nist.gov/speech/atlas/>) (Bird et al., 2000), en el que participan tanto el LDC como el NIST, y cuyo objetivo es construir una arquitectura general para gestionar anotaciones lingüísticas, así como diseñar un formato general de intercambio.

2.2.4. Un formato de anotación de fenómenos de habla espontánea.

La alternativa a diseñar un nuevo formato era utilizar alguno de los existentes. Sin embargo, ninguno de los formatos revisados resultaba adecuado, bien por ofrecer una escasa cobertura de los fenómenos que pretendíamos anotar, bien por abarcar una gama demasiado extensa de fenómenos. Por otra parte, la existencia de proyectos como MATE y especialmente ATLAS que, como se ha dicho, tratan de integrar esquemas y formatos específicos en una arquitectura general de representación de anotaciones, y proporcionan numerosas utilidades para generar y procesar anotaciones, permite definir con total libertad un nuevo esquema de anotación, sin por ello renunciar a dichas utilidades.

Siguiendo la tendencia observada en anotación lingüística, el nuevo formato de anotación se ha definido como una aplicación XML. Esta elección permitirá por un lado que el esquema *crezca* y se *especialice*, ya que el inventario puede ampliarse con marcas nuevas, con marcas

más específicas, o con nuevos atributos y valores; y por otro lado, se podrán utilizar todo tipo de herramientas XML para visualizar, validar, procesar y realizar búsquedas en nuestras bases de datos de habla espontánea. El formato define marcas para todos los fenómenos descritos en los apartados 2.1.3, 2.1.4, 2.1.5 y 2.1.6, pero también incluye marcas adicionales que permiten conformar la estructura de las anotaciones y manejar determinados eventos como solapamientos, fragmentos ruidosos o ininteligibles y cortes en la grabación.

Cada fichero de anotación representa un diálogo completo y tiene su correlato en un fichero de señal, al cual se hace referencia desde el propio fichero de anotación. Cada anotación consta de una descripción, una lista de hablantes y la secuencia de turnos. Cada turno, a su vez, contiene la transcripción de la intervención de un hablante. Así pues, todas las marcas comparten el mismo documento XML, y se sincronizan unas con otras implícitamente mediante relaciones de anidamiento y secuencia, si bien se han previsto también mecanismos explícitos de sincronización, por un lado mediante los atributos *t_ini* y *t_fin*, que marcan el inicio y el final de un fenómeno dentro del fichero de señal, y por otro mediante el elemento XML `sync`, que puede aparecer en cualquier sitio.

A continuación mostramos en detalle la estructura de estas anotaciones. De cada elemento XML se indican el tipo y estructura de sus contenidos, sus principales atributos y posibles valores. Por defecto los atributos son opcionales, y cuando no se da la lista de posibles valores, se supone que consisten en texto sin más restricciones. Cuando son obligatorios, se señala explícitamente. En cuanto a los atributos *t_ini* y *t_fin*, son admitidos como atributos opcionales por todos los elementos, salvo indicación explícita en sentido contrario. La especificación formal completa de esta aplicación XML se adjunta en el apéndice A.

Elemento anotación.

Descripción Elemento básico de las anotaciones. Representa un diálogo completo.

Estructura Un elemento de tipo `descripción` (opcional), una secuencia de elementos de tipo `hablante` (al menos uno) y una secuencia de elementos de tipo `turno` (al menos uno).

Atributos Sólo admite dos:

- ▷ *anotador*: nombre del anotador.
- ▷ *fecha*: fecha de creación de la anotación.

Elemento descripción.

Descripción Sus atributos aportan información sobre el diálogo: lengua, tipo de habla, base de datos a la que pertenece, frecuencia de muestreo, etc.

Estructura Elemento vacío.

Atributos La lista completa es:

- ▷ *base_de_datos*: nombre de la base de datos a la que pertenece la anotación.
- ▷ *lengua*: español, euskera o inglés.
- ▷ *tipo_habla*: leída, planeada o espontánea.
- ▷ *registro*: familiar, coloquial o formal.
- ▷ *tipo_discurso*: tarea, monólogo, entrevista o conversación.
- ▷ *dominio*: descripción del dominio semántico.
- ▷ *fichero_audio*: nombre del fichero con la señal.
- ▷ *frecuencia_kHz*: frecuencia de muestreo en kHz: 8 o 16.
- ▷ *codificación*: mulaw o lineal.
- ▷ *bits_muestra*: 8, 16 o 32.
- ▷ *lugar*: laboratorio, oficina o exterior.
- ▷ *canal*: micrófono o teléfono.
- ▷ *nivel_ruido*: nulo, bajo, alto o variable.

Elemento hablante.

Descripción Sus atributos aportan información sobre uno de los hablantes que intervienen en el diálogo.

Estructura Elemento vacío.

Atributos La lista completa es:

- ▷ *id*: identificador; obligatorio.
- ▷ *nombre*: nombre completo del hablante.
- ▷ *sexo*: hombre o mujer.
- ▷ *edad*: niño, adulto o anciano.
- ▷ *acento*: descripción del acento y otras características especiales del hablante.

Elemento turno.

Descripción Contiene la transcripción de la intervención de un hablante.

Estructura Un segmento solapado (elemento de tipo `o`), seguido de un segmento libre de solapamientos (elemento de tipo `t`), más otro segmento solapado. Finalmente, cada turno puede finalizar con un segmento no transcrito (elemento de tipo `no_transcrito`) o un corte en la grabación (elemento de tipo `corte`), y una marca de continuación (elemento de tipo `continua`). Todos los elementos mencionados son opcionales, es decir, pueden aparecer una vez o ninguna.

- Atributos**
- ▷ *id*: identificador; obligatorio.
 - ▷ *hablante*: referencia a un identificador de hablante; obligatorio.
 - ▷ *fichero_audio*: nombre del fichero de señal con el turno completo cortado.

Elemento t.

Descripción Segmento libre de solapamientos.

Estructura La transcripción ortográfica de la intervención de un hablante, en la que pueden aparecer cualesquiera de los siguientes elementos, en cualquier número y orden: a, n, p, f, l, r, b, d, N y sync.

Atributos Aparte de *t_ini* y *t_fin*, no tiene.

Elemento o.

Descripción Segmento solapado: segmento donde las voces de dos o más hablantes se solapan.

Estructura La transcripción ortográfica de la intervención de un hablante, en la que pueden aparecer cualesquiera de los siguientes elementos, en cualquier número y orden: a, n, p, f, l, r, b, d, N, sync, no_transcrito y corte.

Atributos Aparte de *t_ini* y *t_fin*, no tiene.

Elemento N.

Descripción Segmento ruidoso: segmento donde la voz de un hablante se solapa con ruido de fondo.

Estructura La transcripción ortográfica de la intervención de un hablante, en la que pueden aparecer cualesquiera de los siguientes elementos, en cualquier número y orden: a, n, p, f, l, r, b, d, sync, no_transcrito y corte.

- Atributos**
- ▷ *fuelle*: en el futuro podría tomar varios valores; por el momento, se le asigna un valor fijo: *exterior*.
 - ▷ *tipo*: en el futuro podría tomar varios valores; por el momento, se le asigna un valor fijo: *genérico*.

Elemento n.

Descripción Ruido aislado (no solapado con voz).

Estructura Elemento vacío.

- Atributos**
- ▷ *fuelle*: hablante o exterior; obligatorio.
 - ▷ *tipo*: aspiración, labios, tos o genérico. Los tres primeros valores se asignan sólo cuando la fuente es el hablante, y el último sólo cuando la fuente es exterior. El atributo es obligatorio.

Elemento a.

Descripción Alargamiento de un sonido.

Estructura El grafema de un sonido.

Atributos Aparte de *t_ini* y *t_fin*, no tiene.

Elemento p.

Descripción Pausa de silencio.

Estructura Elemento vacío.

Atributos Aparte de *t_ini* y *t_fin*, no tiene.

Elemento f.

Descripción *Pausa hablada.*

Estructura Elemento vacío.

Atributos ▷ *tipo*: *a*, *e*, *m* o *sin_identificar*. Estos valores representan los grafemas de los sonidos con los que se pueden asimilar las pausas habladas más frecuentes en castellano. Cuando el sonido no es asimilable a *a*, *e* o *m*, se asigna el valor *sin_identificar*. Este atributo es obligatorio.

Elemento l.

Descripción Distorsiones léxicas de todo tipo.

Estructura Distinguimos tres casos:

- ▷ en el caso de palabras cortadas o mal pronunciadas, la transcripción ortográfica de una palabra tal como ha sido pronunciada;
- ▷ en el caso de acrónimos y palabras extranjeras, la transcripción ortográfica habitual de dichos términos; y
- ▷ en el caso de sonidos guturales de afirmación o negación, el elemento está vacío.

Atributos ▷ *tipo*: *cortada*, *mal_pronunciada*, *gutural*, *acrónimo* o *extranjera*; obligatorio.

▷ *palabra*: en el caso de palabras cortadas o mal pronunciadas, este atributo contiene la transcripción ortográfica completa habitual; en el caso de sonidos guturales, el valor es *sí* cuando se trata de afirmaciones, y *no* cuando se trata de negaciones.

▷ *pronunciación*: contiene la transcripción ortográfica castellanizada de un acrónimo o de una palabra extranjera. En el caso de los acrónimos, se utilizan guiones para separar aquellos grafemas que se deletrean.

Elemento r.

Descripción Reformulación.

Estructura Un elemento de tipo *m* (*reparandum*), seguido de un elemento de tipo *s* (*señal de edición*) opcional, y un elemento de tipo *c* (*corrección*).

Atributos ▷ *tipo*: repetición, sustitución, inserción o borrado; obligatorio.

Elemento m.

Descripción *Reparandum*: segmento que el hablante va a repetir o corregir.

Estructura Transcripción ortográfica de una secuencia de palabras, mezcladas con cualesquiera de los siguientes elementos, en cualquier número y orden: a, n, p, f, l, r, d, N y sync.

Atributos Aparte de *t_ini* y *t_fin*, no tiene.

Elemento s.

Descripción *Señal de edición*: serie de palabras, pausas y/o ruidos con los que el hablante marca la presencia de una reformulación.

Estructura Transcripción ortográfica de una secuencia de palabras, mezcladas con cualesquiera de los siguientes elementos, en cualquier número y orden: a, n, p, f, l, r, d, N y sync.

Atributos Aparte de *t_ini* y *t_fin*, no tiene.

Elemento c.

Descripción *Corrección*: segmento que sustituye al *reparandum* en una reformulación.

Estructura Transcripción ortográfica de una secuencia de palabras, mezcladas con cualesquiera de los siguientes elementos, en cualquier número y orden: a, n, p, f, l, r, d, N y sync.

Atributos Aparte de *t_ini* y *t_fin*, no tiene.

Elemento b.

Descripción Frase abandonada.

Estructura Transcripción ortográfica de una secuencia de palabras, mezcladas con cualesquiera de los siguientes elementos, en cualquier número y orden: a, n, p, f, l, r, d, N y sync.

Atributos Aparte de *t_ini* y *t_fin*, no tiene.

Elemento d.

Descripción Marcador de discurso.

Estructura La transcripción ortográfica de una palabra o secuencia de palabras, en la que pueden aparecer cualesquiera de los siguientes elementos, en cualquier número y orden: a, n, p, f, l, r, N y sync.

Atributos ▷ *tipo*: abrir, cerrar, aceptar, rechazar, editar, pedir, rellenar o exclamar; obligatorio.

Elemento corte.

Descripción Corte en la grabación. Su realización acústica suele consistir en un pequeño *click* seguido de una pausa. Este fenómeno implica una ruptura de la continuidad y por tanto la desaparición de una parte del diálogo.

Estructura Elemento vacío.

Atributos Aparte de *t_ini* y *t_fin*, no tiene.

Elemento no_transcrito.

Descripción Segmento no transcrito o ininteligible.

Estructura Elemento vacío.

Atributos Aparte de *t_ini* y *t_fin*, no tiene.

Elemento continua.

Descripción Indica, como última marca de un turno, que dicho turno, cortado debido a un solapamiento o un a un ruido de fondo, continúa en el siguiente turno del mismo hablante.

Estructura Elemento vacío.

Atributos No tiene atributos.

Elemento sync.

Descripción Marca de sincronización. Se puede introducir en cualquier punto y podría servir como referencia en entrenamiento o en test.

Estructura Elemento vacío.

Atributos ▷ *t*: *offset* con respecto al inicio del fichero de señal, en segundos; obligatorio.

2.3. La herramienta de anotación.

Estrechamente ligada al formato, la herramienta de anotación consiste en una interfaz gráfica ergonómica que permite minimizar el tiempo empleado en la creación de un corpus, ya que los anotadores no tienen que operar directamente sobre los ficheros donde se almacenan las anotaciones —cuyo formato puede ser complejo o incluso ilegible—, concentrándose en marcar las características relevantes que observen en la señal. La herramienta debe proporcionar acceso a la señal de forma gráfica, de manera que el anotador pueda recorrerla y escuchar selectivamente determinados tramos para generar las anotaciones correspondientes. Finalmente, la herramienta debe ser configurable, para poder anotar diferentes bases de datos con diferentes conjuntos de marcas.

En el listado del LDC aparecen más de 30 herramientas de anotación, de las cuales más de 20 pueden descargarse libremente. Sin embargo, muchas de estas herramientas sirven sólo para anotar un determinado tipo de fenómenos, utilizan un formato muy específico, o sólo pueden compilarse o ejecutarse sobre determinadas plataformas, lo cual impide su reutilización para otras tareas, con otros formatos o sobre otras plataformas. Sólo algunas herramientas admiten por diseño su adaptación a distintas tareas y el manejo de distintos formatos sobre varias plataformas. Entre éstas destacamos las siguientes:

- ▷ **GATE** (*General Architecture for Text Engineering*) es una arquitectura que integra distintos componentes básicos en la elaboración de software para el procesamiento del lenguaje natural: una biblioteca de clases, un entorno gráfico de desarrollo, soporte para almacenamiento distribuido, un modelo de anotación basado en documentos y enlaces externos XML —similar al utilizado en el proyecto ATLAS y compatible con el XCES—, traductores a otros formatos (HTML, RTF, SGML, etc.), así como una colección de herramientas gráficas, entre ellas una herramienta de anotación y una herramienta para la definición de formatos de anotación (Cunningham, 2000, 2002).
- ▷ El proyecto **MATE** proporciona herramientas gráficas para la anotación, visualización, búsqueda y demás tareas de procesamiento de bases de datos de diálogo hablado, a varios niveles: prosódico, morfosintáctico, funcional, discursivo, etc. Entre ellas, destacan una herramienta de búsqueda y extracción de información y una herramienta que permite diseñar nuevos formatos e integrarlos en la arquitectura. La arquitectura MATE se basa en el XML para definir los distintos niveles de anotación, y al igual que GATE, utiliza profusamente los enlaces externos XML para relacionar unas anotaciones con otras, lo que se conoce como *standoff markup* (Thompson and McKelvie, 1997). Una única herramienta de anotación se adapta a cada nivel de anotación, ya que cada uno de éstos, además de un formato XML específico, debe incluir asimismo una hoja de estilo XSLT (XSLT, 1999) para transformar el XML en otro formato —normalmente HTML o XHTML— y visualizar sus contenidos de manera eficaz.
- ▷ **TASX** (*Time Aligned Signal data eXchange format*) proporciona un conjunto de herramientas para la creación y el procesamiento de bases de datos multimodales (Milde and Gut, 2002). TASX utiliza un formato de anotación XML y pone especial énfasis en la descripción de fenómenos lingüísticos. Actualmente el elemento central de TASX es una herramienta de anotación de audio y video, llamada *TASX Annotator*, que permite definir distintos niveles de anotación sincronizados entre sí, importar y exportar anotaciones en

distintos formatos, reproducir fragmentos de audio y/o video, realizar espectrogramas, extraer la frecuencia fundamental de señales de voz, etc.

- ▷ **Transcriber** proporciona un entorno gráfico para llevar a cabo la transcripción, segmentación y anotación de bases de datos de habla espontánea: noticias de radio o televisión, diálogos hombre-máquina, conversaciones entre personas, etc. (Barras et al., 2001). El formato de anotación está basado en XML y ofrece soporte Unicode para distintas lenguas. Esta herramienta ha evolucionado en estrecho contacto con el NIST y el LDC, por lo que es capaz de importar y exportar anotaciones en formato UTF. Asimismo permite manejar una base de datos interna de locutores, definir segmentos, turnos y secciones, anotar solapamientos y cambios en las condiciones ambientales, así como una amplia colección de fenómenos de habla espontánea: pequeños ruidos, pausas, distorsiones léxicas, etc. Hay una lista predefinida de eventos que puede ampliarse según las necesidades de la aplicación. La gran mayoría de estos eventos han de sincronizarse con la señal, lo cual se lleva a cabo definiendo el segmento correspondiente mediante cursores sobre la señal. Es decir, *Transcriber no sólo transcribe sino que también y sobre todo segmenta.*

Como puede verse, todas estas herramientas se basan en el uso del XML para definir el lenguaje de marcado. En el mejor de los casos, la especificación del formato mostrada en el apéndice A debería bastar para adaptar una de tales herramientas a nuestras necesidades.

No es el caso del Transcriber. Aunque en la documentación se contempla ampliar el inventario de fenómenos, no queda claro como podría efectuarse esa ampliación, y hasta qué punto la interfaz gráfica es sensible a tales ampliaciones. Al parecer, para adaptar la herramienta a otro formato XML, es necesario generar una versión especial, con una nueva especificación de formato y menús especiales. Actualmente, la herramienta está pensada para transcribir noticias de radio y televisión, el formato es cerrado y la interfaz gráfica es afín a dicho formato.

En cuanto al proyecto MATE, si bien incluye una herramienta que permite especificar nuevos formatos, en la versión *beta* final no se llegó a desarrollar un módulo imprescindible: la generación semi-automática de hojas de estilo para los nuevos formatos. Ello impedía la inserción de anotaciones de manera eficaz, ya que la herramienta de anotación no podía presentarlas. Tampoco llegó a concretarse la forma de definir las relaciones de los nuevos formatos con otros formatos preexistentes y con la señal.

En el caso de GATE, la herramienta está diseñada para llevar a cabo tareas de anotación, visualización y procesamiento de *texto*, y no parece contar con una interfaz que permita manejar señales de voz y generar anotaciones sobre las mismas. Esto quiere decir que habría que construirla sobre la infraestructura que proporciona la herramienta, lo cual, aunque factible, resultaría demasiado costoso. Finalmente, en el caso de TASX, la herramienta no había empezado a desarrollarse en el momento de tomar una decisión sobre la herramienta de anotación (principios de 2000).

Así pues, en lugar de adaptar una de estas herramientas, se ha optado por una solución más sencilla: por un lado, añadir las marcas mediante un editor de texto genérico; y por otro, visualizar, escuchar y cortar las señales de forma independiente, mediante una herramienta también genérica de visualización y reproducción de audio. Este entorno de anotación permite llevar a cabo la tarea sobre cualquier plataforma (UNIX, Windows, Mac), ya que los elementos necesarios se pueden encontrar en todas ellas. En concreto, se ha preparado un entorno de anotación sobre FreeBSD, utilizando por un lado el editor de texto *Vim* (Oualine, 2001), adecuadamente

configurado para colorear las marcas, y por otro la herramienta *WaveSurfer* (Sjölander and Beskow, 2000), que permite recorrer, reproducir y analizar señales de voz. Ambos elementos son de libre distribución y de código abierto.

Finalmente, como el formato XML definido en el apartado 2.2.4 no es lo bastante compacto para una tarea de anotación manual, con objeto de facilitar y acelerar el proceso de anotación se ha definido un formato simplificado, más legible y manejable. Este formato debe entenderse como auxiliar, ya que las anotaciones son finalmente traducidas a XML de forma automática. Para ello se ha desarrollado un analizador/traductor específico, que además permite localizar errores en las anotaciones y obtener distintos tipos de transcripciones ortográficas.

2.3.1. Formato de anotación simplificado.

En formato simplificado cada fichero almacena un diálogo en forma de secuencia de turnos. Cada turno va precedido del identificador de hablante y el identificador de turno, seguidos del carácter `:`. A continuación se suministra su transcripción ortográfica en la que determinados segmentos son anotados con fenómenos de habla espontánea. Los segmentos se acotan mediante paréntesis de apertura y de cierre. Tras el paréntesis de apertura se coloca una marca formada en el caso más general por dos letras: la primera se refiere a un elemento XML y la segunda a los valores de sus atributos. Entre estas marcas y el paréntesis de cierre va el texto marcado, salvo que el elemento XML original esté vacío, como sucede, por ejemplo, en el caso de las pausas de silencio y las pausas habladas. La excepción a esta regla la encontramos en las distorsiones léxicas, que requieren añadir junto al texto marcado el valor de uno de sus atributos. En la Tabla 2.1 se muestran las equivalencias entre los elementos y atributos XML correspondientes a fenómenos de habla espontánea y las marcas en formato simplificado.

Algunas marcas especiales, como las que indican un segmento ininteligible o un corte en la grabación, se marcan de manera especial mediante corchetes, tal como se muestra en la Tabla 2.2.

Las anotaciones en formato simplificado no admiten marcas de sincronización, ni explícitas ni en forma de atributos, ya que el formato está diseñado para generar *transcripciones enriquecidas*, no segmentaciones. Si posteriormente fuera necesario añadir marcas de sincronización, éstas se agregarían —bien de forma automática, bien de forma manual— sobre la versión XML.

En el apéndice B se muestran las transcripciones en versión XML y en versión simplificada de un diálogo hombre-máquina (B.1) y de una conversación natural entre personas (B.2), pertenecientes a dos bases de datos de habla espontánea que serán descritas con detalle en el capítulo 3.

2.3.2. Guías de anotación.

Las tareas de anotación y cortado de las señales se reparten habitualmente entre un equipo de anotadores. Esto acelera el proceso, pero también incrementa la variabilidad de las anotaciones, ya que un mismo fenómeno puede ser anotado o no, o puede ser interpretado de manera distinta por cada anotador. Para reducir dicha variabilidad se hace necesario editar una guía práctica que establezca los criterios básicos sobre qué se anota y qué no, y cómo se deben interpretar los fenómenos. En este trabajo se han producido dos guías, una para cada una de las bases de datos manejadas.

Tabla 2.1: Equivalencias entre los elementos y atributos XML y las marcas en formato simplificado.

Elemento XML	Atributo:Valor	Atributo:Valor	Marca simplificada
o	–	–	(o <i>TEXTO</i>)
N	fFuente:exterior	tipo:genérico	(nw <i>TEXTO</i>)
n	fFuente:hablante	tipo:aspiración	(na)
n	fFuente:hablante	tipo:labios	(nl)
n	fFuente:hablante	tipo:tos	(nt)
n	fFuente:exterior	tipo:genérico	(nw)
a	–	–	(a <i>grafema</i>)
p	–	–	(p)
f	–	tipo:a	(fa)
f	–	tipo:e	(fe)
f	–	tipo:m	(fm)
f	–	tipo:sin_identificar	(fb)
l	palabra:X	tipo:cortada	(lu X <i>TEXTO</i>)
l	palabra:X	tipo:mal_pronunciada	(lm X <i>TEXTO</i>)
l	palabra:sí	tipo:gutural	(lg sí)
l	palabra:no	tipo:gutural	(lg no)
l	pronunciación:X	tipo:acrónimo	(ls <i>TEXTO</i> X)
l	pronunciación:X	tipo:extranjera	(ls <i>TEXTO</i> X)
r	–	tipo:repetición	(rr <i>TEXTO</i>)
r	–	tipo:sustitución	(rs <i>TEXTO</i>)
r	–	tipo:inserción	(ri <i>TEXTO</i>)
r	–	tipo:borrado	(rd <i>TEXTO</i>)
m	–	–	(m <i>TEXTO</i>)
s	–	–	(s <i>TEXTO</i>)
c	–	–	(c <i>TEXTO</i>)
b	–	–	(b <i>TEXTO</i>)
d	–	tipo:abrir	(do <i>TEXTO</i>)
d	–	tipo:cerrar	(dc <i>TEXTO</i>)
d	–	tipo:aceptar	(da <i>TEXTO</i>)
d	–	tipo:rechazar	(dr <i>TEXTO</i>)
d	–	tipo:editar	(de <i>TEXTO</i>)
d	–	tipo:pedir	(dq <i>TEXTO</i>)
d	–	tipo:rellenar	(df <i>TEXTO</i>)
d	–	tipo:exclamar	(dx <i>TEXTO</i>)

Tabla 2.2: Marcas especiales.

XML	Formato simplificado
<corfe/>	[CORTE]
<no_transcrito/>	[NO TRANSCRITO]
<continua/>	[CONTINUA]

2.3.2.1. Guía para diálogos hombre-máquina.

En la base de datos de diálogos hombre-máquina los turnos del usuario y de la máquina se almacenan en canales separados, aunque sólo los primeros son anotados y procesados, y las señales correspondientes cortadas y almacenadas en ficheros separados. Los anotadores parten de una transcripción ortográfica preliminar en la que se han incluido algunos ruidos y pausas habladas. Se ha editado una guía básica en la que se describen los fenómenos de habla espontánea más importantes, cada uno de ellos ilustrado con ejemplos. También se especifican las marcas, tanto en formato XML como en formato simplificado, las convenciones ortográficas y los pasos que deben seguirse para generar las anotaciones y cortar las señales (Rodríguez et al., 2000).

Convenciones ortográficas. Por un lado todos los números, cantidades y ordinales deben aparecer en su versión ortográfica expandida; por otro lado, los signos de puntuación se han reducido a puntos, comas, signos de admiración y signos de interrogación, y van separados de las palabras adyacentes por un espacio en blanco.

Comas y puntos se colocan no como correlato de la señal, es decir, no como reflejo de pausas, sino como marcadores de unidades sintácticas —nótese que las pausas de silencio se marcan explícitamente. Se aplican dos criterios: (1) minimizar el uso de comas, y (2) dividir cada turno cuanto sea posible en frases independientes con sentido. En concreto, marcadores de afirmación y negación (*sí, bien, no*, etc.) y frases similares, constituidas en general por marcadores de discurso (*buenos días, muchas gracias*, etc.), van separados siempre mediante puntos, salvo que vayan seguidos de una frase subordinada (del tipo *sí, pero...* o *no, porque...*), en cuyo caso van separados por comas. Precisamente, las comas se utilizan con dos objetivos: (1) parentizar segmentos que el hablante inserta para aclarar o matizar contenidos, y (2) marcar el comienzo de una frase subordinada o coordinada.

En lo que se refiere a los signos de admiración e interrogación, se utilizan como correlato ortográfico de la entonación, es decir, acotan fragmentos de texto cuya entonación en la señal se ha identificado claramente como exclamativa o interrogativa, respectivamente. Por otra parte, para compactar la representación ortográfica, ni delante ni detrás de los símbolos de admiración e interrogación se colocan comas o puntos. Así pues, los símbolos de admiración e interrogación actúan en sí mismos como comas o puntos, según convenga, separando frases o bloques semánticos.

Procedimiento de anotación y cortado de las señales. Todos los niveles de marcado se procesan en la misma sesión, si bien se sugiere llevar a cabo las anotaciones en un cierto orden:

1. Adaptar la transcripción ortográfica de partida a las convenciones descritas en los apartados anteriores.
2. Anotar los fenómenos acústicos: ruidos, pausas, alargamientos, etc. poniendo especial cuidado en ubicarlos correctamente. Además, determinados fenómenos, como las pausas de silencio o los alargamientos, sólo se anotan cuando su duración no deja lugar a dudas —más de 200 milisegundos. Con respecto a los silencios iniciales y finales, en el momento de cortar las señales se han reducido a alrededor de 50 milisegundos, y no se marcan como tales. En cuanto a las pausas habladas, cuando su realización no corresponde a

ninguno de los tres sonidos más frecuentes en castellano (*a*, *e* o *m*), o se trata de versiones muy distorsionadas de los mismos, se anotan como *no_identificado*.

3. Anotar las distorsiones léxicas.
4. Anotar los marcadores de discurso.
5. Finalmente, por tratarse del nivel de anotación más complejo, anotar las reformulaciones y las frases abandonadas. En el caso de las reformulaciones, la mayor dificultad estriba por un lado en identificar los tres segmentos: *reparandum*, *señal de edición* y *corrección*, y por otro en definir con exactitud los niveles de anidamiento. Para identificar los segmentos se aplican dos criterios: (1) minimizar el tamaño de los mismos alrededor de la *señal de edición*, de manera que al eliminar ésta y el *reparandum* resulte una frase gramaticalmente correcta; y (2) observar la equivalencia gramatical entre *reparandum* y *corrección*. Con respecto a los niveles de anidamiento, en la práctica es raro encontrar más de dos, y suele tratarse siempre de una reformulación en la que se anidan una o más repeticiones. En algún caso, podríamos encontrar dos reformulaciones acopladas, es decir, no estrictamente anidadas. El formato de anotación no permite este tipo de estructuras, por lo que tales casos —muy raros en la práctica— deben resolverse mediante una aproximación basada en anidamientos estrictos. Por último, en el caso de frases o segmentos de frase abandonados sólo es necesario identificar el segmento abandonado.

2.3.2.2. Guía para conversaciones naturales entre personas.

En general, las descripciones y directrices contenidas en la guía básica de anotación de diálogos hombre-máquina son válidas y se aplican también a conversaciones entre personas. Así pues, se ha editado una segunda guía complementaria donde se incluyen aquellos fenómenos que sólo aparecen en conversaciones entre personas, como solapamientos, o aquellos fenómenos que no habían sido observados en diálogos hombre-máquina, como acrónimos y palabras extranjeras (Rodríguez, 2002). En esta guía se hace especial énfasis en el procedimiento de cortado de las señales, ya que a diferencia de los diálogos hombre-máquina, las conversaciones entre personas se almacenan en un único canal. La guía contiene además una lista completa y resumida de las marcas en formato simplificado, así como una descripción del entorno de anotación.

Cortado de las señales. Se han aplicado dos criterios:

1. Cortar sólo señales acústicamente aceptables. Esto significa que si una señal presenta ruido o música de fondo apreciables, o contiene voces solapadas, se descarta, ya que no puede utilizarse para estimar modelos acústicos.
2. Anotar todo lo que se escuche: ruidos de fondo, música de fondo, cortes, solapamientos, respiraciones, etc. incluso en tramos ruidosos o solapados, ya que para estimar el modelo de lenguaje es necesario disponer de la transcripción completa.

Para compatibilizar ambos criterios, las anotaciones deben indicar dónde empiezan y dónde terminan los tramos *conflictivos*, de manera que aún disponiendo de toda la transcripción, sea posible quedarse únicamente con los fragmentos de transcripción para los que se tenga señal. Para facilitar esta tarea, cada vez que se encuentra un tramo ruidoso o solapado, el final de

éste determina el final de un turno, de modo que el siguiente tramo acústicamente aceptable tiene lugar ya en el turno siguiente. Ello permite asociar a cada señal cortada un índice de turno distinto. Por ejemplo, si se produce un solapamiento entre el final de un turno y el comienzo del siguiente, el segmento solapado queda fuera de los dos ficheros de señal que se generan. Tal como muestra la Figura 2.1, en un fichero estaría la señal hasta el inicio del solapamiento y en el otro la señal a partir del final del mismo.

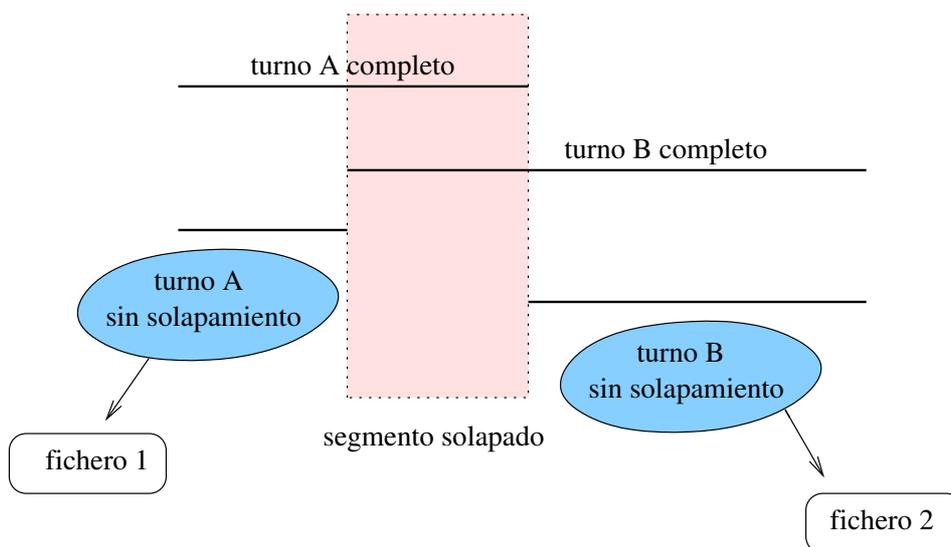


Figura 2.1: Cortado de señales en presencia de solapamientos.

Los segmentos ruidosos o solapados plantean dificultades a la hora de determinar dónde empiezan y dónde acaban. La definición de los puntos de inicio y fin se hace de forma que abarquen palabras completas. Es decir, si una palabra está afectada por un solapamiento, aunque sólo sea muy al principio o muy al final, dicha palabra queda incluida en el segmento solapado.

Procedimiento de anotación. Las conversaciones entre personas se han anotado en dos fases, con objeto de iniciar la modelización acústica lo antes posible. La primera fase se ha realizado junto con el proceso de cortado, e incluye fenómenos acústicos y léxicos. La segunda fase, prescindible desde el punto de vista de la modelización acústica, incluye los fenómenos sintácticos y pragmáticos.

En la **primera fase** los anotadores disponen de dos fuentes de información: la señales, que pueden recorrer, analizar y escuchar, y unas anotaciones preliminares en formato simplificado, obtenidas a partir de las anotaciones originales de manera automática, conversión que será descrita con detalle en el capítulo 3. Por otra parte, se aplican las mismas convenciones ortográficas y el mismo procedimiento de anotación (pasos 1, 2 y 3) definidos para diálogos hombre-máquina. En los siguientes párrafos se describen y discuten los criterios de anotación de los fenómenos no considerados previamente.

Acrónimos. Sólo se marcan como tales aquellas instancias que no encajan en el patrón de pronunciación habitual —acrónimos como CAMPSA o REPSOL no se marcan. Los casos que nos interesan consisten en palabras que se deletrean total o parcialmente. Como regla, cada

letra que se pronuncia aisladamente se separa del resto con un guión. Por ejemplo, CSIC suele pronunciarse *ze-sik*, y en formato simplificado se anotaría:

(ls CSIC C-SIC)

Palabras extranjeras. Se marcan como tales todas aquellas palabras que no forman parte del léxico castellano. Las anotaciones incluyen su transcripción ortográfica castellanizada. Por ejemplo, la palabra inglesa *light* se anotaría:

(lx light lait)

Solapamientos. Se distinguen dos tipos:

1. los que marcan la transición entre dos turnos:
 - a) suelen ser bastante largos,
 - b) en ambos casos la señal solapada forma parte (es final o inicio) de un turno,
 - c) el turno entrante (el del hablante que interrumpe) continua después de manera normal, es decir, sin solapamiento, y
 - d) normalmente estos solapamientos están ya anotados en la transcripción original de los diálogos;
2. los que suceden en el interior del turno de un hablante:
 - a) suelen ser cortos,
 - b) tienen una función fática, el hablante que interrumpe trata de mostrar acuerdo, confirmar una idea, etc. mediante una palabra, una breve expresión o incluso un sonido gutural,
 - c) dada su corta duración, estas intervenciones pueden incluso suceder entre dos palabras de la intervención principal, pero también pueden superponerse,
 - d) no forman parte de ningún turno, y
 - e) no suelen estar anotados en la transcripción original de las entrevistas.

Ambos tipos de solapamientos han recibido el mismo tratamiento: la parte solapada del primer hablante marca el final de un turno, y la parte solapada del segundo hablante marca el principio del turno siguiente. Por ejemplo:

H1[15]: no es normal que los jóvenes se reúnan en un parque con unas bolsas del supermercado y se pongan a beber como cosacos . Los padres (o deberían hacer algo) .

H2[16]: (o y la educación en la escuela) . la escuela es importante .

H1[17]: pero los padres tienen mayor responsabilidad .

Cuando el solapamiento no está marcado en la transcripción original, el turno donde aparece se divide en dos partes, una anterior y otra posterior a la interrupción, lo cual implica además crear un nuevo turno para el hablante que interrumpe. En el siguiente ejemplo H2 aprueba lo que dice H1 diciendo *claro* de forma superpuesta con *el esfuerzo*. Sin embargo, en la transcripción original no hay ninguna referencia a esa intervención de H2:

H1[6]: quiero decir que no merece la pena el esfuerzo . llevaría demasiado tiempo .

Aplicando lo dicho anteriormente nos quedaría lo siguiente:

H1[6a]: quiero decir que no merece la pena (o el esfuerzo) .

H2[6b]: (o claro) .

H1[6c]: llevaría demasiado tiempo .

La marca especial [CONTINUA]. Nótese la diferencia entre los dos ejemplos anteriores. En el primer ejemplo, el hablante H1 es interrumpido pero puede terminar su intervención y pasa a escuchar lo que H2 ha empezado a decirle. Por su parte, H2 empieza su turno hablando a la vez que H1 pero luego continúa su intervención en solitario. Cuando H1 vuelve a intervenir, lo hace para replicar a lo que acaba de decir H2 en el turno 16, pero el turno 17 no es continuación del turno 15. Por el contrario, en el segundo ejemplo H1 no llega a ser interrumpido por H2, sino que continúa su intervención como si éste último no hubiera dicho nada. Es decir, el turno 6c es continuación del turno 6a. Esta circunstancia ha de ser anotada porque desde el punto de vista del modelo de lenguaje la intervención completa de H1 estaría formada por la concatenación de los turnos 6a y 6c:

H1[6a]: quiero decir que no merece la pena (o el esfuerzo) [CONTINUA] .

H2[6b]: (o claro) .

H1[6c]: llevaría demasiado tiempo .

Ruidos de fondo, segmentos no transcritos y cortes en la grabación. La división del turno natural de un hablante en varios pequeños turnos artificiales se produce también cuando la intervención del hablante se ve solapada con ruido ambiente (risas, aplausos, música, etc.), o cuando aparecen las marcas [NO TRANSCRITO] o [CORTE]. En tales casos se utiliza también la marca [CONTINUA] para indicar la continuación del turno. Esta regla surge, una vez más, para mantener la coherencia de las anotaciones con las reglas de cortado: cada señal ha de corresponder a un turno distinto y ha de tenerse la transcripción completa de todas las señales.

Resulta difícil justificar la marca de continuación cuando se encuentra un segmento no transcrito o un corte en la grabación. Estrictamente no sabemos qué relación puede unir a los segmentos anterior y posterior a un segmento no transcrito o a un corte en la grabación. No obstante, si en la transcripción original formaban parte del mismo turno, mantendremos esa relación mediante la marca de continuación.

Por ejemplo, el siguiente turno:

H3[34]: pues hoy estábamos revisando algunas (nw de las canciones de Serrat)
cuando hemos recibido varias llamadas telefónicas felicitándonos .

se divide en los dos siguientes:

H3[34a]: pues hoy estábamos revisando algunas (nw de las canciones de Serrat)
[CONTINUA] .

H3[34b]: cuando hemos recibido varias llamadas telefónicas felicitándonos .

Esto significa que tendremos dos señales, una para el turno 34a, que no incluirá el segmento contaminado con ruido, y otra para el turno 34b. Obsérvese la utilidad de la marca [CONTINUA] para enlazar ambos turnos, puesto que efectivamente uno es continuación del otro.

Análogamente, el siguiente turno:

H1[15]: lo cierto es que mi madre acostumbraba a [NO TRANSCRITO] pueblo
que no me gustaba nada, aunque pasaba allí todo el verano .

se desdoblara en los dos siguientes:

H1[15a]: lo cierto es que mi madre acostumbraba a [NO TRANSCRITO]
[CONTINUA] .

H1[15b]: pueblo que no me gustaba nada, aunque pasaba allí todo el verano .

Nuevamente generamos dos ficheros de señal distintos, uno para el turno 15a, correspondiente únicamente a la parte transcrita, y otro para el turno 15b. Nótese cómo a pesar de que mantenemos la relación de continuidad entre los turnos, no hay relación aparente entre los bloques de texto anterior y posterior al segmento no transcrito.

En la **segunda fase** de anotación las señales ya están cortadas y los fenómenos acústicos y léxicos debidamente anotados, de modo que el anotador puede concentrarse en la detección y caracterización de los fenómenos sintácticos (reformulaciones y frases abandonadas) y pragmáticos (marcadores de discurso). En esta segunda fase se aplica el mismo procedimiento descrito para diálogos hombre-máquina (pasos 4 y 5), al cual nos remitimos. En un caso extremo, el anotador podría utilizar como único punto de partida el texto resultante de la primera fase. Sin embargo, en situaciones de duda se recomienda recurrir a la señal para confirmar la presencia o no de un cierto fenómeno.

Resumen del capítulo.

Tras una revisión bibliográfica y una descripción de los principales esquemas de anotación propuestos en la literatura, así como de los formatos y herramientas asociados, desarrollados todos ellos para la lengua inglesa, se ha propuesto un esquema de anotación de fenómenos de habla espontánea para diálogos hombre-máquina y entre personas en castellano. El esquema consta de un inventario de fenómenos, un formato basado en XML, un formato simplificado que facilita y acelera el proceso de anotación y mejora la legibilidad de las anotaciones, herramientas de libre distribución para la edición de marcas y la visualización de señales, y una guía de anotación que trata de resolver las posibles ambigüedades o problemas de interpretación que puedan surgir, mediante una descripción clara de los fenómenos —ejemplos incluidos—, todo ello encaminado a incrementar la coherencia de las anotaciones.

Capítulo 3

Estudio de los fenómenos de habla espontánea.

Para evaluar el impacto que la presencia de fenómenos de habla espontánea puede tener en el rendimiento del reconocedor es necesario obtener en primer lugar sus estadísticas de aparición. También es importante conocer la distribución de los fenómenos dentro de cada intervención, comprobando si el número de fenómenos está relacionado más con la longitud de ésta o con la identidad del hablante. De esta forma podremos determinar la utilidad potencial de modelar este tipo de dependencias, para, por ejemplo, estimar y aplicar distintos modelos acústicos y/o de lenguaje en función del hablante.

Desde el punto de vista del proceso comunicativo, algunos de los fenómenos de habla espontánea descritos en el capítulo anterior —los que entrarían en la categoría de *disfluencias*— tienen una gran importancia, ya que permiten detenerse, retroceder, corregir o enfatizar determinados contenidos. Es raro que aparezcan de manera aislada, sino que más bien se agrupan y combinan de distintas formas. Así, por ejemplo, tras una palabra cortada es frecuente encontrar pausas y reformulaciones. Sin embargo, los sistemas de reconocimiento suelen tratar los fenómenos de habla espontánea como desviaciones o problemas propios de la lengua hablada, e incorporan estrategias para detectar y eliminar este tipo de fenómenos, con el único objetivo de *limpiar* la cadena de palabras reconocida de elementos *no gramaticales*. Este filtrado es necesario ya que normalmente el modelo de lenguaje no considera dependencias en las que intervengan estos fenómenos. Esta aproximación, que es apropiada en el caso de los ruidos, sean externos o producidos por el propio hablante, no lo parece tanto para el resto de fenómenos. Como apuntan algunos autores, las disfluencias pueden cambiar y/o matizar el significado de un enunciado, y de hecho, pueden ayudar decisivamente a resolver determinadas ambigüedades de interpretación. De este planteamiento se deduce que en lugar de filtrarlos, estos fenómenos se pueden integrar en el modelo de lenguaje y participar en la comprensión de los enunciados.

En cualquier caso, es imprescindible modelar este tipo de fenómenos para evitar errores graves en la salida del reconocedor. Los fenómenos que plantean más problemas son las reformulaciones, ya que rompen completamente la estructura gramatical de las frases. Ciertos fenómenos —en particular ruidos, pausas o alargamientos— pueden ser filtrados fácilmente sin alterar el sentido del enunciado. Otros pueden ser absorbidos sin problemas por el modelo de lenguaje, como es el caso de los marcadores de discurso. Por ello, los esfuerzos en el tratamiento específico del habla espontánea se han dirigido sobre todo al procesamiento de las reformulaciones. La detección de reformulaciones podría basarse potencialmente en la pre-

sencia de determinados fenómenos de tipo acústico (pausas, alargamientos), léxico (palabras cortadas), sintáctico (palabras repetidas) o pragmático (marcadores de discurso). No obstante, antes de plantear un esquema de este tipo, conviene conocer hasta qué punto estos fenómenos pueden ser considerados como marcas fiables de la presencia de reformulaciones.

En este capítulo, en primer lugar se describen dos bases de datos de habla espontánea, una de ellas ligada a una tarea específica, y por tanto con un vocabulario y una gramática bastante restringidos, y la otra completamente genérica, no ligada a ningún dominio semántico en particular, por lo que tanto el vocabulario como el modelo de lenguaje son generales. A continuación se realiza un estudio pormenorizado de las estadísticas de aparición de los fenómenos de habla espontánea en cada una de ellas, y se exploran las dependencias con respecto a la duración de las intervenciones y con respecto al hablante. Finalmente, se estudia hasta qué punto la presencia de fenómenos acústicos, léxicos y pragmáticos puede determinar la presencia de reformulaciones.

3.1. INFOTREN: una base de datos de diálogos hombre-máquina.

La base de datos que en adelante llamaremos INFOTREN fue adquirida en el marco de un proyecto conjunto entre varias universidades españolas, denominado BASURDE (Bonafonte et al., 2000), que seguía la estela de otros proyectos similares llevados a cabo en la Unión Europea, como el proyecto ARISE (den Os et al., 1999). El objeto del proyecto era diseñar y desarrollar una aplicación informática que permitiera acceder automáticamente a información sobre horarios, trayectos y precios de viajes en tren entre ciudades españolas, mediante una interfaz hablada basada en diálogos abiertos con los usuarios.

Como es lógico, al iniciar el proyecto no se disponía de un sistema de diálogo y por tanto las señales necesarias para entrenar los distintos módulos del sistema debían recogerse sobre un *simulacro* de sistema. Esto se conoce como estrategia del *Mago de Oz*. Se recluta un cierto número de voluntarios, y a cada uno se le asigna un guión. Se le dice, por ejemplo, que tiene que viajar de Barcelona a Bilbao, saliendo un viernes por la noche y volviendo antes del martes. El usuario interactúa con el sistema pensando que se trata de una máquina, en las mismas condiciones en las que se supone lo hará cuando todos los elementos estén desarrollados. Sin embargo, independientemente de que determinadas partes del sistema estén realmente automatizadas, es un operador humano el que maneja los mandos detrás del telón. El operador humano, al que en adelante llamaremos *mago*, escucha los requerimientos del usuario y, ayudado por una aplicación informática, accede en tiempo real a la información de la compañía ferroviaria y decide el contenido y hasta cierto punto la forma de la respuesta que ha de llegarle al usuario. En ocasiones el mago puede simular un error de comprensión y pedirle al usuario que repita su pregunta o requerimiento. Un módulo de síntesis es el que se encarga de producir dicha respuesta, lo cual contribuye a reforzar en el usuario la sensación de que interactúa con una máquina.

La respuesta del mago solía demorarse entre treinta y sesenta segundos, de ahí que no resultasen diálogos naturales, sino más bien una serie de requerimientos puntuales del usuario que eran atendidos uno a uno por el sistema. Puesto que disponía de un guión, el usuario había planificado su discurso con antelación y trataba de especificar todas sus necesidades en el primer turno. Dicho turno era por lo general mucho más largo que el resto y contenía también el número más alto de fenómenos de habla espontánea. Por otra parte, durante los prolongados tiempos de

espera el usuario podía planificar de nuevo sus requerimientos. Así pues, ni el contenido ni la forma de los enunciados planteaban incertidumbres a los usuarios. Por otra parte, tampoco encontraban problemas de sincronización con la máquina, puesto que: (1) los usuarios no podían interrumpir un turno del sistema una vez comenzado; y (2) el sistema esperaba a la finalización del turno del usuario para intervenir. Esto significa que determinados fenómenos de tipo interactivo comunes en diálogos entre personas, como solapamientos de voces, sonidos guturales de aceptación, interpelaciones, etc. no van a aparecer en absoluto en estos diálogos.

Al menos en principio, las condiciones de interacción descritas en el párrafo anterior parecen limitar notablemente la posibilidad de que se produzcan fenómenos de habla espontánea como pausas habladas, palabras cortadas, repeticiones, reformulaciones, etc. ya que no se dan las circunstancias —presión cognitiva o presión interactiva— para que aparezcan. Sin embargo, en la práctica sí que se produce un número apreciable de fenómenos, lo cual quiere decir que dichos fenómenos son *parte intrínseca*, es decir, *recursos* de la lengua hablada, y que no se producen únicamente como respuesta a *perturbaciones externas*. Por otra parte, esta aplicación se planteaba como una primera aproximación a la problemática del habla espontánea. De ahí que este reducido grado de espontaneidad no deba verse como una limitación, sino como una característica deseable para poder abordar el reconocimiento automático del habla espontánea con la tecnología disponible.

INFOTREN consta de 227 diálogos, que suman 1657 turnos de usuario con una duración total de 150 minutos. Los diálogos fueron mantenidos a través del teléfono en entorno de oficina, lo cual quiere decir que aunque no absolutamente libres de ruidos, las señales —adquiridas a 8 kHz— tienen una calidad aceptable. Además de las señales —sistema y usuario en canales separados— la base de datos contiene las transcripciones ortográficas, que han sido enriquecidas con fenómenos de habla espontánea. La anotación de estos fenómenos se realizó en dos *pasadas*. En la primera pasada los 227 diálogos se repartieron entre tres anotadores, quienes ya habían participado en la especificación del inventario de fenómenos, en la definición del esquema de anotación y en la elaboración de la guía descrita en el apartado 2.3.2.1, que fue aplicada rigurosamente. En la segunda pasada un único anotador —más concretamente, el propio autor— se encargó de revisar todas las anotaciones, con el propósito de unificar criterios y eliminar las incongruencias que pudieran hallarse.

En los diálogos intervienen 75 locutores, 43 hombres y 32 mujeres, a cada uno de los cuales le fueron asignados tres guiones, con fechas, horarios y otras condiciones para viajes de largo recorrido en tren entre dos ciudades españolas. Para favorecer la espontaneidad de las interacciones, se indicó a los locutores que hablaran de forma natural, ya que el sistema automático iba a ser capaz de reconocerles. Estas indicaciones fueron interpretadas de distinto modo por cada locutor, y como resultado tanto las duraciones de los turnos como el número de turnos por diálogo presentan una gran variabilidad.

Aunque el número medio de turnos por diálogo es de 7.3, en la práctica este número oscila entre 2 y 35 turnos (véase la Figura 3.1). Se observan, por tanto, dos tendencias extremas. Por un lado, la de aquellos voluntarios que no ponen demasiado interés en la tarea, o que rápidamente obtienen una respuesta satisfactoria. Y por otro, la de los voluntarios que tratan de *exprimir* las posibilidades del sistema y mejorar sus opciones de viaje. También, aunque la duración promedio de un turno es de 5.43 segundos, las duraciones de los turnos oscilan entre 0.5 y 55 segundos (véase la Figura 3.2). Los turnos cortos corresponden a respuestas simples, como *Sí* o *No*, *gracias*, y los turnos largos, a requerimientos complejos que constan de varias frases y que, en general, muestran un gran número de fenómenos de habla espontánea.

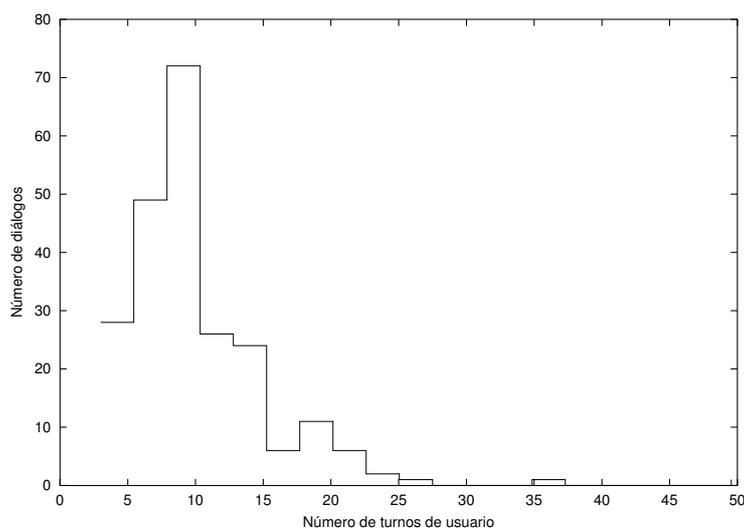


Figura 3.1: Histograma de los diálogos con respecto al número de turnos en INFOTREN.

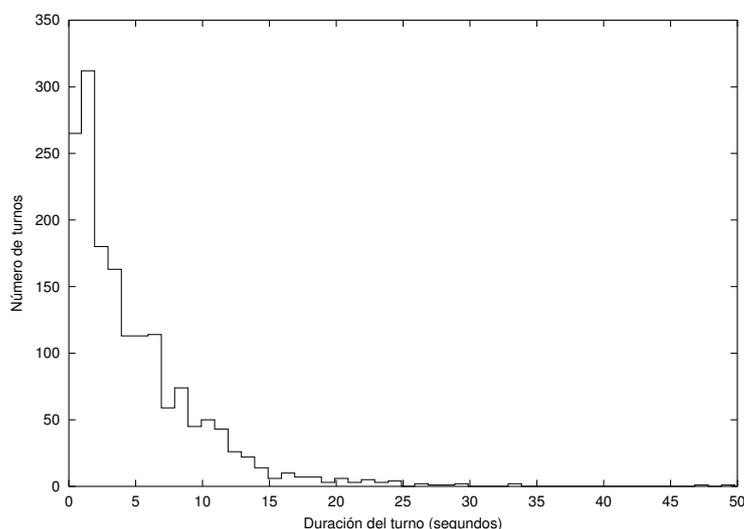


Figura 3.2: Histograma de los turnos de usuario con respecto a la duración de los mismos en INFOTREN.

Las señales producidas por el *magó*, así como sus transcripciones, son almacenadas de cara a entrenar modelos de diálogo, pero lógicamente no serán utilizadas para entrenar ni los modelos acústicos ni el modelo de lenguaje del reconocedor, ya que éstos deberán operar sólo sobre los turnos de usuario. En la Tabla 3.1 se resumen las características de INFOTREN.

3.2. CORLEC: una base de datos de diálogos naturales entre personas.

La espontaneidad de los hablantes que participan en una tarea como INFOTREN puede diferir tanto en tipología como en intensidad de la que de forma natural se produce en situaciones cotidianas. Así pues, creemos que puede ser interesante estudiar del habla espontánea

Tabla 3.1: Principales características de la base de datos de diálogos hombre-máquina INFOTREN.

Diálogos	227	Hablantes	75
Turnos	1657	Hombres	43
Duración	2.5 horas	Mujeres	32
Promedio turnos/diálogo	7.3	Número de palabras	18734
Promedio duración de turno	5.43 segundos	Tamaño del vocabulario	788
Frecuencia de muestreo	8 kHz	Promedio muestras/palabra	23.77
Canal	teléfono		
Entorno	oficina		

desde un punto de vista general, con objeto de disponer de una referencia con la que comparar el habla obtenida en tareas específicas. Para llevar a cabo dicho estudio se requiere la adquisición y anotación de habla espontánea en circunstancias no controladas: monólogos, conversaciones entre varias personas, entrevistas, etc.

La tarea de adquisición y anotación de una base de datos de las características mencionadas supone un trabajo de preparación y postproceso enorme que puede llevar varios años. Tal es el caso de la conocida base de datos de habla espontánea genérica en lengua inglesa *Switchboard* (Godfrey et al., 1992). Por esta razón se ha preferido utilizar una parte del *Corpus Oral de Referencia de la Lengua Española Contemporánea de la Universidad Autónoma de Madrid* (CORLEC, 1992) —CORLEC en lo sucesivo—, cuyas características encajan casi perfectamente con nuestras necesidades. El corpus contiene tanto las señales como las transcripciones, y éstas a su vez incluyen muchos de los fenómenos que deseamos anotar, como palabras cortadas, palabras mal pronunciadas, pausas, etc. (Ballester et al., 1993). No obstante, CORLEC presenta un gran inconveniente: la calidad de las grabaciones es por lo general muy baja, ya que fueron realizadas en condiciones ambientales adversas, con material muy básico (grabadora y cinta de audio) y depositando la señal en un solo canal, lo cual dificulta enormemente el proceso de transcripción. De hecho, para el grupo de investigación que llevó a cabo la adquisición, una vez realizadas las transcripciones las señales tenían apenas valor de referencia, ya que su interés se centraba en aspectos morfológicos, sintácticos y semántico-pragmáticos de la lengua hablada, no tanto en aspectos acústicos, fonéticos o fonológicos. En definitiva, CORLEC muestra un alto grado de espontaneidad pero también condiciones ambientales y de canal adversas.

3.2.1. Definición de un corpus reducido: CORLEC-EHU.

CORLEC está formado por diálogos entre dos o más personas, grabados en radio, en televisión, en la calle, en aulas de clase, en mesas redondas, en casa, etc. Se trata, por tanto, de un corpus de habla espontánea en castellano, que abarca diversos dominios semánticos y pragmáticos, ya que los diálogos se desarrollan en distintos contextos socio-culturales y en distintas situaciones, algunas más formales, otras más familiares.

Nuestro objetivo es disponer de una base de datos genérica de habla espontánea, lo suficientemente grande como para obtener modelos acústicos robustos, con un vocabulario y una sintaxis no restringidos a una tarea. Sin embargo, CORLEC resulta demasiado grande, por lo que no se ha procesado todo el corpus, sino sólo una parte, cuyas dimensiones se ajustan —creemos— al objetivo enunciado más arriba.

Las cuentas realizadas sobre las anotaciones originales arrojan un total de 941386 palabras, con un tamaño del vocabulario de 39785 palabras. La base de datos consta de 17 bloques o

secciones, creados atendiendo al área temática o al tipo de registro del habla. En la Tabla 3.2 se muestra el identificador de cada uno de los bloques, el tipo de conversaciones que contiene, así como el número de palabras total, el tamaño del vocabulario y el promedio de muestras por palabra. Este último es un dato importante para el modelo de lenguaje del reconocedor.

Tabla 3.2: Estadísticas de los diferentes bloques que conforman la base de datos CORLEC: número de palabras, tamaño del léxico y promedio de muestras por palabra (M/P). El tamaño del vocabulario es de 39785 palabras, sobre un total de 941386.

Bloque	Contenido	Palabras	Léxico	M/P
adm	Administrativos	6322	1080	5.85
cie	Científicos	35172	4857	7.24
con	Conversacionales	207748	14808	14.03
deb	Debates	81928	8557	9.57
dep	Deportivos	47165	5597	8.43
doc	Documentales	26779	4721	5.67
edu	Educativos	59240	6429	9.21
ent	Entrevistas	147468	12813	11.51
hum	Humanísticos	53432	7150	7.47
ins	Instrucciones	7175	1321	5.43
jur	Jurídicos	34386	4247	8.10
lud	Lúdicos	50347	6356	7.92
not	Noticiario	65373	8389	7.79
pol	Políticos	48604	5864	8.29
pub	Publicitarios	24896	3864	6.44
rel	Religiosos	11162	2298	4.86
tec	Técnicos	34687	4333	8.01

Al no tratarse de diálogos enfocados todos ellos a una misma tarea, sino cada uno sobre un tema distinto, tendremos como resultado una gran dispersión léxica, es decir, muy pocas muestras de cada palabra, lo cual podría traducirse en un modelo de lenguaje infraentrenado. Como dato de referencia, el número promedio de muestras por palabra obtenido para la tarea INFOTREN era de aproximadamente 23. Observando la Tabla 3.2, vemos que —sorprendentemente— son dos bloques no temáticos: conversaciones y entrevistas, los que arrojan un mayor *ratio* de muestras por palabra (14.03 y 11.51, respectivamente). Si unimos los dos bloques en uno solo, se obtiene un total de 355216 palabras, y un tamaño del léxico de 21537 palabras, de modo que el *ratio* de muestras por palabra resulta aún mayor: 16.49.

Por otra parte, dado que nuestro interés principal es estudiar los fenómenos de habla espontánea, para cada bloque se han contado los fenómenos ya presentes en las anotaciones originales de CORLEC: palabras cortadas, fonemas borrados, pausas habladas, silencios, sonidos guturales o interjecciones con función fática, ruidos, etc. Resulta un total de 92412 fenómenos, y un promedio de 0.098 fenómenos por palabra. Los datos desglosados por bloques se muestran en la Tabla 3.3, donde por un lado vemos que el 41.63 % de los fenómenos se producen en las conversaciones y entrevistas, y por otro que ambos bloques muestran un promedio de fenómenos por palabra alrededor de la media: las conversaciones claramente por encima y las entrevistas ligeramente por debajo. No obstante, en cuanto a la *densidad* de fenómenos no hay grandes diferencias: si bien entre el mínimo (0.058) y el máximo (0.180) hay un factor de 3, nótese que 13 de los 17 bloques muestran valores comprendidos entre 0.08 y 0.12.

A partir de los datos anteriores, se escogieron los bloques de conversaciones y entrevistas

Tabla 3.3: Estadísticas de fenómenos de habla espontánea presentes en las anotaciones originales de CORLEC, desglosadas por bloques: número absoluto de fenómenos (NFHE), porcentaje que este número representa sobre el total (%) y *ratio* de fenómenos por palabra (FHE/P).

Bloque	NFHE	%	FHE/P
adm	1136	1.23	0.1797
cie	3634	3.93	0.1033
con	25095	27.16	0.1208
deb	6651	7.20	0.0812
dep	4326	4.68	0.0917
doc	2139	2.31	0.0799
edu	5778	6.25	0.0975
ent	13376	14.47	0.0907
hum	4723	5.11	0.0884
ins	809	0.88	0.1128
jur	3624	3.92	0.1054
lud	5066	5.48	0.1006
not	3804	4.12	0.0582
pol	4072	4.41	0.0838
pub	2846	3.08	0.1143
rel	769	0.83	0.0689
tec	4564	4.94	0.1316

como candidatos para conformar un corpus reducido. El bloque de conversaciones consta de diálogos por lo general abiertos, que suelen implicar a más de dos personas e incluyen múltiples solapamientos, ya que no hay un moderador que conceda turnos. Las entrevistas son conversaciones más formales, normalmente entre dos personas, una de ellas haciendo las veces de entrevistador/moderador. En general la calidad acústica de las entrevistas es aceptable, ya que en su mayor parte han sido tomadas de la radio, mientras que entre las conversaciones hay muchas grabadas en la calle, en bares o en reuniones caseras, con muchos ruidos, eco, etc.

Por tanto, se plantea un segundo criterio de selección basado en la calidad acústica de las señales. Se trata de descartar, por un lado, diálogos demasiado ruidosos, y por otro, diálogos con un gran número de solapamientos. Tras escuchar y evaluar subjetivamente todas las señales candidatas, se consideraron acústicamente aceptables 67 de un total de 79 entrevistas y 65 de un total de 126 conversaciones, que conforman un corpus de 132 diálogos, al cual nos referiremos en adelante como CORLEC-EHU.

3.2.2. Filtrado de las transcripciones originales de CORLEC-EHU.

Las anotaciones originales de CORLEC se ubican en ficheros de texto con una pequeña cabecera seguida de las transcripciones. Tanto la cabecera como las transcripciones contienen marcas de tipo SGML, tomadas de un inventario de marcas distinto al que se ha descrito en el capítulo 2, y aplicando también distintas convenciones ortográficas (Ballester et al., 1993). Así pues, ha sido necesario escribir y aplicar un pequeño *script* Perl para pasar las 132 transcripciones de CORLEC-EHU a formato simplificado, según el siguiente conjunto de reglas:

- ▷ La cabecera se elimina.
- ▷ La transcripción de cada turno irá precedida por el identificador del locutor y el índice del turno.

- ▷ Las dobles comillas (") y los puntos suspensivos aislados (...) se eliminan.
- ▷ Los caracteres dos_puntos (:) y punto_y_coma (;) son sustituidos por comas (,).
- ▷ Las palabras cortadas, así como las siglas, palabras extranjeras y silencios simplemente cambian de formato.
- ▷ Los borrados de fonemas se anotan como palabras mal pronunciadas.
- ▷ Los puntos suspensivos (...) cuando van inmediatamente tras una palabra se anotan como alargamiento de fonema más pausa de silencio, si la palabra termina en vocal o en 'n', 'l' o 's', o simplemente como pausa de silencio en caso contrario.
- ▷ Los sonidos fáticos de afirmación y negación se transcriben como afirmaciones y negaciones guturales, respectivamente.
- ▷ El resto de sonidos fáticos, así como las marcas de vacilación y la secuencia "eh..." se transcriben como pausas habladas.
- ▷ Las marcas "<ininteligible>" y "<texto no transcrito>" se transcriben como segmentos no transcritos.
- ▷ Los borrados involuntarios se transcriben como cortes.
- ▷ Las marcas "<simultáneo> X" y "X </simultáneo>" se transcriben ambas como solapamientos.
- ▷ Las marcas que indican el tipo de habla o el modo en que se produce el habla: "<texto leído>", "<cantando>", "<onomatopéyico>" y "<argot>", se eliminan.
- ▷ Toses, carraspeos y demás marcas de ruidos (aplausos, risas, música, etc.) se transcriben todas como ruido de fondo genérico.

3.2.3. El subcorpus CORLEC-EHU-1.

Solamente ha sido posible anotar una parte del corpus CORLEC-EHU, que llamaremos CORLEC-EHU-1. La primera fase del proceso, que incluye el cortado de las señales y la anotación de fenómenos acústicos y léxicos (véase el apartado 2.3.2.2), ha corrido a cargo de un único anotador, contratado y entrenado a tal efecto. En esta fase se han revisado 47 de las 67 entrevistas de CORLEC-EHU, de las cuales 42 han sido anotadas y 5 descartadas debido a sus deficientes condiciones de grabación. En total se han empleado 180 horas de tiempo real para anotar 6.41 horas de señal, lo que significa que cada hora de diálogo ha llevado cerca de 30 horas de trabajo de cortado y anotación. Obviamente, estos números podrían reducirse aumentando la ergonomía de las herramientas de anotación, pero baste el dato para poner de manifiesto una vez más los costes enormes asociados a la creación de bases de datos de habla. La segunda fase de anotación, que incluye fenómenos sintácticos y pragmáticos, ha corrido a cargo de dos anotadores (uno de ellos, el propio autor), a partes iguales. Finalmente el propio autor se ha encargado de revisar las anotaciones para asegurar la coherencia de las mismas y corregir posibles errores.

Tabla 3.4: Principales características de la base de datos de diálogos naturales entre personas CORLEC-EHU-1.

Diálogos	42 (27 radio, 15 TV)
Turnos	2873 (2137 útiles)
Duración	6.41 horas (5.67 útiles)
Promedio turnos/diálogo	50.88
Promedio duración de turno	9.54 segundos
Frecuencia de muestreo	16 kHz
Canal	grabadora analógica
Entorno	estudios de radio/TV
Hablantes	116 (105 útiles)
Hombres	80 (72 útiles)
Mujeres	36 (33 útiles)
Hablantes radio	70 (57 hombres, 13 mujeres)
Hablantes TV	46 (23 hombres, 23 mujeres)
Número de palabras	72462
Tamaño del vocabulario	8253 palabras
Promedio muestras/palabra	8.78

En la Tabla 3.4 se resumen las características del subcorpus CORLEC-EHU-1. Las 42 entrevistas que lo conforman contienen 2873 turnos, de los cuales 2137 son turnos *acústicamente útiles*, es decir, no completamente ruidosos o solapados. Estos 2137 turnos suman una duración de 20394 segundos (5.67 horas), es decir, un 88.34 % de la duración total de los diálogos. El número de turnos útiles por diálogo oscila entre 17 y 165, la media es 50.88 y la desviación típica 29 (véase su distribución en la Figura 3.3). La duración media de cada turno útil es de 9.54 segundos, pero con una desviación típica muy alta (12.14). De hecho, el histograma de duraciones (Figura 3.4) muestra un pico muy claro de turnos muy cortos (de 0 a 5 segundos) —con poblaciones de más de 100 turnos—, que va seguido de poblaciones paulatinamente menores, hasta llegar a los 50 segundos, donde la población se hace prácticamente nula. En números: por un lado, hay 1102 turnos de duración inferior a 5 segundos (un 51.57 % de todos los turnos), que suman una duración de tan sólo 2001 segundos (un 9.81 % de la duración total); por otro lado, sólo hay 37 turnos de duración superior a 50 segundos (un 1.73 % de todos los turnos), que suman una duración de 2115 segundos (un 10.37 % de la duración total).

Los datos anteriores reflejan el esquema de interacción típico de una entrevista, con preguntas bastante concisas por parte del entrevistador, seguidas de largos monólogos a cargo del entrevistado, con eventuales gestos de asentimiento, acompañados a veces de sonidos guturales y otras veces de palabras como *Claro*, *Por supuesto*, *Sí*, *Ya*, etc.

3.3. Distribución de los fenómenos de habla espontánea.

3.3.1. Distribución de fenómenos en INFOTREN.

Las transcripciones ortográficas de INFOTREN contienen 18734 palabras, sin contar ruidos, pausas de silencio, pausas habladas y signos de puntuación. Por otra parte, el *número efectivo* de palabras —aquí se sigue la definición dada en (Shriberg, 1994), es decir, se excluyen las palabras que aparecen en los *reparandum* y en las *señales de edición* de las reformulaciones, y las palabras que aparecen dentro de frases abandonadas— es de 17884. Finalmente, el número

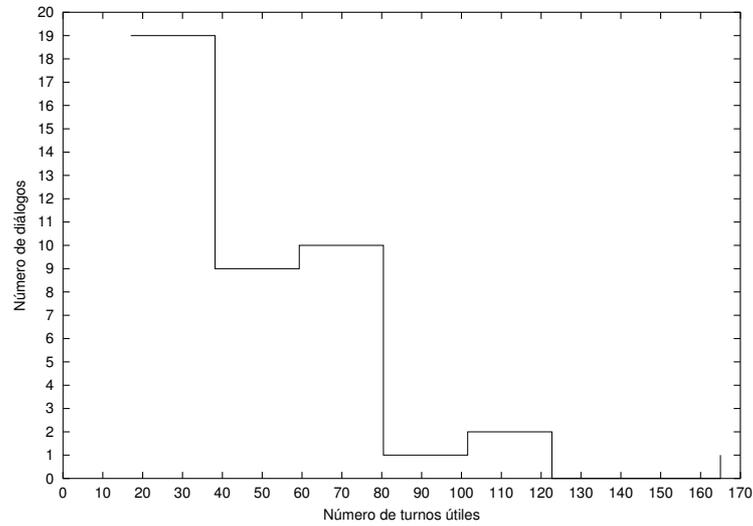


Figura 3.3: Histograma de los diálogos con respecto al número de turnos útiles en CORLEC-EHU-1.

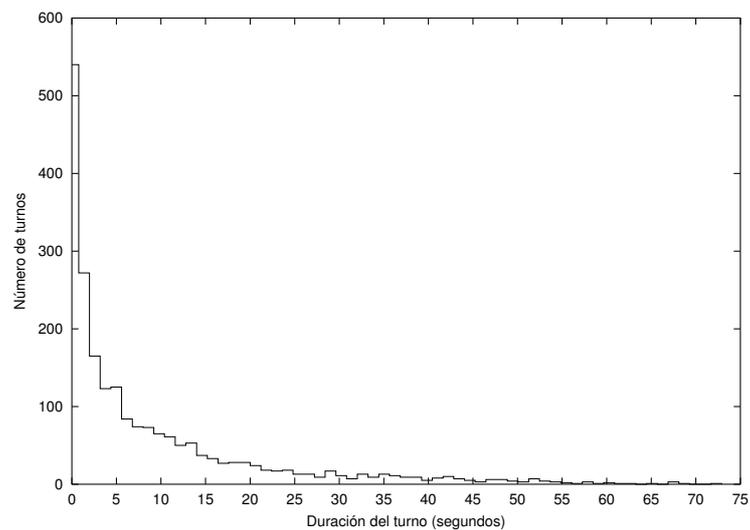


Figura 3.4: Histograma de los turnos útiles con respecto a la duración de los mismos en CORLEC-EHU-1.

total de fenómenos anotados es de 7084. Nótese que algunos de estos fenómenos están anidados dentro de otros. Por ejemplo, algunas de las pausas habladas y muchas de las palabras cortadas aparecen dentro de reformulaciones.

En la Tabla 3.5 se muestran las estadísticas correspondientes a los fenómenos de habla espontánea anotados en INFOTREN: el número absoluto de apariciones, el número promedio de fenómenos por cada 100 palabras efectivas, el porcentaje de fenómenos con respecto a la categoría a la que pertenecen y el porcentaje que representa cada categoría con respecto al total. Los fenómenos acústicos se han separado en dos bloques. Por un lado se consideran los ruidos, y por otro, el resto de fenómenos acústicos: pausas de silencio, pausas habladas y alargamientos, a los que en adelante nos referiremos como *disfluencias acústicas*. También utilizaremos el nombre de *disfluencias sintácticas* para referirnos a los fenómenos de tipo sintáctico.

3.3.1.1. Ruidos.

Los ruidos, que muchos autores ni siquiera consideran en sus estudios, representan un 37.75 % del total de fenómenos: se producen 15 ruidos cada 100 palabras efectivas. Si extrajéramos los ruidos del inventario, el número de fenómenos se vería reducido a 4410, pero todavía tendríamos un promedio de 24.65 fenómenos por cada 100 palabras efectivas.

3.3.1.2. Disfluencias acústicas.

Las pausas de silencio, las pausas habladas y los alargamientos representan en conjunto el 39.53 % de los fenómenos: se producen más de 15 disfluencias acústicas por cada 100 palabras efectivas. Este dato pone de manifiesto la importancia que estos fenómenos tienen en el habla espontánea. Las pausas habladas y los alargamientos presentan cuentas casi idénticas: 36.71 % y 36.39 % de las disfluencias acústicas, respectivamente. Esto significa, desde nuestro punto de vista, que ambos mecanismos resultan igualmente útiles a los hablantes, y que su importancia como recursos de la lengua hablada es similar. Las pausas de silencio son menos frecuentes (26.89 % de las disfluencias acústicas) y, aunque no queda reflejado en las anotaciones, muestran una gran variabilidad en su duración, que va desde los 200 milisegundos a los 2 segundos. En general, las pausas de silencio no son consideradas en sí mismas como disfluencias, sino que más bien intervienen o aparecen como parte de otros fenómenos que sí lo son, típicamente como *señal de edición* en reformulaciones. Si excluyéramos del inventario tanto los ruidos como las pausas de silencio, el número de fenómenos se reduciría a 3657, de los cuales 2047 (un 56 % del total) serían disfluencias acústicas, y aún tendríamos alrededor de 20 fenómenos por cada 100 palabras efectivas.

Pausas habladas. Las pausas habladas pueden realizarse acústicamente de distintas maneras. Estas realizaciones cambian de unas lenguas a otras, y pueden variar también entre locutores. En inglés, por ejemplo, la realización más común es la que se transcribe como *uh* (Shriberg, 1994). En castellano, tal como muestra la Tabla 3.5, hay una clara preferencia por la vocal /e/ (53.11 % de los casos), seguida de una especie de nasalización (17.41 % de los casos), que en inglés se suele transcribir como *um* y que aquí transcribimos con el sonido /m/. Finalmente, aparece la vocal /a/ (9.05 % de los casos). También se cuenta un número importante de realizaciones difíciles de identificar o distorsionadas (20.43 % de los casos).

Tabla 3.5: Estadísticas de los fenómenos de habla espontánea anotados en INFOTREN: número absoluto de apariciones (#FHE), número promedio de fenómenos por cada 100 palabras efectivas (%FHE/PE), porcentaje de fenómenos con respecto a la categoría a la que pertenecen (%FHE/C) y porcentaje que representa cada categoría con respecto al total (%C/Total).

Categoría / Subcategoría		Tipo	#FHE	%FHE/PE	%FHE/C	%C/Total
Ruidos		genérico	661	3.70	24.72	37.75
		aspiración	1404	7.85	52.51	
		labios	600	3.35	22.44	
		tos	9	0.05	0.34	
Disfluencias acústicas	Pausas de silencio	–	753	4.21	26.89	39.53
	Pausas habladas	a	93	0.52	3.32	
		e	546	3.05	19.50	
		m	179	1.00	6.39	
Alargamientos	–	210	1.17	7.5		
Distorsiones léxicas		cortada	95	0.53	47.50	2.82
		mal pronunciada	105	0.59	52.50	
Disfluencias sintácticas	Reformulaciones	repetición	292	1.63	53.58	7.69
		sustitución	141	0.79	25.87	
		inserción	37	0.21	6.79	
		borrado	5	0.03	0.92	
	Frases abandonadas	–	70	0.39	12.84	
Marcadores de discurso		abrir	150	0.84	17.34	12.21
		cerrar	189	1.06	21.85	
		aceptar	78	0.44	9.02	
		rechazar	45	0.25	5.20	
		explicar/editar	71	0.40	8.21	
		pedir	92	0.51	10.64	
		rellenar	225	1.26	26.01	
exclamar	15	0.08	1.73			

Tabla 3.6: Cuentas absolutas y relativas de los alargamientos anotados en INFOTREN.

Tipo de sonido	Sonido	Cuenta absoluta	%
Vocal	/i/	118	11.58
	/e/	364	35.72
	/a/	235	23.06
	/o/	100	9.81
	/u/	1	0.10
Consonante	/l/	105	10.30
	/m/	3	0.29
	/n/	47	4.61
	/s/	46	4.51

Alargamientos. Los alargamientos se producen en su mayoría sobre vocales (80.27 % de los casos), de nuevo con una clara preferencia por la vocal /e/. También se observa un número importante de alargamientos de algunas consonantes, especialmente /l/, /m/ y /s/, que constituyen el 19.43 % del total de alargamientos (véase la Tabla 3.6). En particular, la consonante /l/ es alargada un mayor número de veces de lo que cabía esperar. En nuestra opinión, ello puede ser debido al proceso de reclutamiento de voluntarios para la base de datos, que tuvo lugar en la Universidad Politécnica de Cataluña, y que por tanto estuvo integrado en su mayor parte por hablantes del castellano en su variedad catalana. Una de las particularidades de esta variedad es precisamente un mayor énfasis en la pronunciación del sonido /l/.

Si analizamos la posición de los alargamientos dentro de las palabras (inicial, intermedia o final), se comprueba que un 8.54 % de los alargamientos son iniciales, un 12.07 % intermedios y un 67.04 % finales. Por tanto, parece clara la tendencia a alargar sonidos al final de las palabras. Por último, un 12.37 % de los casos corresponde a palabras monofonémicas, especialmente la conjunción *y* y la preposición *a*. Precisamente, si se analiza qué tipo de palabras son las que contienen alargamientos, se observa que se trata en su mayor parte de conjunciones, preposiciones, artículos y pronombres. Este hecho puede ser utilizado en la práctica para detectar alargamientos.

En la Tabla 3.7 se muestran los 20 alargamientos más frecuentes en INFOTREN, teniendo en cuenta no solamente el sonido alargado sino también la palabra y la posición dentro de la palabra. Por ejemplo, el alargamiento más frecuente en INFOTREN es el del sonido /e/ al final de la palabra "de" (el sonido alargado aparece subrayado). Nótese que los 20 casos mostrados dan cuenta de casi el 60 % de los alargamientos. De hecho, los 7 casos más frecuentes dan cuenta del 40 % de los alargamientos.

3.3.1.3. Distorsiones léxicas.

El número de distorsiones léxicas es muy pequeño en INFOTREN si lo comparamos con el número de disfluencias acústicas. Como promedio se observa alrededor de una distorsión léxica cada 100 palabras efectivas. En la mitad de los casos se trata de palabras mal pronunciadas, y en la otra mitad, de palabras cortadas (véase la Tabla 3.5). Podemos concluir, por tanto, que estos fenómenos, a diferencia de pausas habladas y alargamientos, aparecen únicamente como errores, no como recursos del habla espontánea.

Tabla 3.7: Lista de los 20 alargamientos más frecuentes en INFOTREN.

Alargamiento	Número de apariciones
de	113
el	77
y	72
para	45
a	38
el	36
que	31
qué	22
saber	18
la	16
día	16
las	15
del	15
sobre	14
quería	13
o	13
si	12
este	12
el	12
sí	11

3.3.1.4. Disfluencias sintácticas.

Se ha observado una media de 3.05 disfluencias sintácticas cada 100 palabras efectivas. Entre ellas destacan las repeticiones (53.58 % de los casos) y las sustituciones (25.87 % de los casos). Por otra parte, sólo se han anotado 70 casos de frases abandonadas, lo que representa un 12.84 % del total de disfluencias sintácticas.

Algunas repeticiones se producen en forma múltiple, es decir, la misma palabra se repite varias veces. Estos casos han sido anotados como repeticiones anidadas, aplicando un esquema de anidamiento en forma de árbol binario y de izquierda a derecha, similar al utilizado en (Shriberg, 1994). Por ejemplo, el siguiente caso:

Sí, sí, sí.

se anotaría en formato simplificado como dos repeticiones, una dentro de otra:

(rr (m Sí) (c (rr (m sí) (c sí)))) .

Sin embargo, algunos autores considerarían el ejemplo anterior como una sola repetición. En la Tabla 3.8 se muestra el número de reformulaciones simples (o no anidadas) y compuestas (o anidadas) en INFOTREN. Las repeticiones compuestas corresponden siempre —salvo en un caso— a repeticiones múltiples. Por otra parte, sólo 26 repeticiones sobre 292 (es decir, un 8.9 %) son repeticiones múltiples. De hecho, las reformulaciones compuestas no suponen más que un 13.21 % del total (72 sobre 545). Si las repeticiones compuestas se cuentan como simples, aún el número de disfluencias sintácticas es de 519, y se tienen como promedio 2.90 disfluencias sintácticas cada 100 palabras efectivas.

Tabla 3.8: Número de disfluencias sintácticas simples y compuestas en INFOTREN.

Tipo	Simples	Compuestas	Total
Repeticiones	265	27	292
Sustituciones	111	30	141
Reformulaciones con inserción	30	7	37
Reformulaciones con borrado	5	0	5
Frasas abandonadas	62	8	70

Tabla 3.9: Los 10 marcadores de discurso más frecuentes en INFOTREN representan un 65 % del total.

Expresión	Categoría	Número de apariciones
gracias	cerrar	108
por favor	pedir	80
hola	abrir	76
muchas gracias	cerrar	66
buenos días	abrir	60
pues	rellenar	38
nada más	rechazar	37
mire	rellenar	36
a ver	rellenar	31
bueno	rellenar	31

3.3.1.5. Marcadores de discurso.

Los marcadores de discurso suponen el 12.21 % de los fenómenos. Aparecen como promedio 4.84 marcadores de discurso cada 100 palabras efectivas. Como se muestra en la Tabla 3.5, las funciones más frecuentes son las de *rellenar* (26.01 % de los casos), *cerrar* (21.85 %) y *abrir* (17.34 %). En general, muy pocas expresiones representan la mayor parte de los casos. Por ejemplo, dentro de la categoría *abrir* aparecen tan sólo cuatro expresiones distintas; dentro de la categoría *cerrar* aparecen siete; y en el caso más extremo, la expresión *por favor* da cuenta del 86.96 % de los casos en la categoría *pedir*. Sólo diez expresiones distintas dan cuenta del 65 % de los casos en todas las categorías, tal como muestra la Tabla 3.9.

3.3.2. Distribución de fenómenos en CORLEC-EHU-1.

Las transcripciones ortográficas de CORLEC-EHU-1 contienen 75116 palabras, de las cuales 72182 son palabras efectivas, es decir, palabras que no están ni en el *reparandum* ni en la *señal de edición* de una reformulación, ni en frases abandonadas. En el cálculo de estos números y de todos los que siguen se han considerado las señales completas, tanto los segmentos *limpios* como los ruidosos y/o solapados.

Se han anotado 1808 solapamientos —que corresponden a 904 segmentos de señal descartados para entrenar modelos acústicos— y 249 segmentos ruidosos —también descartados para entrenar modelos acústicos—, que suponen, como ya se ha dicho, el 11.66 % de la señal. En lo que respecta a marcas especiales, se han anotado 70 segmentos no transcritos, 10 cortes en

la grabación y 473 marcas de continuación, que normalmente van asociadas a breves interrupciones que se solapan con el discurso de un hablante. Dentro de lo que llamamos distorsiones léxicas, se han anotado 36 siglas y 186 palabras extranjeras, que no se consideran fenómenos de habla espontánea. Del resto de fenómenos, los que en este trabajo sí consideramos propiamente como fenómenos de habla espontánea, se ha anotado un número total de 17579, que aparece desglosado por categorías en la Tabla 3.10.

No podemos decir que las entrevistas de CORLEC-EHU-1 muestren una mayor densidad de fenómenos que los diálogos hombre-máquina de INFOTREN. De hecho, los números hablan más bien en sentido contrario. En INFOTREN se contabilizaron 7084 fenómenos sobre 17884 palabras efectivas y 150 minutos de señal, lo que significa 39.61 fenómenos cada 100 palabras efectivas y 0.7871 fenómenos por segundo. En CORLEC-EHU-1 hemos contabilizado 17579 fenómenos sobre 72182 palabras efectivas y 23076 segundos de señal, lo que significa 24.35 fenómenos cada 100 palabras efectivas y 0.7617 fenómenos por segundo.

Si consideramos únicamente pausas habladas, alargamientos, distorsiones léxicas y disfluencias sintácticas, es decir, los fenómenos que en la literatura suelen tratarse como disfluencias, los números se equilibran. En INFOTREN contabilizamos 2792 fenómenos, lo que significa 15.61 fenómenos cada 100 palabras efectivas y 0.3102 fenómenos por segundo. En CORLEC-EHU-1 contabilizamos 9192 fenómenos, lo que significa 12.73 fenómenos cada 100 palabras efectivas y 0.3983 fenómenos por segundo.

La diferencia más notable entre una y otra base de datos se refiere a la distribución de fenómenos. El porcentaje de ruidos es muy inferior en CORLEC-EHU-1, mientras que los porcentajes de distorsiones léxicas y de disfluencias sintácticas aumentan muy notablemente. Las disfluencias acústicas mantienen en CORLEC-EHU-1 la proporción observada en INFOTREN, mientras que los marcadores de discurso aumentan ligeramente su presencia. En los siguientes párrafos analizamos cada categoría por separado.

3.3.2.1. Ruidos.

En conjunto, los ruidos representan el 20.26 % de los fenómenos, y se producen a razón de casi 5 ruidos por cada 100 palabras efectivas. Es notable la cantidad de aspiraciones anotadas (3050, del orden de una cada 25 palabras, más del 93 % de los ruidos producidos por los hablantes). En cuanto a los ruidos externos, sólo se han anotado 287, cantidad anormalmente pequeña teniendo en cuenta las condiciones ambientales de las grabaciones. Esto se explica por el hecho de que solamente estamos contabilizando ruidos externos *aislados* que suceden en segmentos por lo demás relativamente libres de ruido. Los segmentos de señal afectados por una secuencia ininterrumpida de *clicks* o por un ruido o música de fondo demasiado fuertes se anotan por entero como segmentos ruidosos y se contabilizan aparte.

3.3.2.2. Disfluencias acústicas.

Las disfluencias acústicas representan el 41.30 % de los fenómenos (10 cada 100 palabras efectivas). Destaca el gran número de alargamientos (3539, 4.98 cada 100 palabras efectivas, frente a las 5.70 de INFOTREN). Esto es debido en cierta medida a un exceso de celo por parte del anotador, ya que algunos de los alargamientos son prácticamente inapreciables y podrían haber sido absorbidos como segmentos normales. El número de alargamientos resulta significativo sobre todo si lo comparamos con el número de pausas de silencio (1863, también quizá sobredimensionado, ya que algunas de estas pausas no alcanzan los 200 milisegundos) y con

Tabla 3.10: Estadísticas de los fenómenos de habla espontánea anotados en CORLEC-EHU-1: número absoluto de apariciones (#FHE), número promedio de fenómenos por cada 100 palabras efectivas (%FHE/PE), porcentaje de fenómenos con respecto a la categoría a la que pertenecen (%FHE/C) y porcentaje que representa cada categoría con respecto al total (%C/Total).

Categoría / Subcategoría		Tipo	#FHE	%FHE/PE	%FHE/C	%C/Total
Ruidos		genérico	287	0.40	8.06	20.26
		aspiración	3050	4.23	85.65	
		labios	184	0.25	5.17	
		tos	40	0.06	1.12	
Disfluencias acústicas	Pausas de silencio	–	1863	2.58	25.66	41.30
	Pausas habladas	a	33	0.05	0.45	
		e	794	1.10	10.94	
		m	335	0.46	4.61	
	sin identificar	642	0.89	8.84		
	Alargamientos	–	3593	4.98	49.49	
Distorsiones léxicas		cortada	222	0.31	14.92	8.46
		mal pronunciada	968	1.34	65.05	
		afirmación gutural	298	0.41	20.03	
Disfluencias sintácticas	Reformulaciones	repetición	1657	2.30	71.82	13.12
		sustitución	337	0.47	14.61	
		inserción	94	0.13	4.07	
		borrado	16	0.02	0.69	
	Frasas abandonadas	–	203	0.28	8.80	
Marcadores de discurso		abrir	103	0.14	3.48	16.86
		cerrar	128	0.18	4.32	
		aceptar	728	1.01	24.57	
		rechazar	101	0.14	3.41	
		explicar/editar	234	0.32	7.90	
		pedir	224	0.31	7.56	
		rellenar	1380	1.91	46.57	
exclamar	65	0.09	2.19			

el número total de pausas habladas (1804). Esto pone de manifiesto, en este caso también, que los tres tipos de fenómenos resultan igualmente útiles como recursos del habla espontánea. Los números de la Tabla 3.10 dan un reparto de 25.66 % para pausas de silencio, 24.85 % para pausas habladas y 49.49 % para alargamientos. Esta distribución difiere de la de INFOTREN sobre todo en que los alargamientos ganan peso frente a las pausas habladas.

Pausas habladas. La distribución interna de las pausas habladas muestra que la realización /e/ es la más frecuente en CORLEC-EHU-1 (794 instancias, un 44.01 % del total), seguida de /m/ (335, un 18.57 % del total) y la realización /a/, cuya presencia es prácticamente anecdótica (33, 1.83 % del total). Con respecto a los números de INFOTREN, es significativo el aumento en el porcentaje de realizaciones no identificadas (642, un 35.59 % del total). Se ha comprobado que estas realizaciones corresponden por lo general a glotalizaciones o distorsiones de sonidos vocálicos, que suelen suceder al final de alargamientos o de pausas habladas de otro tipo, o también en determinados contextos que no permiten articular una de las realizaciones más ortodoxas.

Alargamientos. Los alargamientos de vocales representan un 79.29 % de los casos, cifra casi idéntica a la obtenida para INFOTREN. La distribución interna de los alargamientos de vocales presenta, con respecto a INFOTREN, una disminución de los sonidos /e/ y /a/ y un aumento importante del sonido /o/ (véase la Tabla 3.11). En cuanto a las consonantes, el sonido /l/, que en INFOTREN representaba un 10.30 % de los alargamientos, en este caso alcanza tan sólo un 3.17 %. Su lugar lo ocupan los sonidos /n/ y /s/, que representan el 8.79 % y el 7.96 % de los alargamientos, respectivamente. También, como novedad con respecto a INFOTREN, aparecen 22 alargamientos del sonido /rr/.

Si analizamos la posición de los alargamientos dentro de las palabras (inicial, intermedia o final), se comprueba que un 3.12 % de los alargamientos son iniciales, un 6.21 % intermedios y un 80.77 % finales. Por tanto, la tendencia a alargar sonidos al final de las palabras se incrementa con respecto a INFOTREN. Por último, un 9.91 % de los casos corresponde a palabras monofonémicas, entre las que destacan la conjunción *y* (233 casos), la preposición *a* (68 casos) y la conjunción *o* (53 casos). De hecho, al igual que sucedía en INFOTREN, la mayor parte de las palabras con alargamientos son conjunciones, preposiciones, artículos y pronombres. En la Tabla 3.12 se muestran los 20 alargamientos más frecuentes, que representan el 45.92 % de los casos, frente al 60 % que representaban los 20 alargamientos más frecuentes en INFOTREN. De hecho, CORLEC-EHU-1 contiene una gran cantidad de palabras distintas con alargamientos (1056 frente a las 208 de INFOTREN), y muchos de tales alargamientos aparecen una sola vez: 741 casos se dan en CORLEC-EHU-1 (70.17 %) frente a los 103 de INFOTREN (49.52 %). Esto es debido a que el vocabulario de INFOTREN está restringido a una tarea, mientras que el de CORLEC-EHU-1 es completamente general.

3.3.2.3. Distorsiones léxicas.

El número de distorsiones léxicas en CORLEC-EHU-1 (1488, un 8.46 % del total de fenómenos) es significativamente mayor que en INFOTREN (200, un 2.82 % del total). Destaca el gran número de *malas* pronunciaciones (968, un 65.05 % de las distorsiones léxicas). Esto es debido a las estrictas condiciones impuestas al anotador con respecto a la pronunciación de

Tabla 3.11: Cuentas absolutas y relativas de los alargamientos anotados en CORLEC-EHU-1.

Tipo de sonido	Sonido	Cuenta absoluta	%
Vocal	/i/	362	10.10
	/e/	972	27.05
	/a/	715	19.90
	/o/	756	21.04
	/u/	44	1.22
Consonante	/l/	114	3.17
	/m/	6	0.17
	/n/	316	8.79
	/r/	22	0.61
	/s/	286	7.96

Tabla 3.12: Lista de los 20 alargamientos más frecuentes en el subcorpus CORLEC-EHU-1. Los sonidos alargados aparecen subrayados.

Alargamiento	Número de apariciones
que <u>ē</u>	304
<u>y</u>	233
de <u>ē</u>	211
la <u>ā</u>	106
en <u>ē</u>	81
no <u>ō</u>	77
<u>a</u>	68
un <u>ū</u>	66
el <u>ē</u>	58
pero <u>ō</u>	57
porque <u>ē</u>	55
<u>o</u>	53
para <u>ā</u>	52
una <u>ā</u>	51
es <u>ē</u>	34
pues <u>ē</u>	33
se <u>ē</u>	29
como <u>ō</u>	29
yo <u>ō</u>	28
sí <u>ī</u>	25

Tabla 3.13: Número de disfluencias sintácticas simples y compuestas en CORLEC-EHU-1.

Tipo	Simples	Compuestas	Total
Repeticiones	1457	200	1657
Sustituciones	268	69	337
Reformulaciones con inserción	81	13	94
Reformulaciones con borrado	14	2	16
Frasas abandonadas	185	18	203

las palabras, ya que debía anotar cualquier desviación de la pronunciación *standard*; por ejemplo, *Madrí* en lugar de *Madrid*, *pasao* en lugar de *pasado*, *desir* en lugar de *decir*, etc. Estas condiciones buscan obtener modelos acústicos más ajustados, si bien ello implica una modelización explícita de las variantes de pronunciación al construir el vocabulario del reconocedor. Es destacable también el número de afirmaciones guturales (298 instancias), lo cual confirma la necesidad de modelar a nivel acústico tales fenómenos.

3.3.2.4. Disfluencias sintácticas.

Se han observado 2307 disfluencias sintácticas en CORLEC-EHU-1, lo que representa un 13.12 % del total de fenómenos, cifra que casi duplica el porcentaje de INFOTREN. Es decir, en CORLEC-EHU-1 las disfluencias sintácticas ocupan un lugar mucho más destacado dentro del conjunto de fenómenos. Sin embargo, las frecuencias con respecto al tamaño de la base de datos son muy similares: 3.20 fenómenos cada 100 palabras efectivas en CORLEC-EHU-1, y 3.05 en INFOTREN.

Destaca el gran número de repeticiones (1657, un 71.82 % de los casos). De hecho, esta gran cantidad de repeticiones parece indicar, una vez más, que aparecen no como errores sino como recursos del habla espontánea (Clark, 2002). Los hablantes, mientras planifican el discurso, pueden optar o bien por repetir una o más veces la misma palabra, o bien por insertar una pausa hablada. El objetivo sería el mismo en ambos casos: mantener el turno. El resto de disfluencias sintácticas tiene una presencia mucho menos significativa, lo cual sugiere que aparecen como consecuencia de errores que el propio hablante detecta y tiene que corregir sobre la marcha.

En la Tabla 3.13 se muestra el número de reformulaciones simples y compuestas en CORLEC-EHU-1. Las repeticiones compuestas corresponden siempre —salvo en cuatro casos— a repeticiones múltiples. Por otra parte, sólo 196 repeticiones sobre 1657 (un 11.83 %) son repeticiones múltiples. De hecho, las reformulaciones compuestas no suponen más que un 13.09 % del total (302 sobre 2307), porcentaje casi idéntico al obtenido sobre INFOTREN. Si las repeticiones múltiples se contaran como simples, el número de disfluencias sintácticas se reduciría a 2111, y se tendrían como promedio 2.92 disfluencias sintácticas cada 100 palabras efectivas (resultado que es, de nuevo, casi idéntico al obtenido para INFOTREN).

3.3.2.5. Marcadores de discurso.

Los marcadores de discurso suponen el 16.86 % de los fenómenos. Aparecen como promedio 4.10 marcadores de discurso cada 100 palabras efectivas. Como se muestra en la Tabla 3.5, la función más frecuente es la de *rellenar* (46.57 % de los casos), seguida de *aceptar* (24.57 %) y, ya más lejanas, *explicar/editar* (7.90 %) y *pedir* (7.56 %). Al igual que en INFOTREN, muy

Tabla 3.14: Marcadores de discurso en CORLEC-EHU-1 desglosados por categorías: número de expresiones distintas, número total de instancias, número de instancias correspondientes a las 5 expresiones más frecuentes, y porcentaje sobre el total que representan éstas últimas.

Categoría	#Expresiones	#Instancias	#Instancias(5)	%Instancias(5)
abrir	12	103	87	84.47
cerrar	30	128	79	61.72
aceptar	31	728	651	89.51
rechazar	11	101	95	94.06
explicar/editar	37	234	171	73.08
pedir	19	224	198	88.39
rellenar	44	1380	1152	83.48
exclamar	23	65	39	60.00

pocas expresiones representan la mayor parte de los casos. En los 2963 marcadores de discurso de CORLEC-EHU-1 sólo aparecen 207 expresiones distintas, muchas de ellas muy parecidas, como *gracias*, *muchas gracias* y *muchísimas gracias*. De hecho, las cinco expresiones más frecuentes en cada categoría dan cuenta del 83.43 % de todos los marcadores de discurso (véase la Tabla 3.14).

3.4. Dependencia del número de fenómenos con respecto a la duración de las intervenciones.

En este apartado se estudia en qué medida el número de fenómenos que aparecen en un turno depende de la duración del mismo. Se plantea como hipótesis que aparezca un mayor número de fenómenos cuanto más largo sea el turno, y que esta relación sea lineal.

3.4.1. Dependencia del número de fenómenos con respecto a la duración en INFOTREN.

En primer lugar se ha contado el número de fenómenos en cada uno de los 1657 turnos de usuario de INFOTREN. Tal como muestra la Tabla 3.15, por ejemplo, el número promedio de disfluencias acústicas en cada turno de usuario es de 1.69, aunque la desviación típica es relativamente alta (2.52), lo que significa que hay una variabilidad importante. De hecho, el número máximo de disfluencias acústicas en un turno es de 32.

Nuestro interés se centra, sin embargo, en estudiar hasta qué punto esta variabilidad es debida a la variabilidad en la duración de los turnos, es decir, hasta qué punto el número de fenómenos en un turno depende de la duración del mismo. Para ello se calcula el coeficiente de correlación de Pearson entre el número de fenómenos y la duración de cada turno (véase Tabla 3.16). Recuérdese que el coeficiente de correlación $\rho(x, y)$ entre dos variables es una medida normalizada de en qué medida el valor de la variable x se puede predecir en función del valor de la variable y , suponiendo que entre ambas existe una relación lineal. Si $\rho(x, y) = 0$, las dos variables estarían completamente decorreladas, es decir, serían independientes. Si $|\rho(x, y)| = 1$, entonces se podría establecer una relación lineal exacta entre los valores de x e y . Aunque no

Tabla 3.15: Media, desviación típica y valor máximo del número de fenómenos de habla espontánea en los 1657 turnos de usuario de INFOTREN.

Categoría	Media	Desviación típica	Máximo
Ruidos	1.61	1.67	23
Disfluencias acústicas	1.69	2.52	32
Distorsiones léxicas	0.12	0.40	3
Disfluencias sintácticas	0.33	0.83	9
Marcadores de discurso	0.52	0.93	8

Tabla 3.16: Coeficientes de correlación de Pearson para el número de fenómenos en un turno con respecto a la duración de dicho turno, calculados sobre los 1657 turnos de INFOTREN (two-tailed t-test, $p=0.0000$). R: ruidos, A: disfluencias acústicas, L: distorsiones léxicas, S: disfluencias sintácticas, D: marcadores de discurso, t: duración del turno.

$\rho(R, t)$	$\rho(A, t)$	$\rho(L, t)$	$\rho(S, t)$	$\rho(D, t)$
0.6867	0.8606	0.2844	0.6288	0.4035

deben extraerse conclusiones definitivas del análisis de correlación, ya que una alta correlación entre dos variables no implica forzosamente una dependencia entre ellas, el cuadrado del coeficiente de correlación de Pearson, $\rho^2(x, y)$, suele interpretarse como la fracción de varianza de x que se explica con la varianza de y . El número de disfluencias acústicas muestra una clara correlación con la duración del turno. Esto puede ser debido a que estos fenómenos son recursos del habla espontánea que los hablantes utilizan con regularidad. Por otra parte, los ruidos muestran una correlación con la duración más baja de lo que cabía esperar, si tenemos en cuenta que su distribución en el discurso debería ser más o menos aleatoria. Sin embargo, los ruidos producidos por los hablantes no se distribuyen de forma aleatoria, sino que tienden a aparecer al comienzo de los turnos y alrededor de las pausas, de manera que un turno de un segundo puede contener el mismo número de ruidos que otro de diez segundos. El mismo razonamiento se puede aplicar para explicar la relativamente baja correlación con la duración de los marcadores de discurso: muchos de ellos son expresiones de aceptación o rechazo, de bienvenida o despedida, que aparecen en turnos muy cortos. Para explicar la baja correlación con la duración de las distorsiones léxicas aportamos otro razonamiento: su baja frecuencia. Se trata de fenómenos que aparecen raramente, y cuando aparecen, lo pueden hacer en turnos largos o cortos, indistintamente. Finalmente, con respecto a las disfluencias sintácticas, hay que tener en cuenta que un 61.47 % de las reformulaciones son repeticiones, las cuales muestran cierta tendencia a aparecer en turnos cortos, como recursos enfatizadores, en respuestas como *Sí, sí, sí* o *No, a Bilbao, a Bilbao*. El resto de reformulaciones y también las frases abandonadas, cuya aparición responde siempre a la detección de un error por parte del hablante, sí tienden a aparecer en turnos largos. De ahí que la correlación con la duración del turno sea moderadamente alta. Sin embargo, habrá que buscar otras fuentes de variabilidad, como la identidad del hablante, que exploraremos en el apartado 3.5.

Tabla 3.17: Media, desviación típica y valor máximo del número de fenómenos de habla espontánea en los 2873 turnos de CORLEC-EHU-1.

Categoría	Media	Desviación típica	Máximo
Ruidos	1.24	2.26	20
Disfluencias acústicas	2.53	3.99	31
Distorsiones léxicas	0.60	1.32	16
Disfluencias sintácticas	0.80	1.53	21
Marcadores de discurso	1.03	1.33	10

Tabla 3.18: Coeficientes de correlación de Pearson para el número de fenómenos en un turno con respecto al número efectivo de palabras de dicho turno, calculados sobre los 2873 turnos de CORLEC-EHU-1. R: ruidos, A: disfluencias acústicas, L: distorsiones léxicas, S: disfluencias sintácticas, D: marcadores de discurso, PE: número efectivo de palabras.

$\rho(R, PE)$	$\rho(A, PE)$	$\rho(L, PE)$	$\rho(S, PE)$	$\rho(D, PE)$
0.8605	0.7934	0.4352	0.5272	0.4428

3.4.2. Dependencia del número de fenómenos con respecto a la duración en CORLEC-EHU-1.

Tal como se muestra en la Tabla 3.17, salvo en el caso de los ruidos, el número promedio de fenómenos en cada turno de CORLEC-EHU-1 es claramente superior al observado en INFOTREN (Tabla 3.15). Esto es debido en gran medida a que los turnos son más largos. Las desviaciones típicas son también mayores en todos los casos. Sin embargo, los valores máximos no difieren mucho de los de INFOTREN, salvo en el caso de las distorsiones léxicas. De hecho, el número de fenómenos en esta categoría experimenta un aumento espectacular con respecto a INFOTREN, debido sobre todo al gran número de palabras *mal pronunciadas*. Independientemente de que dichas anotaciones puedan ser muy estrictas, lo que se pone de manifiesto con ello es una mayor relajación en la articulación de los sonidos.

Para comprobar hasta qué punto el número de fenómenos en un turno depende únicamente de la duración del mismo, se ha calculado el coeficiente de correlación entre el número de fenómenos y el número de palabras efectivas de cada turno, en las cinco categorías habituales (véase Tabla 3.18). No se ha podido utilizar el valor de la duración, porque sólo se dispone de ese dato para los turnos cortados (sin segmentos ruidosos y/o solapados). Como se verá en el apartado siguiente, la relación entre el número efectivo de palabras y la duración de un turno es lineal, por lo que a efectos prácticos es indiferente tomar uno y otro valor como variable independiente. En INFOTREN la longitud media de los turnos de usuario es de 10.79 palabras efectivas, con una desviación típica de 10.14, y el rango de valores es [1,100]. En CORLEC-EHU-1 los números son mucho mayores. La longitud media de un turno es de 25.12 palabras efectivas, con una desviación típica de 33.72, y el rango de valores es [0,226]. Tanto el número de ruidos como el de disfluencias acústicas muestran una clara correlación con la longitud del turno. En el caso de los ruidos, esto se debe al predominio de las aspiraciones (85.65 % de todos los ruidos), que se producen a intervalos fijos, como pausas técnicas que el hablante se ve obligado a realizar para tomar aire. En INFOTREN las aspiraciones y los chasquidos de labios —que en esa base de datos suelen aparecer simultáneamente— suponen también un 75 % de los ruidos, pero la

longitud de los turnos es mucho menor, y la periodicidad no llega a manifestarse significativamente, sino que más bien pesan las aspiraciones y chasquidos iniciales, que se producen tanto en turnos cortos como en turnos largos. De ahí que la correlación entre el número de ruidos y la longitud de los turnos sea menor. En el caso de las disfluencias acústicas, la explicación es la misma que para INFOTREN: se trata de recursos del habla espontánea que los hablantes utilizan con regularidad. Por último, es significativo el aumento de la correlación correspondiente a las distorsiones léxicas. Este aumento sólo puede ser debido a que la frecuencia de estos fenómenos es mucho mayor en CORLEC-EHU-1 que en INFOTREN. El resto de fenómenos muestra una correlación positiva pero más pequeña, similar a la obtenida para INFOTREN, y se puede explicar con los mismos argumentos aportados anteriormente.

3.4.3. Comparación con trabajos previos realizados sobre bases de datos en inglés.

En (Oviatt, 1995) se muestra cómo el número promedio de fenómenos (en realidad, sólo se consideran disfluencias sintácticas y pausas habladas) se puede expresar como una función lineal de la duración del turno, con una correlación considerable ($\rho = 0,77$). En (Shriberg, 1994) se muestra cómo la probabilidad de que una intervención sea fluida —es decir, no contenga ninguna disfluencia— se puede escribir como una función que decrece exponencialmente con la duración del turno y que depende de un parámetro que se interpreta como *tasa de fluencia*. Este parámetro puede ser distinto para cada base de datos, y depende sobre todo —siempre según (Shriberg, 1994)— de que el diálogo se produzca entre personas o entre una persona y una máquina. Ambos trabajos manejan bases de datos en habla inglesa, y utilizan el número efectivo de palabras como medida de la longitud del turno. Los resultados se obtuvieron sobre intervenciones relativamente cortas, formadas por frases sencillas de menos de 20 palabras efectivas.

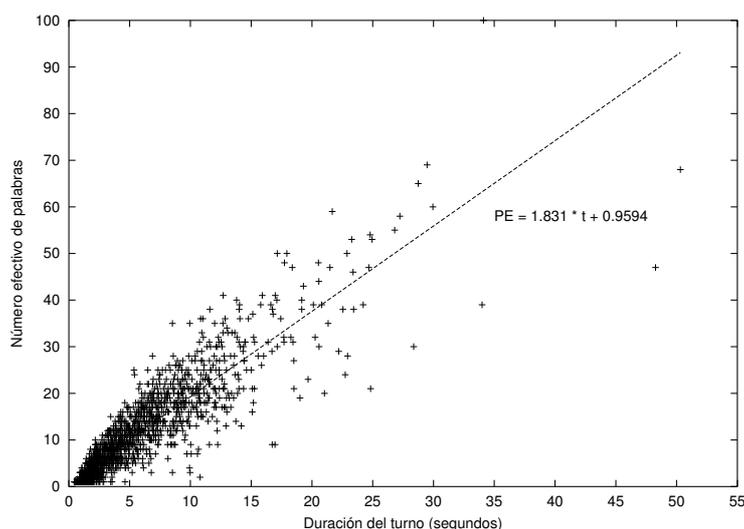


Figura 3.5: Número efectivo de palabras (PE) en los turnos de INFOTREN con respecto a la duración de los mismos (t , en segundos). Un análisis por regresión lineal resulta en el siguiente ajuste: $PE = 1.831 * t + 0.9594$ (two tailed F-test, $F(1,1655)=6616.07$, $p=0.0000$).

En el estudio de INFOTREN realizado hasta este punto se ha utilizado la duración (en

segundos) como medida de la longitud del turno. Sin embargo, tanto el número total como el número efectivo de palabras muestran una alta correlación con la duración de los turnos. En INFOTREN se obtienen coeficientes de correlación de $\rho = 0,9167$ (two tailed t-test, $t(1655)=93.36$, $p=0.0000$) y $\rho = 0,8944$ (two tailed t-test, $t(1655)=81.34$, $p=0.0000$), respectivamente (véase la Figura 3.5). Por otra parte, tanto en INFOTREN como en CORLEC-EHU-1 se han considerado turnos completos (con una o más frases), en lugar de frases sencillas, y las tasas de fenómenos no están normalizadas por la longitud de los turnos.

Para comparar los resultados de este trabajo con los de (Oviatt, 1995), se ha establecido el número efectivo de palabras como variable independiente, y la suma de disfluencias acústicas y disfluencias sintácticas, normalizada por el número de palabras efectivas, que denominaremos *tasa de disfluencias por palabra*, como variable dependiente.

En el caso de INFOTREN se ha obtenido una correlación casi nula entre ambas variables ($\rho = 0,0570$, two tailed t-test, $t(1655)=2.3217$, $p=0.0204$). De hecho, tal como muestra la Figura 3.6, el número de disfluencias acústicas y sintácticas en cada turno de INFOTREN no muestra en absoluto una dependencia lineal con el número efectivo de palabras, y un análisis por regresión lineal resulta en una pendiente sólo ligeramente positiva.

En el caso de CORLEC-EHU-1 se ha obtenido una correlación ligeramente negativa ($\rho = -0,0359$, two tailed t-test, $t(2807)=-1.902$, $p=0.0572$). Tal como muestra la Figura 3.7, la tasa de disfluencias por palabra no parece depender linealmente del número efectivo de palabras, y del análisis por regresión lineal resulta una recta de pendiente ligerísimamente negativa. Por tanto, ambos resultados contradicen las conclusiones de Oviatt. Al menos en lo que respecta al habla espontánea en castellano, el número promedio de disfluencias no aumenta con la longitud del turno, como concluye Oviatt, sino que más bien es una constante, probablemente distinta para cada hablante.

Para comparar los resultados de este trabajo con los de (Shriberg, 1994), en primer lugar se define *turno fluido* como aquél que no contiene ni disfluencias acústicas ni disfluencias sintácticas; en segundo lugar, la longitud de los turnos la mediremos en términos del número efectivo de palabras; y por último, la probabilidad de que un turno de longitud n sea fluido se define como el número de turnos fluidos de longitud n dividido por el número total de turnos de longitud n .

Para ambas bases de datos se obtiene idéntico resultado: el logaritmo de la probabilidad de que un turno sea fluido parece depender linealmente de la longitud del turno, confirmando los resultados de Shriberg para habla espontánea en lengua inglesa. Esta relación lineal parece romperse cuando el número efectivo de palabras es mayor que 20 (véanse las Figuras 3.8 y 3.9). En realidad, lo que sucede es que o bien no se encuentran turnos con esa longitud, o bien se encuentran tan pocos que la estimación de la probabilidad no es fiable. De ahí que el análisis que sigue se efectúe sólo para turnos de longitudes entre 0 y 25 palabras efectivas.

En el caso de INFOTREN la correlación entre el logaritmo de la probabilidad de que un turno sea fluido y el número efectivo de palabras es $\rho = -0,9119$ (two tailed t-test $t(23)=10,65$, $p=0,0000$). Del análisis por regresión lineal obtenemos que dicha probabilidad decae exponencialmente con la longitud del turno, según la siguiente expresión: $Prob_{fluido}(n) = Cb^n$, donde $C = 0,9296$ y la *tasa de fluencia por palabra* es $b = 0,8838$. De aquí deducimos que la tasa de disfluencias por palabra, es decir, la probabilidad de que se produzca una disfluencia tras una palabra es, en el caso de INFOTREN, de $1 - b = 0,1162$.

En el caso de CORLEC-EHU-1 la correlación entre el logaritmo de la probabilidad de que un turno sea fluido y el número efectivo de palabras es $\rho = -0,9205$ (two tailed t-test $t(23)=-11,3$, $p=0,0000$), y del análisis por regresión lineal se obtiene $Prob_{fluido}(n) = Cb^n$,

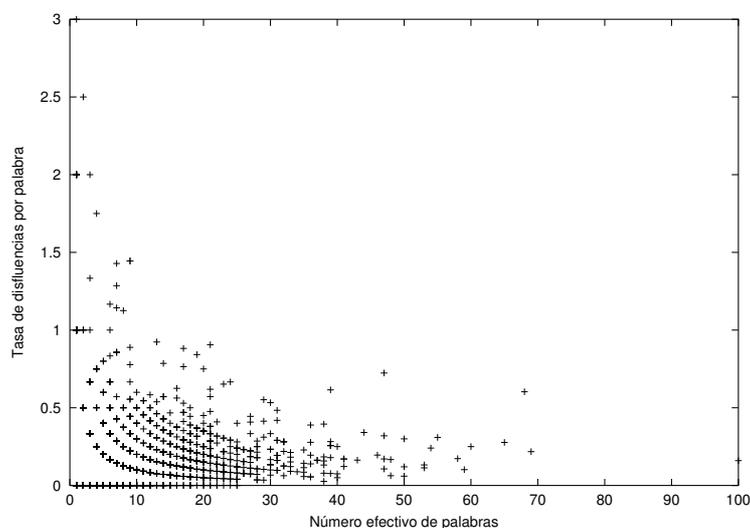


Figura 3.6: Número de disfluencias acústicas y sintácticas (D), normalizado por el número efectivo de palabras (PE), en función del número efectivo de palabras en cada turno de INFOTREN. Análisis por regresión lineal: $D/PE = 0.001509 * PE + 0.156375$ (two tailed F-test, $F(1,1655)=5.3904$, $p=0.0204$).

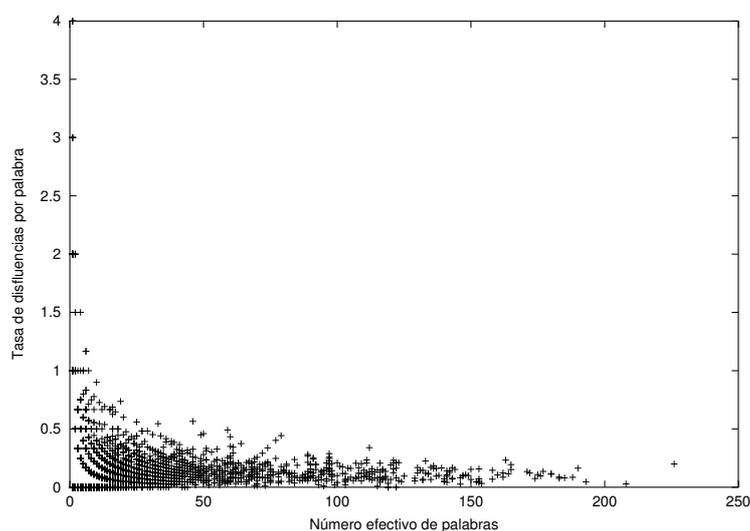


Figura 3.7: Número de disfluencias acústicas y sintácticas (D), normalizado por el número efectivo de palabras (PE), en función del número efectivo de palabras en cada turno de CORLEC-EHU-1. Análisis por regresión lineal: $D/PE = -0.0003222 * PE + 0.154957$ (two tailed F-test, $F(1,2807)=3.6188$, $p=0.0572$).

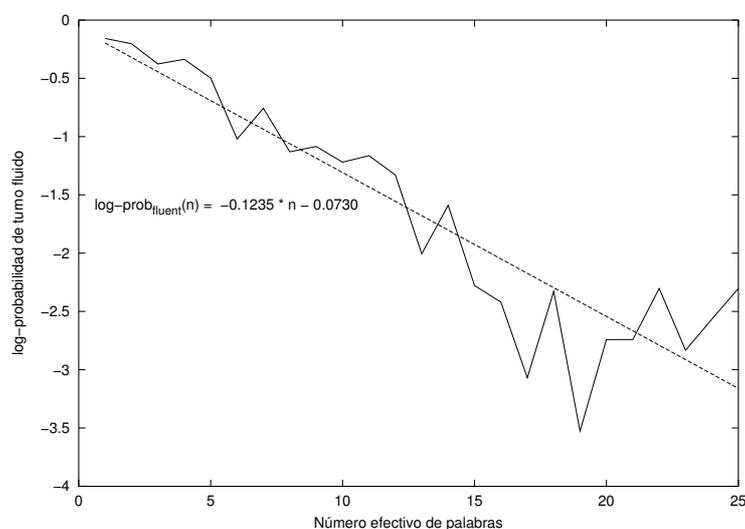


Figura 3.8: Logaritmo de la probabilidad de que un turno sea fluido en función de la longitud del turno, en el caso de INFOTREN. Ajuste por regresión lineal: $\log Prob_{fluent}(n) = -0,1235 * n - 0,0730$ (two tailed F-test, $F(1,23)=113.51$, $p=0.0000$).

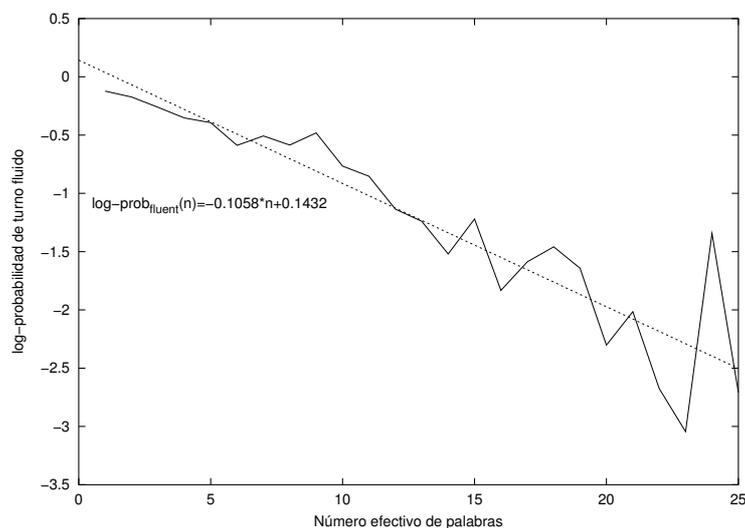


Figura 3.9: Logaritmo de la probabilidad de que un turno sea fluido en función de la longitud del turno, en el caso de CORLEC-EHU-1. Ajuste por regresión lineal: $\log Prob_{fluent}(n) = -0,1058 * n + 0,1432$ (two tailed F-test, $F(1,23)=127.7135$, $p=0.0000$).

Tabla 3.19: Media, desviación típica y valor máximo del número de fenómenos para los 75 hablantes de INFOTREN, en las 5 categorías habituales.

Categoría	Media	Desviación típica	Máximo
Ruidos	35.65	26.15	160
Disfluencias acústicas	37.33	31.31	203
Distorsiones léxicas	2.67	3.44	17
Disfluencias sintácticas	7.27	9.80	54
Marcadores de discurso	11.53	10.03	56

donde $C = 1,1540$ y la *tasa de fluencia por palabra* es $b = 0,8996$. Así pues, en CORLEC-EHU-1 la probabilidad de que se produzca una disfluencia tras una palabra es $1 - b = 0,1004$, ligeramente inferior a la obtenida para INFOTREN.

Los valores de la *tasa de disfluencias* obtenidos para INFOTREN y CORLEC-EHU-1 son, respectivamente, más altos que los obtenidos por Shriberg para ATIS ($1 - b = 0,01$) —que, al igual que INFOTREN, es una base de datos de diálogos hombre-máquina— y Switchboard ($1 - b = 0,055$) —que, como CORLEC-EHU-1, contiene diálogos naturales entre personas. Esto se explica en gran medida por el hecho de que, a diferencia de (Shriberg, 1994), en este trabajo las pausas de silencio y los alargamientos se cuentan como disfluencias.

3.5. Dependencia del número de fenómenos con respecto a la identidad del hablante.

En este apartado trataremos de determinar hasta qué punto la presencia de fenómenos de habla espontánea está ligada a la identidad del hablante. A priori, parece claro que unos hablantes podrían tener más problemas de planificación del discurso que otros, o ser más sensibles a cualquier tipo de presión, sea cognitiva o interactiva. Esta es nuestra hipótesis de partida.

3.5.1. Dependencia del número de fenómenos con respecto al hablante en INFOTREN.

Se ha contado el número de fenómenos para cada hablante de INFOTREN, en las cinco categorías consideradas en la Tabla 3.5: ruidos, disfluencias acústicas, distorsiones léxicas, disfluencias sintácticas y marcadores de discurso. Como ya se ha dicho, cada hablante ha llevado a cabo tres diálogos con el sistema automatizado de información, por lo que en nuestra opinión disponemos de datos suficientes como para evaluar su tendencia o no a insertar fenómenos de habla espontánea. En la Tabla 3.19 se muestran la media, la desviación típica y el valor máximo del número de fenómenos para los 75 voluntarios que han participado en INFOTREN.

Las desviaciones típicas son muy altas si las comparamos con los valores medios. También se puede observar que los valores máximos son cinco veces o más los valores medios. Esto pone de manifiesto la gran variabilidad que presenta el número de fenómenos en el conjunto de hablantes. Así, por ejemplo, de las distribuciones de disfluencias acústicas (Figura 3.10) y disfluencias sintácticas (Figura 3.11) podemos deducir que unos pocos hablantes acumulan un gran número de fenómenos, mientras que otros, muchos más, apenas suman unos pocos.

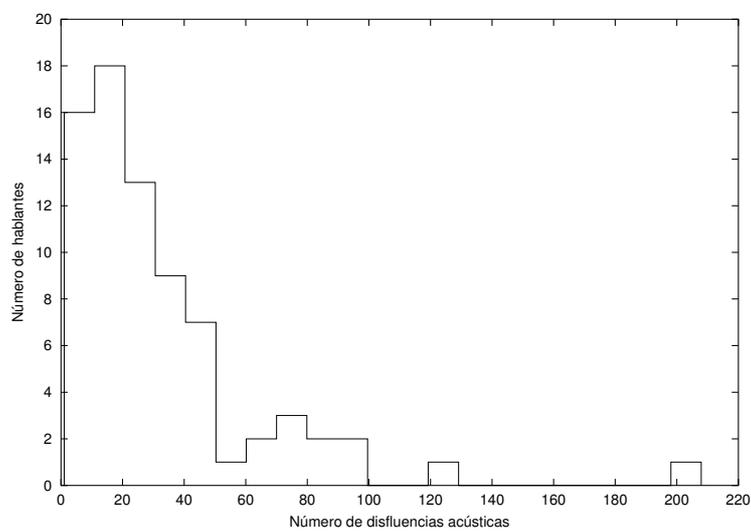


Figura 3.10: Histograma de los hablantes de INFOTREN con respecto al número de disfluencias acústicas.

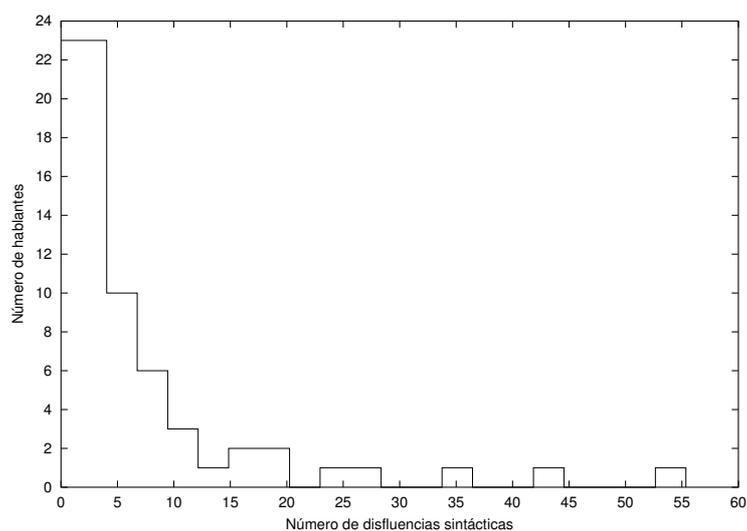


Figura 3.11: Histograma de los hablantes de INFOTREN con respecto al número de disfluencias sintácticas.

Tabla 3.20: Coeficientes de correlación entre el número de fenómenos y el tiempo de elocución, calculados sobre el conjunto de hablantes de INFOTREN (two tailed t-test, $p=0.0000$). R: ruidos, A: disfluencias acústicas, L: distorsiones léxicas, S: disfluencias sintácticas, D: marcadores de discurso, t: tiempo total de elocución de un hablante.

$\rho(R, t)$	$\rho(A, t)$	$\rho(L, t)$	$\rho(S, t)$	$\rho(D, t)$
0.8929	0.8743	0.5934	0.8138	0.6002

Esta variabilidad podría provenir simplemente del hecho de que unos hablantes han llevado a cabo diálogos mucho más largos que otros. Es decir, los hablantes con un mayor número de fenómenos podrían ser simplemente *aquéllos que hablan durante más tiempo*. Para comprobarlo, calculamos el coeficiente de correlación entre el número de fenómenos y el tiempo de elocución de un hablante, para los 75 hablantes de INFOTREN. El resultado se muestra en la Tabla 3.20. La correlación es relativamente alta en el caso de ruidos, disfluencias acústicas y disfluencias sintácticas, y más baja en el caso de distorsiones léxicas y marcadores de discurso, confirmando los resultados presentados en el apartado 3.4.

Así pues, una vez más se confirma que el número de fenómenos observados depende fuertemente del tiempo de elocución. No obstante, trataremos de demostrar cómo algunos hablantes muestran una mayor tendencia a insertar fenómenos de habla espontánea. Para ello, considere la Tabla 3.21, en la que se han seleccionado 6 de los 75 hablantes de INFOTREN. Los hablantes con índices 9 y 11 presentan tiempos de elocución mucho mayores que la media, y también un gran número de fenómenos, aunque si examinamos los valores normalizados vemos que en el hablante 9 se dan 0.74 fenómenos por segundo, por debajo de la media, que es 0.80, mientras que el hablante 11 alcanza 1.05 fenómenos por segundo, claramente por encima de la media. Los hablantes 6 y 56 presentan tiempos de elocución muy pequeños y también un escaso número de fenómenos, por debajo de la media, pero los valores normalizados son radicalmente distintos (0.50 y 1.16). Finalmente, los hablantes 26 y 27 presentan tiempos de elocución muy similares y próximos a la media de INFOTREN, pero una vez más uno de ellos presenta un número de fenómenos muy superior. Estos casos confirman en parte las conclusiones extraídas de las correlaciones, pero vemos que el número de fenómenos varía mucho de unos hablantes a otros, aún cuando sus tiempos de elocución sean similares. Podemos concluir que, aunque la tendencia general es encontrar un número de fenómenos directamente proporcional al tiempo de elocución, algunos hablantes son intrínsecamente más *disfluentes* que otros. Todo esto podría justificar la utilización de mecanismos de adaptación que permitieran personalizar hasta cierto punto los modelos acústicos y/o el modelo de lenguaje del reconocedor en un sistema de diálogo hombre-máquina.

3.5.2. Dependencia del número de fenómenos con respecto al hablante en CORLEC-EHU-1.

Mientras que en INFOTREN todos los hablantes llevan a cabo tres diálogos completos con un sistema de información, lo cual garantiza un tiempo mínimo de elocución, los hablantes que participan en las entrevistas de CORLEC-EHU-1 se implican en ellas de manera muy distinta. Así, algunos de ellos apenas dicen unas frases mientras que otros muestran una gran locuacidad. Para obtener unas estadísticas fiables se van a considerar sólo aquellos hablantes que han producido más de 120 palabras efectivas. Según la relación entre número efectivo de palabras

Tabla 3.21: Tiempo de elocución (t) y número de fenómenos de habla espontánea para 6 hablantes especialmente seleccionados de INFOTREN. Se muestran asimismo el número de fenómenos por unidad de tiempo para cada hablante, así como la media y la desviación típica de las cuentas de fenómenos, para una mejor evaluación de los datos.

Hablante	t (segundos)	R	A	L	S	D	Total	Total/t
9	378.86	124	88	2	12	56	282	0.74
11	394.45	135	203	4	54	17	413	1.05
6	51.91	13	9	0	0	4	26	0.50
56	55.37	27	23	1	3	10	64	1.16
26	138.06	27	22	0	3	28	80	0.58
27	144.34	55	49	4	10	24	142	0.98
Media	118.63	35.65	37.33	2.67	7.27	11.53	94.45	0.80
Desviación típica	76.16	26.15	31.31	3.44	9.80	10.03	70.49	–

Tabla 3.22: Media, desviación típica y valor máximo del número de fenómenos en las 5 categorías habituales, para los 98 hablantes de CORLEC-EHU-1 que suman más de 120 palabras efectivas.

Categoría	Media	Desviación típica	Máximo
Ruidos	36.10	35.28	227
Disfluencias acústicas	73.49	62.99	380
Distorsiones léxicas	17.23	22.12	126
Disfluencias sintácticas	23.40	20.99	114
Marcadores de discurso	29.89	22.97	106

y tiempo de elocución estimada para INFOTREN (véase Figura 3.5), 120 palabras efectivas corresponden aproximadamente a un minuto de señal, tiempo que consideramos suficiente para evaluar la tendencia de un hablante a insertar fenómenos de habla espontánea.

De los 118 hablantes que participan en CORLEC-EHU-1, 98 han producido más de 120 palabras efectivas. Para cada uno de ellos se ha contado el número de fenómenos en las cinco categorías habituales. En la Tabla 3.22 se muestran la media, la desviación típica y el valor máximo del número de fenómenos para esos 98 hablantes.

Al igual que en INFOTREN, estos datos ponen de manifiesto la gran variabilidad que presenta el número de fenómenos en el conjunto de hablantes. Esta desigual distribución de fenómenos se puede observar con claridad en los histogramas de hablantes con respecto al número de disfluencias acústicas y sintácticas (Figuras 3.12 y 3.13).

Para comprobar hasta qué punto esta variabilidad proviene simplemente de la variabilidad en los tiempos de elocución de los hablantes, se ha calculado el coeficiente de correlación entre el número de fenómenos y el número efectivo de palabras de cada hablante, para los 98 hablantes *estadísticamente útiles* de CORLEC-EHU-1 (véase la Tabla 3.23). La correlación es relativamente alta en el caso de ruidos y disfluencias acústicas, con valores casi idénticos a los de INFOTREN. Sin embargo, el número de distorsiones léxicas muestra en CORLEC-EHU-1 una menor correlación con la longitud del turno que en INFOTREN. Lo mismo sucede con las disfluencias sintácticas. Esto podría indicar que en la producción de este tipo de fenómenos las

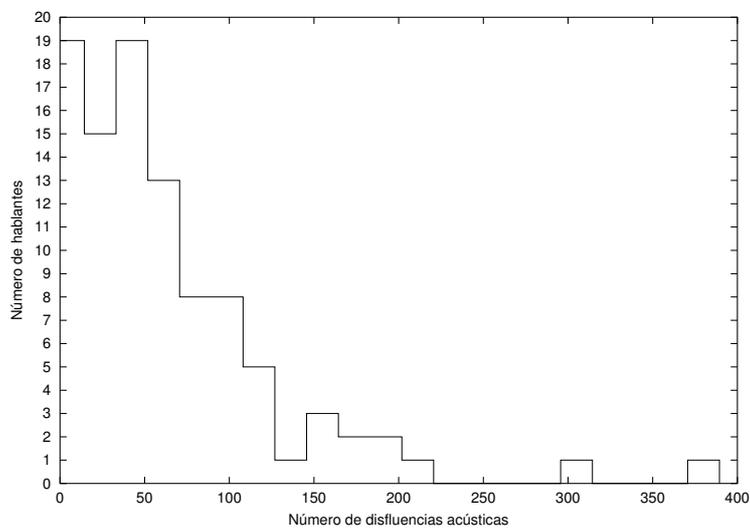


Figura 3.12: Histograma de los hablantes de CORLEC-EHU-1 con respecto al número de disfluencias acústicas.

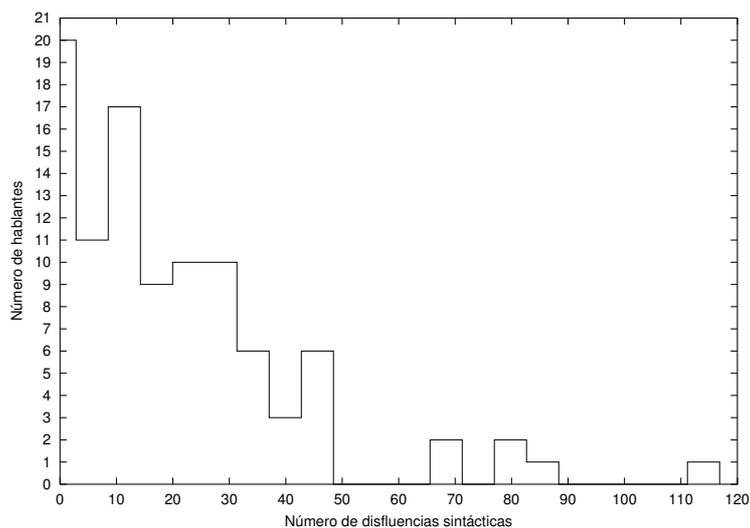


Figura 3.13: Histograma de los hablantes de CORLEC-EHU-1 con respecto al número de disfluencias sintácticas.

Tabla 3.23: Coeficientes de correlación entre el número de fenómenos y el número efectivo de palabras de cada hablante, calculados sobre el conjunto de 98 hablantes estadísticamente útiles de CORLEC-EHU-1 (two tailed t-test, $p=0.0000$). R: ruidos, A: disfluencias acústicas, L: distorsiones léxicas, S: disfluencias sintácticas, D: marcadores de discurso, PE: número efectivo de palabras.

$\rho(R, PE)$	$\rho(A, PE)$	$\rho(L, PE)$	$\rho(S, PE)$	$\rho(D, PE)$
0.8952	0.8329	0.4506	0.6462	0.6506

Tabla 3.24: Número efectivo de palabras (PE) y número de fenómenos de habla espontánea para 6 hablantes especialmente seleccionados de CORLEC-EHU-1. Se muestran asimismo el número de fenómenos total (FHE) y el número de fenómenos por palabra efectiva (FHE/PE) para cada hablante, así como la media y la desviación típica de cada columna, para una mejor evaluación de los datos.

Hablante	PE	R	A	L	S	D	FHE	FHE/PE
bent026f.H2	156	6	13	17	21	51	108	0.6923
aent033a.H8	159	9	9	2	3	1	24	0.1509
cent001a.H2	649	13	122	17	50	50	252	0.3883
aent001e.H2	669	28	26	4	4	5	67	0.1001
aent007a.H2	1131	43	317	15	114	106	595	0.5261
cent002d.H3	1109	67	40	10	2	12	131	0.1181
Media	729.94	36.10	73.49	17.23	23.40	29.89	180.11	0.2504
Desviación típica	523.34	35.28	62.99	22.12	20.99	22.97	136.50	0.0872

características propias del hablante tienen más peso en CORLEC-EHU-1 que en INFOTREN. Finalmente, la correlación del número de marcadores de discurso con la longitud del turno es mayor en CORLEC-EHU-1 que en INFOTREN, aunque la diferencia es pequeña. Este aumento puede deberse al predominio de la categoría *rellenar*, que supone el 46 % de los marcadores de discurso (frente al 26.01 % de INFOTREN). Este tipo de marcadores se producen de manera regular en el discurso de los hablantes y, por tanto, deben mostrar una alta correlación con el tiempo de elocución, mientras que otras categorías, como *abrir*, *cerrar* o *aceptar*, que predominan en INFOTREN, tienen una mayor presencia en turnos cortos.

Los datos de correlación indican que las características propias de los hablantes pesan más en CORLEC-EHU-1 que en INFOTREN. De hecho, si calculamos la correlación entre el número total de fenómenos y el tiempo de elocución total de cada hablante —o el número total de palabras efectivas, en el caso de CORLEC-EHU-1—, obtenemos $\rho = 0,9472$ para INFOTREN y $\rho = 0,8976$ para CORLEC-EHU-1. Así pues la existencia de hablantes intrínsecamente más disfluentes que otros debería ser aún más notable en diálogos naturales entre personas que en diálogos hombre-máquina. Para ilustrar más claramente esta circunstancia, en la Tabla 3.24 se muestran 6 casos concretos, organizados por pares de hablantes con un número similar de palabras efectivas pero con tasas de disfluencias muy distintas.

3.6. Evaluación del uso de fenómenos acústicos, léxicos y pragmáticos en la detección de disfluencias sintácticas.

Las disfluencias sintácticas constituyen uno de los problemas más importantes tanto para el reconocimiento como para la comprensión del habla espontánea. Ello es debido a la ruptura local del contexto sintáctico, que no se ha tenido en cuenta en el modelo de lenguaje y que puede introducir ambigüedades o incluso confusiones graves en la comprensión del enunciado. Como varios autores han sugerido (Nakatani and Hirschberg, 1994; Shriberg et~al., 1997), con objeto de integrar las disfluencias sintácticas en un sistema de reconocimiento y comprensión del habla, debemos ser capaces de reconocerlas utilizando no sólo información léxica y sintáctica, sino también informaciones acústicas, prosódicas y pragmáticas. Las primeras aproximaciones se concentraron en mejorar el modelo de lenguaje únicamente a partir de texto, utilizando modelos explícitos para las disfluencias sintácticas y para eventos acústicos como ruidos, pausas habladas, etc. (Rosenfeld et~al., 1995; Stolcke and Shriberg, 1996). Estas aproximaciones consiguieron reducir la perplejidad del modelo de lenguaje, pero no tanto las tasas de error. Otros trabajos más recientes han optado bien por enriquecer el modelo de lenguaje con marcas de función sintáctica, bordes de entonación y disfluencias sintácticas (Heeman, 1999), bien por combinar información acústica y prosódica con un modelo de lenguaje más complejo (Stolcke et~al., 1999). En ambos casos se han obtenido prometedoras —aunque pequeñas— mejoras en las tasas de reconocimiento.

Con el fin de asentar sobre una base sólida cualquier estrategia que se adopte en el futuro con respecto a las disfluencias sintácticas, en este apartado trataremos de evaluar hasta qué punto su detección podría estar basada en la presencia de fenómenos acústicos, distorsiones léxicas y/o marcadores de discurso.

3.6.1. Fenómenos acústicos y disfluencias sintácticas.

3.6.1.1. Resultados sobre INFOTREN.

De las 545 disfluencias sintácticas de INFOTREN, 358 (un 65.69 %) contienen fenómenos acústicos (es decir, ruidos o disfluencias acústicas). Por otra parte, de los 5474 fenómenos acústicos de INFOTREN sólo 541 (un 9.88 %) aparecen dentro de disfluencias sintácticas, con la siguiente distribución: 84 ruidos (15.53 %), 288 alargamientos (53.23 %), 76 pausas de silencio (14.05 %) y 93 pausas habladas (17.19 %). La distribución detallada de fenómenos acústicos dentro de disfluencias sintácticas se muestra en la Tabla 3.25. Nótese que los fenómenos acústicos pueden aparecer, por ejemplo, en el *reparandum* de una reformulación que a su vez está anidada en la *corrección* de otra reformulación. Estos fenómenos se contarían dos veces, de ahí que los números de la Tabla 3.25 no se correspondan con los de tablas anteriores. Aún así, creemos que resultan de extrema utilidad para saber qué fenómenos y dónde aparecen dentro de las disfluencias sintácticas.

La primera conclusión que podemos extraer de la Tabla 3.25 es que las pausas de silencio, las pausas habladas y los ruidos (en particular, aspiraciones producidas por el hablante), cuando aparecen dentro de disfluencias sintácticas, lo hacen casi exclusivamente como *señales de edición* en reformulaciones. Por otra parte, la frecuencia relativa de los alargamientos dentro de las disfluencias sintácticas es mucho mayor que fuera de ellas (véase la Tabla 3.5). Esto sugiere que precisamente los alargamientos pueden constituir una pista importante en la detección de

Tabla 3.25: Distribución de fenómenos acústicos dentro de las disfluencias sintácticas en INFOTREN. R: ruidos, S: pausas de silencio, H: pausas habladas, A: alargamientos.

	Reparandum				Señal de edición				Corrección			
	R	S	H	A	R	S	H	A	R	S	H	A
repeticiones	5	0	0	153	42	37	52	0	3	0	0	15
sustituciones	8	6	4	67	12	25	32	0	9	7	3	25
reformulaciones con inserción	1	1	0	19	4	2	6	0	3	3	1	6
reformulaciones con borrado	3	2	0	3	0	0	1	0	0	0	0	1
frases abandonadas	6	5	3	47	-				-			
Totales	23	14	7	289	58	64	91	0	15	10	4	47

disfluencias sintácticas. De hecho, más de la mitad de las disfluencias sintácticas de INFOTREN contienen alargamientos. Es de destacar también que estos alargamientos casi siempre afectan al último fonema de la última palabra del reparandum. Sin embargo, no hay que olvidar que el 70 % de los alargamientos suceden fuera de las disfluencias sintácticas, de ahí que no podamos considerarlos como pistas definitivas.

Para obtener un perfil más detallado de la inter-dependencia entre fenómenos acústicos y disfluencias sintácticas, hemos considerado el número de disfluencias acústicas y el número de disfluencias sintácticas en cada turno de INFOTREN. A continuación, hemos calculado el coeficiente de correlación de Pearson para ese conjunto de valores. Muchos turnos no contienen ni disfluencias acústicas ni disfluencias sintácticas. Para obtener una relación más realista entre ambas variables, sólo se han considerado aquellos turnos que contienen uno u otro tipo de disfluencia. En estas condiciones se ha obtenido una correlación de $\rho = 0,5816$ (two tailed t-test, $t(961)=22.1649$, $p=0.0000$), que es positiva pero no muy elevada: sólo el 33.83 % de la varianza del número de disfluencias sintácticas se puede atribuir a la varianza del número de disfluencias acústicas.

3.6.1.2. Resultados sobre CORLEC-EHU-1.

De las 2307 disfluencias sintácticas de CORLEC-EHU-1, 1361 (un 58.99 %) contienen fenómenos acústicos. Por otra parte, de los 10821 fenómenos acústicos de CORLEC-EHU-1 sólo 1786 (un 16.50 %) aparecen dentro de disfluencias sintácticas, con la siguiente distribución: 174 ruidos (9.74 %), 1111 alargamientos (62.21 %), 260 pausas de silencio (14.56 %) y 241 pausas habladas (13.49 %). La distribución detallada de fenómenos acústicos dentro de las disfluencias sintácticas se muestra en la Tabla 3.26.

La Tabla 3.26 es muy similar a la Tabla 3.25 y podemos extraer las mismas conclusiones: las pausas y los ruidos aparecen casi exclusivamente como *señales de edición*, mientras que los alargamientos suelen afectar al último fonema de la última palabra del *reparandum*, por lo que se convierten en una pista fundamental para la detección de reformulaciones y frases abandonadas. De 2104 reformulaciones, 983 (un 46.72 %) contienen alargamientos, casi todos ubicados en el *reparandum*. Por otra parte, de 203 frases abandonadas, 112 (un 55.17 %) contienen alargamientos, de los cuales 95 son finales, es decir, marcan el *abandono* de la frase. Al igual que en INFOTREN, no hay que olvidar que la mayor parte de los alargamientos (2482 de 3593, un 69.08 %) sucede fuera de las disfluencias sintácticas, de ahí que no podamos considerar los alargamientos más que como posibles indicadores de la presencia de disfluencias

Tabla 3.26: Distribución de fenómenos acústicos dentro de las disfluencias sintácticas en CORLEC-EHU-1. R: ruidos, S: pausas de silencio, H: pausas habladas, A: alargamientos.

	Reparandum				Señal de edición				Corrección			
	R	S	H	A	R	S	H	A	R	S	H	A
repeticiones	5	3	1	687	132	187	151	2	12	11	7	138
sustituciones	5	5	10	180	22	45	58	3	7	6	7	40
reformulaciones con inserción	0	2	1	32	6	9	14	0	4	0	6	7
reformulaciones con borrado	1	1	0	12	0	4	3	0	0	0	0	2
frases abandonadas	5	8	5	121	-				-			
Totales	16	19	17	1032	160	245	226	5	23	17	20	187

sintácticas.

Por último, el coeficiente de correlación entre el número de disfluencias acústicas y el número de disfluencias sintácticas, calculado para los 1696 turnos de CORLEC-EHU-1 que contienen al menos uno de estos tipos de fenómenos, es $\rho = 0,5213$ (two tailed t-test, $t(1694)=25.1443$, $p=0.0000$). Al igual que en INFOTREN, se trata de una correlación positiva pero no muy elevada: sólo el 27.18 % de la varianza del número de disfluencias sintácticas se puede atribuir a la varianza del número de disfluencias acústicas.

3.6.2. Distorsiones léxicas y disfluencias sintácticas.

3.6.2.1. Resultados sobre INFOTREN.

De las 200 distorsiones léxicas que encontramos en INFOTREN, 90 aparecen dentro y 110 fuera de disfluencias sintácticas. Por otra parte, de 545 disfluencias sintácticas sólo 90 (un 16.51 %) contienen distorsiones léxicas. Para ilustrar en mayor medida la escasa dependencia entre unos y otros fenómenos, se ha calculado el coeficiente de correlación entre el número de distorsiones léxicas y el número de disfluencias sintácticas en cada turno de usuario de INFOTREN, considerando sólo aquellos turnos donde se produce uno y/u otro fenómeno. El resultado es $\rho = -0,0098$ (two tailed t-test, $t(397)=0.1959$, $p=0.8448$), una correlación prácticamente nula y negativa.

Sin embargo, un estudio más detallado revela que de las distorsiones léxicas que aparecen dentro de disfluencias sintácticas, 76 son palabras cortadas y 14 palabras mal pronunciadas, mientras que de las 110 que aparecen fuera, 19 son palabras cortadas y 92 palabras mal pronunciadas. En resumen, un 80 % de las palabras cortadas aparece dentro de disfluencias sintácticas, normalmente como parte del *reparandum* y constituyendo el elemento desencadenante de una reformulación (véase la Tabla 3.27). Esto confirma la caracterización del fenómeno que dimos en el apartado 2.1.4.2. Así pues, cada vez que se detecte una palabra cortada —ése es otro problema— hay una alta probabilidad de que se produzca una reformulación, en concreto $\text{Prob}(\text{reformulación}|\text{palabra_cortada})=0.8$.

3.6.2.2. Resultados sobre CORLEC-EHU-1.

De 1488 distorsiones léxicas, 265 (un 17.81 %) aparecen dentro y 1223 fuera de disfluencias sintácticas. Por otra parte, de 2307 disfluencias sintácticas sólo 250 (un 10.84 %) contienen distorsiones léxicas. El coeficiente de correlación entre el número de distorsiones léxicas

Tabla 3.27: Distorsiones léxicas: palabras mal pronunciadas (M) y palabras cortadas (C), dentro de disfluencias sintácticas en INFOTREN.

	Reparandum		Señal de edición		Corrección	
	M	C	M	C	M	C
repeticiones	10	29	0	0	0	2
sustituciones	3	33	0	0	2	3
reformulaciones con inserción	0	8	0	0	0	1
reformulaciones con borrado	0	1	0	0	0	0
frases abandonadas	0	10	–	–	–	–
Totales	13	81	0	0	2	6

Tabla 3.28: Distorsiones léxicas: palabras mal pronunciadas (M), palabras cortadas (C) y afirmaciones guturales (G) dentro de disfluencias sintácticas en CORLEC-EHU-1.

	Reparandum			Señal de edición			Corrección		
	M	C	G	M	C	G	M	C	G
repeticiones	34	113	0	1	0	0	23	13	2
sustituciones	3	45	0	1	0	0	5	4	0
reformulaciones con inserción	3	22	0	0	0	0	4	0	0
reformulaciones con borrado	0	0	0	0	0	0	0	0	0
frases abandonadas	13	5	0	–	–	–	–	–	–
Totales	53	185	0	2	0	0	32	17	2

y el número de disfluencias sintácticas en cada turno de CORLEC-EHU-1, considerando sólo los 1485 turnos donde se produce uno u otro fenómeno, es $\rho = 0,1268$ (two tailed t-test, $t(1483)=4.9245$, $p=0.0000$), es decir, una correlación positiva pero muy pequeña.

En lo que respecta a la distribución de las distorsiones léxicas que aparecen dentro de disfluencias sintácticas, 184 son palabras cortadas y 80 palabras mal pronunciadas, mientras que sólo se ha observado una afirmación gutural. De las 1223 que aparecen fuera, sólo 38 son palabras cortadas, mientras que 888 son palabras mal pronunciadas y 297 afirmaciones guturales. Es decir, al igual que en INFOTREN, lo más destacable es que la mayor parte de las palabras cortadas aparece dentro de disfluencias sintácticas, como elemento desencadenante de una reformulación (véase la Tabla 3.28). La probabilidad de que tras una palabra cortada se produzca una disfluencia sintáctica es en este caso de 0,8288.

3.6.3. Marcadores de discurso y disfluencias sintácticas.

3.6.3.1. Resultados sobre INFOTREN.

Cuando en el apartado 2.1.5.1 se describía la estructura superficial de las reformulaciones, se decía que como *señales de edición* podían aparecer, entre otros fenómenos, marcadores de discurso, expresiones como *bueno*, *perdón*, *es decir*, etc. Sin embargo, al examinar los marcadores de discurso en INFOTREN, resulta que de 865 instancias, tan sólo 45 (un 5.20 %) aparecen dentro de disfluencias sintácticas (véase la Tabla 3.29). Por otra parte, sólo 39 de las 545 disfluencias sintácticas (un 7.16 %) contienen marcadores de discurso. Finalmente, el coeficiente de correlación entre marcadores de discurso y disfluencias sintácticas, considerando sólo turnos

Tabla 3.29: Número de marcadores de discurso (MD) que aparecen dentro y fuera de disfluencias sintácticas (DS) en INFOTREN. Se indica también el porcentaje que representan esos números dentro de cada categoría.

Función MD	Dentro DS	Fuera DS
Abrir	1(0.67 %)	149(99.33 %)
Cerrar	0(0.00 %)	189(100.00 %)
Aceptar	1(1.28 %)	77(98.72 %)
Rechazar	3(6.67 %)	42(93.33 %)
Explicar/Editar	20(28.17 %)	51(71.83 %)
Pedir	0(0.00 %)	92(100.00 %)
Rellenar	12(5.33 %)	213(94.67 %)
Exclamar	8(53.33 %)	7(46.67 %)
Totales	45(5.20 %)	820(94.80 %)

de INFOTREN que contengan uno u otro tipo de fenómenos, resulta ser casi nulo y negativo: $\rho = -0,0248$ (two tailed t-test, $t(717)=0.6651$, $p=0.5062$).

En la Tabla 3.29 puede comprobarse que el 44.44 % de los marcadores de discurso que aparecen dentro de disfluencias sintácticas cumplen la función *explicar/editar*, lo cual es destacable, ya que este tipo de marcadores tienen una frecuencia muy inferior cuando aparecen fuera (6.22 %). Sin embargo, sólo el 28.17 % de este tipo de marcadores aparece dentro de disfluencias sintácticas. Así pues, en lo que respecta a INFOTREN no parece que los marcadores de discurso puedan utilizarse como pistas fiables para la detección de disfluencias sintácticas, ni tan siquiera en la categoría más específicamente relacionada con ellas.

3.6.3.2. Resultados sobre CORLEC-EHU-1.

Sobre CORLEC-EHU-1 se obtienen resultados aparentemente más prometedores. Por un lado, de 2963 marcadores de discurso 296 (un 9.99 %, frente al 5.20 % de INFOTREN) aparecen dentro de disfluencias sintácticas; por otro, de 2307 disfluencias sintácticas, 213 (un 9.23 %, frente al 7.16 % de INFOTREN) contienen marcadores de discurso. El coeficiente de correlación entre marcadores de discurso y disfluencias sintácticas, considerando sólo los 2035 turnos que contienen uno u otro tipo de fenómenos, es positivo aunque pequeño: $\rho = 0,3412$ (two tailed t-test, $t(2033)=16.3655$, $p=0.0000$).

Considerese ahora la Tabla 3.30, que nos permite saber qué relación hay entre el número de marcadores que aparecen en el *reparandum*, y el número de ellos que aparecen en la *señal de edición* o en la *corrección*, y si lo hacen en repeticiones, en sustituciones, etc. El número de marcadores de discurso que aparece en dicha tabla no coincide con el número indicado más arriba, debido a que muchos de ellos, por ejemplo los que aparecen en repeticiones múltiples, se cuentan dos o más veces.

Lo primero que observamos en dicha tabla es que la mayor parte de los marcadores de discurso que aparecen dentro de disfluencias sintácticas lo hacen en repeticiones. Se trata de casos en los que el hablante repite el mismo marcador varias veces, como en *Sí, sí, sí*, que podría anotarse como marcador de aceptación, o en *No, no*, que podría anotarse como marcador de rechazo. De hecho, de los 339 marcadores que aparecen en el *reparandum* o en la *corrección*, más del 80 % son marcadores de aceptación o rechazo. Sin embargo, lo que buscamos son

Tabla 3.30: Distribución de los marcadores de discurso que aparecen dentro de disfluencias sintácticas en CORLEC-EHU-1.

	Reparandum	Señal de edición	Corrección
repeticiones	121	46	199
sustituciones	1	29	2
reformulaciones con inserción	0	5	2
reformulaciones con borrado	1	2	1
frases abandonadas	12	–	–
Totales	135	82	204

marcadores de discurso *que indiquen la presencia de una reformulación, no que el propio marcador sea el sujeto de la reformulación*. Precisamente, sólo 82 marcadores aparecen en la *señal de edición*, de los cuales 58 pertenecen a la categoría *editar/explicar* y 20 a la categoría *rellenar*. Por otra parte, de 2104 reformulaciones sólo 675 tienen *señal de edición*, y de éstas sólo 55 (un 8.15 %) contienen marcadores de edición/explicación. Además, aunque los marcadores de edición/explicación tienen una frecuencia relativa más alta dentro que fuera de las disfluencias sintácticas (21.28 % frente a 6.41 %), el hecho es que de los 234 anotados en CORLEC-EHU-1, la mayor parte (171, un 73.08 %) aparecen fuera de disfluencias sintácticas. De ahí que también en este caso no podamos sino descartar los marcadores de discurso como pistas fiables para la detección de reformulaciones.

Resumen del capítulo.

En este capítulo se han analizado las anotaciones de fenómenos de habla espontánea en dos bases de datos, una de ellas (INFOTREN) formada por diálogos hombre-máquina en una tarea de acceso a información sobre trayectos, horarios y precios de viajes en tren, y la otra (CORLEC-EHU-1) formada por entrevistas tomadas de radio y televisión, completamente libres en cuanto al léxico y la gramática, y por tanto mucho menos restringidas que los diálogos hombre-máquina. La presión cognitiva e interactiva es más alta en el caso de las entrevistas, lo cual, *a priori*, debería propiciar la aparición de un mayor número de fenómenos de habla espontánea. Sin embargo, no se observan diferencias significativas en el número relativo de fenómenos que aparecen en diálogos hombre-máquina y en conversaciones naturales entre personas. Las diferencias se refieren más bien a la distribución de fenómenos. La proporción de distorsiones léxicas y reformulaciones es notablemente mayor en el habla espontánea natural: se producen más errores y, por tanto, más correcciones. Muchos de los fenómenos aparecen no como respuesta a perturbaciones externas —la presión cognitiva o interactiva a que aludíamos—, sino como recursos propios del habla espontánea. En este sentido, es destacable el gran número de pausas habladas, alargamientos y repeticiones, que no deben verse como errores o problemas del habla sino que cumplen una función concreta: marcar la apropiación del turno mientras se planifica el discurso. Depende de los usos del hablante el que se utilicen unas u otras formas. En particular, los alargamientos y las pausas habladas muestran porcentajes de aparición comparables, si bien en la literatura las pausas habladas han recibido una atención mucho mayor que los alargamientos.

El número de fenómenos en una intervención de los diálogos muestra una correlación notable con la longitud de la misma —expresada ésta en segundos o en número de palabras

efectivas—, particularmente en el caso de los ruidos producidos por el propio hablante y las disfluencias acústicas. Se han confirmado los resultados de (Shriberg, 1994) en cuanto a que la probabilidad de que una intervención sea fluida decrece exponencialmente con la longitud de la misma, y depende de un parámetro b , conocido como *tasa de fluencia*. Se ha comprobado asimismo que la tasa de fluencia de INFOTREN ($b = 0,8838$) es prácticamente igual a la de CORLEC-EHU-1 ($b = 0,8996$). De estos valores se deduce que la probabilidad de que tras una palabra se produzca una disfluencia es aproximadamente de 0,1 en ambos casos. Por otra parte, el número de fenómenos varía notablemente de unos hablantes a otros. Ello es debido principalmente a que unos hablantes *hablan durante más tiempo* que otros. No obstante, se han presentado ejemplos que demuestran que tanto en INFOTREN como en CORLEC-EHU-1 existen hablantes intrínsecamente más disfluentes que otros —con tiempos de elocución similares. Por último, no parece posible determinar la presencia de disfluencias sintácticas utilizando como indicador únicamente la presencia de disfluencias acústicas, distorsiones léxicas o marcadores de discurso. Así, por ejemplo, entre el 50 % y el 60 % de las disfluencias sintácticas contienen alargamientos (típicamente en el último fonema del *reparandum*), pero alrededor del 70 % de los alargamientos aparece fuera de disfluencias sintácticas. En el mismo sentido, el 80 % de las palabras cortadas aparece como elemento desencadenante de una reformulación, pero sólo una pequeña fracción de las reformulaciones están motivadas por la presencia de una palabra cortada.

Capítulo 4

El sistema de reconocimiento de referencia.

El objetivo tecnológico principal de este trabajo es mejorar las prestaciones de los sistemas de reconocimiento del habla, al menos en lo que se refiere al uso del habla espontánea en las interacciones. Para ello, en primer lugar es necesario establecer un sistema de reconocimiento básico de referencia y a continuación plantear mejoras sobre el mismo. En este capítulo se describen las características del sistema de reconocimiento del *Grupo de Reconocimiento de Formas y Tecnologías del Habla* (GRFyTH) de la UPV/EHU, tal como se aplica sobre habla leída, y se muestran los resultados de decodificación acústico-fonética y de reconocimiento sobre habla espontánea utilizando dicha configuración. También se presentan resultados sobre dos bases de datos de habla leída, con objeto de contrastarlos con los de habla espontánea y establecer posibles fuentes de error. Del análisis de estos resultados deberán extraerse conclusiones que permitan abordar mejoras del sistema.

4.1. Descripción general del sistema de reconocimiento.

El proceso de reconocimiento requiere (véase la Figura 4.1): (1) el procesamiento de la señal de voz para obtener los parámetros acústicos; (2) sólo en el caso de utilizar modelos acústicos discretos, el etiquetado de los vectores de parámetros, que se realiza a partir de un diccionario de vectores calculado previamente mediante cuantificación vectorial; y (3) el reconocimiento propiamente dicho, que en nuestra implementación consiste en un autómata probabilístico sobre el que se efectúa la búsqueda. En este autómata se integran los modelos acústicos y el modelo de lenguaje, que han sido estimados previamente a partir de un *corpus de entrenamiento*.

4.1.1. Los parámetros acústicos.

En este trabajo se utilizan como parámetros acústicos los coeficientes cepstrales, calculados a partir de un banco de filtros cuyas anchuras reproducen la escala de percepción frecuencial del oído humano (Davis and Mermelstein, 1980). Este es el método de análisis que la mayor parte de los sistemas de reconocimiento aplican cuando el entorno no es muy ruidoso (Picone, 1993; Young et al., 1999), y que de hecho ha sido adoptado recientemente como representación estándar para los sistemas de reconocimiento distribuido que operan con habla limpia (ETSI, 2003a). Evidentemente, en entornos ruidosos otros métodos de análisis han demostrado una

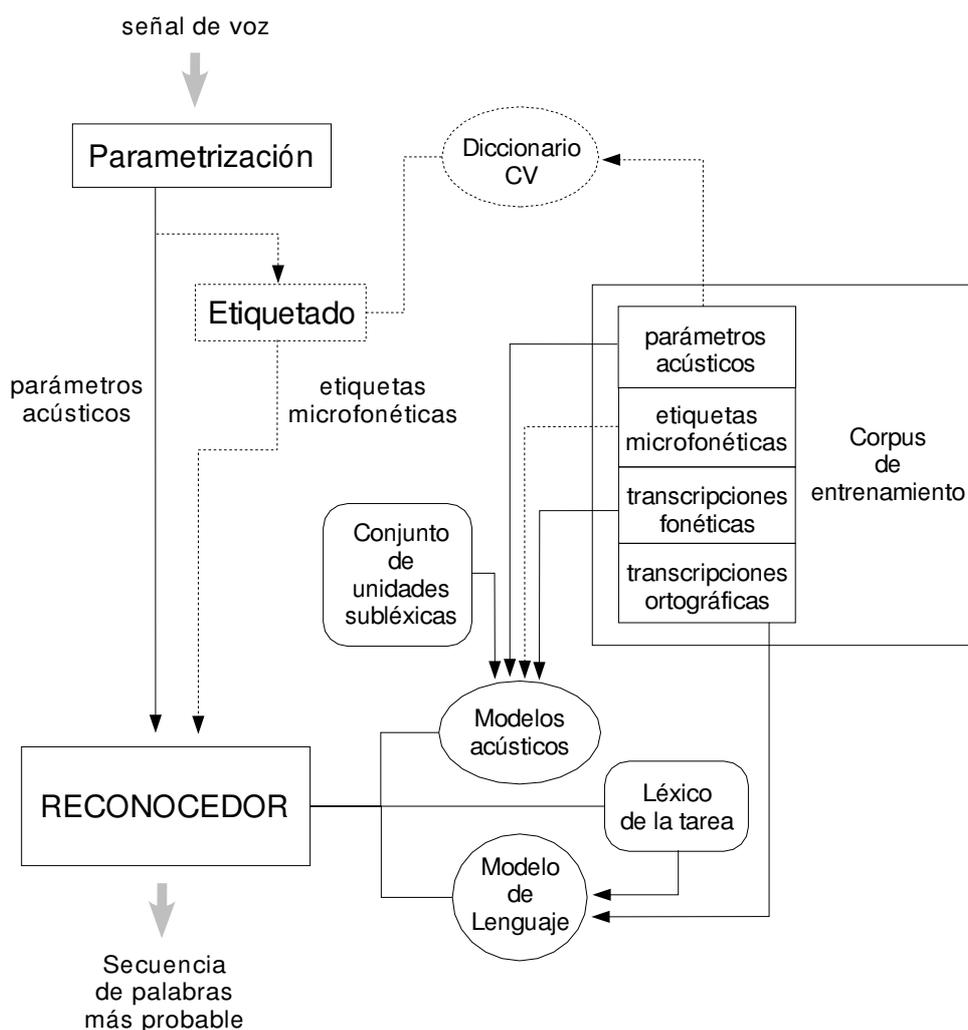


Figura 4.1: El sistema de reconocimiento.

mayor robustez (Hermansky and Morgan, 1994; Nadeu et~al., 2001). En el caso de CORLEC-EHU-1, parece claro que alguno de esos métodos deberá ser aplicado en el futuro. En cualquier caso, el procesamiento básico incluye dos transformaciones de los cepstrales que disminuyen su sensibilidad al canal, al entorno y al hablante:

- ▷ En primer lugar, se aplica a los cepstrales una ventana de tipo *seno realzado* —lo que en inglés se conoce como *liftering*, una especie de filtro frecuencial— que reduce la sensibilidad a las condiciones del canal y del hablante, y que mejora el rendimiento del reconocedor cuando se utiliza una distancia euclídea entre cepstrales, como es el caso de los modelos discretos (Juang et~al., 1987; Tokhura, 1987). Sin embargo, no afecta al rendimiento de los modelos continuos, ya que en ese caso cada componente de los cepstrales es normalizada por su varianza en el cómputo de las gaussianas.
- ▷ A continuación se aplica lo que se conoce como *Cepstral Mean Normalization* (CMN) (Rosenberg et~al., 1994), que consiste en restar a cada componente de los cepstrales su valor medio, calculado sobre un cierto intervalo de tiempo. En este trabajo se calcula el

Tabla 4.1: Características del procesamiento de la señal de voz a 16 y 8 kHz. Se muestran también las características de un procesamiento alternativo a 8 kHz propuesto por el *European Telecommunications Standards Institute* (ETSI) para entornos ruidosos.

	16 kHz	8 kHz	8 kHz (ETSI)
Longitud de tramo	25 ms	25 ms	25 ms
Distancia entre tramos	10 ms	10 ms	10 ms
Eliminar DC	Sí	Sí	Sí
Preénfasis	No	No	Sí
Ventana de análisis	Hamming	Hamming	Hamming
Longitud FFT	512 muestras	256 muestras	256 muestras
Rango de frecuencia	0-8000 Hz	0-4000 Hz	0-4000 Hz
Número de filtros	24	20	24
Tipo de filtro	Triangular	Triangular	Triangular
Tipo ventana cepstral	Seno realzado	Seno realzado	Rectangular
Longitud ventana cepstral	12	10	12
CMN	Sí	Sí	No

valor medio de los cepstrales para cada pronunciación. Esta especie de *filtrado temporal* elimina componentes ruidosas estacionarias convolucionadas con la señal de voz.

También se calcula la energía de cada tramo, normalizando su valor de acuerdo al máximo obtenido para cada pronunciación. Finalmente se calculan las derivadas primera y segunda del vector de cepstrales y la derivada primera de la energía, mediante una fórmula de regresión que típicamente abarca uno o dos tramos alrededor del tramo de análisis. La aproximación habitual consiste en componer todas las informaciones acústicas, estáticas y dinámicas, en un único vector. En este trabajo, por el contrario, los cepstrales y sus derivadas primera y segunda se manejan como tres vectores independientes, mientras que la energía y su primera derivada conforman un cuarto vector de dos componentes.

En la Tabla 4.1 se resumen las características del procesamiento de la señal de voz a 16 y 8 kHz, así como las características de un procesamiento alternativo a 8 kHz que básicamente coincide con el descrito en el estándar del ETSI (*European Telecommunications Standards Institute*) para entornos ruidosos (ETSI, 2003b).

4.1.2. La cuantificación vectorial.

Se ha desarrollado y aplicado una versión del algoritmo LBG (Linde et al., 1980) para generar los diccionarios de vectores correspondientes a las cuatro representaciones acústicas descritas en el apartado anterior. Se ha utilizado únicamente el corpus de entrenamiento de cada base de datos, de manera que el proceso de cuantificación es independiente del corpus de test. Como medida de distorsión se utiliza la distancia euclídea. Del proceso no sólo se obtienen los centroides sino también sus varianzas internas, que servirán, como veremos, para inicializar los modelos continuos, suponiendo matrices de covarianza diagonales. La distorsión promedio que resulta de este proceso, es decir, la distancia promedio de un vector a su centroide, depende del umbral de convergencia, pero sobre todo de la representación acústica que se esté considerando —ya que sus rangos dinámicos son muy distintos— y de la base de datos. Finalmente, con

los diccionarios de centroides se etiquetan el corpus de entrenamiento y el corpus de test. Las secuencias de etiquetas microfonéticas que resultan del etiquetado se guardan en ficheros que pasan a formar parte de la base de datos. En adelante nos referiremos a los diccionarios de centroides —acompañados de sus varianzas internas— como *diccionarios CV*.

4.1.3. El reconocedor.

Los elementos básicos de un reconocedor automático del habla son los modelos acústicos, el modelo de lenguaje y el algoritmo de búsqueda de hipótesis. En este trabajo se utilizan por un lado Modelos Ocultos de Markov (MOM) discretos, continuos y semicontinuos (Lee, 1989; Rabiner, 1989; Huang et al., 1990), y por otro, modelos k-explorables en sentido estricto (k-ESS) (Torres and Varona, 2001), que se integran en un único autómata probabilístico. Sobre este autómata se efectúa una búsqueda en haz mediante un algoritmo de programación dinámica que proporciona la hipótesis más probable (Varona, 2000). En cada paso de esa búsqueda las probabilidades acústicas y del modelo de lenguaje se combinan según un peso α cuyo valor se ajusta de manera heurística. Tanto los MOM como los modelos k-ESS son estimados y optimizados de forma independiente, si bien en última instancia su rendimiento debe verificarse una vez integrados en el autómata de búsqueda. En los párrafos siguientes se describen con más detalle algunos de los elementos necesarios para construir un reconocedor de estas características. Dada su importancia en este trabajo, los modelos acústicos se tratan por separado en el apartado 4.2.

4.1.3.1. El conjunto de unidades subléxicas.

En la elección del conjunto de unidades subléxicas hay que tener en cuenta dos requerimientos básicos: (1) debe ser posible construir cualquier palabra con las unidades de dicho conjunto; y (2) cada una de las unidades ha de disponer de muestras suficientes como para estimar de forma robusta el modelo acústico correspondiente. Además, debería ser posible distinguir acústicamente dos unidades cualesquiera, es decir, el grado de confusión acústica ha de ser mínimo en dicho conjunto. El conjunto más natural es el de los sonidos —los fonemas— de la lengua. Para el castellano se han propuesto otros conjuntos alternativos, fundamentalmente fonemas contextuales (Bonafonte et al., 1995; López de Ipiña et al., 1999), pero también unidades contextuales más pequeñas que el fonema (Mariño et al., 1997). En este trabajo, como aproximación básica, se considera un conjunto reducido de unidades incontextuales, formado por 23 fonemas y una unidad especial para el silencio (véase la Tabla 4.2). Este conjunto ha mostrado un comportamiento óptimo sobre habla leída frente a otros conjuntos de unidades incontextuales más amplios y detallados (Torres, 1992).

4.1.3.2. Las transcripciones fonéticas.

Los modelos acústicos se estiman estadísticamente a partir de las señales que forman el corpus de entrenamiento, en concreto a partir de los parámetros acústicos (vectores o etiquetas) y las transcripciones en términos de unidades subléxicas, que por simplificar denominaremos *transcripciones fonéticas*. Para obtener éstas últimas es necesario procesar las transcripciones ortográficas, bien mediante un conjunto de reglas de transcripción, bien mediante un diccionario de pronunciaciones, bien mediante una estrategia mixta. Por ejemplo, en inglés es necesario recurrir a un diccionario de pronunciaciones, porque aunque sí existen reglas, el número

Tabla 4.2: Conjunto de unidades subléxicas: codificación SAMPA, codificación interna del GRFyTH, descripción del sonido y ejemplo.

Codificación SAMPA	Codificación GRFyTH	Descripción	Ejemplo
p	p	oclusiva, bilabial, sorda	partido
b	b	oclusiva, bilabial, sonora	verde
t	t	oclusiva, dental, sorda	tanque
d	d	oclusiva, dental, sonora	dividir
k	k	oclusiva, velar, sorda	constante
g	g	oclusiva, velar, sonora	ganar
tS	c	africada, prepalatal, sorda	coche
f	f	fricativa, labio-dental, sorda	fideo
T	z	fricativa, interdental, sorda	cena
s	s	fricativa, dorso-alveolar, sorda	signo
x	x	fricativa, velar, sorda	gigante
m	m	nasal, bilabial, sonora	mérito
n	n	nasal, alveolar, sonora	nadie
J	h	nasal, palatal, sonora	niño
l	l	lateral, alveolar, sonora	lección
L	y	lateral, palatal, sonora	amarillo
r	r	vibrante, alveolar, sonora, simple	pareja
rr	@	vibrante, alveolar, sonora, múltiple	ratón
i	i	vocal, palatal, sonora, cerrada	idea
e	e	vocal, palatal, sonora, media	entonces
a	a	vocal, central, sonora, abierta	abeto
o	o	vocal, velar, sonora, media-redondeada	oriente
u	u	vocal, velar, sonora, cerrada-redondeada	último
–	P	silencio	–

de excepciones es muy elevado. En castellano, por el contrario, se pueden aplicar unas reglas básicas de pronunciación, en muchos casos independientes del contexto, que producen lo que llamaríamos una *pronunciación canónica*. Obviamente, esta pronunciación puede sufrir desviaciones, que dependen de la variedad dialectal, de la modalidad del habla, de características propias del hablante, etc.

En el caso del habla leída esta variabilidad es muy pequeña o no se produce en absoluto, de manera que la *pronunciación canónica* es una buena descripción *a priori* de la secuencia de sonidos que presumiblemente se ha producido. En el GRFyTH disponemos de un conversor grafema-fonema para el castellano, cedido por la Universidad Politécnica de Valencia, que produce transcripciones fonéticas en términos del conjunto de unidades presentado en la Tabla 4.2, a partir de una serie de reglas tomadas de (Quilis, 1981).

En el caso del habla espontánea, las transcripciones ortográficas se obtienen a partir de los ficheros de anotaciones en formato EHU simplificado. Estas anotaciones contienen fenómenos acústicos y léxicos que no pueden tenerse en cuenta en una aproximación básica como la que se pretende realizar en este capítulo. En las transcripciones fonéticas resultantes sólo pueden aparecer las unidades de la Tabla 4.2. De ahí que no puedan obtenerse transcripciones exactas, sino tan sólo aproximaciones. Para ello, las anotaciones se procesan turno a turno, y de cada turno se obtiene una transcripción ortográfica, aplicando el siguiente conjunto de reglas:

1. Los segmentos solapados o ruidosos se eliminan por completo, ya que sólo se dispone de las señales correspondientes a los segmentos *limpios*. Si el solapamiento o el ruido de fondo afectan al turno completo, entonces no se genera transcripción ortográfica.
2. Los ruidos aislados de todo tipo, las pausas habladas, las afirmaciones/negaciones guturales y las palabras cortadas se eliminan, es decir, no producen ningún símbolo ortográfico.
3. Las pausas de silencio se transcriben con el símbolo ortográfico *P*.
4. Las marcas de alargamientos y de palabras mal pronunciadas no se tienen en cuenta, es decir, producen la misma secuencia de símbolos ortográficos que obtendríamos en su ausencia.
5. Las palabras extranjeras y las siglas se transcriben ortográficamente tal como han sido pronunciadas, es decir, con los grafemas correspondientes a su pronunciación en castellano.
6. Las marcas de fenómenos sintácticos y pragmáticos se eliminan, ya que no influyen en la transcripción fonética.

Finalmente, la transcripción fonética de cada turno se obtiene aplicando a la transcripción ortográfica el mismo conversor grafema-fonema utilizado para habla leída.

4.1.3.3. El léxico de la tarea.

El léxico de la tarea se construye con las palabras que aparecen tanto en el corpus de entrenamiento como en el corpus de test. Se trata, por tanto, de un vocabulario cerrado, necesario porque el reconocedor sólo puede dar como salida una secuencia de palabras *conocidas*. Para

construir el modelo acústico de una palabra, en primer lugar se aplica el transcriptor ortográfico-fonético a la transcripción ortográfica de la misma, y a continuación se concatenan los MOM correspondientes a la transcripción fonética resultante. Se trata de un modelo lineal que maneja una sola pronunciación, y que, sin embargo, será capaz de aceptar pequeñas desviaciones, porque los modelos acústicos también han absorbido en sus parámetros las distintas variantes de pronunciación halladas en el corpus de entrenamiento.

4.1.3.4. El modelo de lenguaje.

Las restricciones de tipo sintáctico se han modelado mediante un cierto tipo de autómatas de estados finitos estocásticos, capaces de generar una clase de lenguajes regulares conocidos como lenguajes k -explorables en sentido estricto (k -ESS). En la literatura es más común encontrar formalismos puramente estadísticos basados en lo que se conoce como n -gramas (Jelinek, 1990). Se ha demostrado que la distribución de probabilidad representada por los autómatas k -ESS equivale a la obtenida con modelos de n -gramas (García and Vidal, 1990). De hecho, los parámetros n y k tienen el mismo significado: el número de palabras que condicionan la probabilidad de observar una nueva palabra.

El modelo de lenguaje utilizado en este trabajo integra K autómatas k -ESS, con $k = 1, 2, \dots, K$, en un único autómata $M = (\Sigma, Q^K, q_0, q_f, \delta^K)$, donde:

- ▷ Σ es el vocabulario del modelo, es decir, el conjunto de palabras que aparecen en el corpus de entrenamiento.
- ▷ Q^K es el conjunto de estados del autómata. Cada estado representa una cadena de palabras de longitud $k - 1$, $w_{k-1}^1 \equiv w_{k-1}w_{k-2} \dots w_1$, con $k = 1, \dots, K$.
- ▷ q_0 es el estado inicial del autómata.
- ▷ q_f es el estado final del autómata.
- ▷ δ^K es la función de transición, definida $\delta^K : Q^K \times \Sigma \rightarrow Q^K \times [0, 1]$, es decir, dado un estado q del autómata y una palabra $w \in \Sigma$, δ^K permite transitar a un *único* estado $q' \in Q^K$ con una probabilidad $P(w|q) \in [0, 1]$. Cada transición equivale en realidad a un cierto k -grama, ya que cada estado representa, como hemos visto, una *historia*: w_{k-1}^1 para un cierto $k \in [1, K]$.

El autómata así definido es determinista, lo cual permite estimar la probabilidad de cada transición $\delta^K(q, w)$ a partir de un corpus de texto, aplicando el criterio de máxima similitud:

$$P(w|q) = \frac{N(w|q)}{\sum_{w' \in \Sigma} N(w'|q)} \quad (4.1)$$

donde $N(w|q)$ es el número de veces que aparece el k -grama w_{k-1}^1w en el corpus de entrenamiento, siendo w_{k-1}^1 la historia asociada al estado q .

La distribución de probabilidad definida en la ecuación 4.1 debe ser modificada para asignar una probabilidad no nula a las combinaciones de palabras que no han sido observadas en el corpus de entrenamiento. Esto es lo que se conoce como *suavizado* del modelo. Qué fracción de probabilidad se extrae de dicha distribución y cómo se asigna a las combinaciones de palabras no observadas es algo que depende del método de suavizado. En este trabajo el suavizado se

realiza recurriendo a un estado con una historia más corta y, por tanto, más plausible, lo que se conoce como estado de *backoff* (Katz, 1987) —que denominaremos b_q y que llevará asociada la misma secuencia de palabras que q , salvo la primera, es decir, w_{k-2}^1 . La probabilidad de transición, suavizada por *backoff* con descuento de *Witten-Bell* (Witten and Bell, 1991), se calcula como sigue:

$$P(w|q) = \begin{cases} \frac{N(w|q)}{|\Sigma_q| + \sum_{w' \in \Sigma_q} N(w'|q)} & w \in \Sigma_q \\ P(w|b_q)P(b_q|q) & w \notin \Sigma_q \end{cases} \quad (4.2)$$

donde:

$$\Sigma_q = \{w' \in \Sigma | N(w'|q) > 0\} \quad (4.3)$$

$$P(b_q|q) = \frac{|\Sigma_q|}{|\Sigma_q| + \sum_{w' \in \Sigma_q} N(w'|q)} \frac{1}{1 - \sum_{w' \in \Sigma_q} P(w'|b_q)} \quad (4.4)$$

Así pues, el autómata suavizado se construye añadiendo al autómata original las transiciones correspondientes a los estados de *backoff*. Para cada estado $q \in Q^K$, se define $\delta^K(q, U) = (b_q, P(b_q|q))$, donde U representa todas las palabras tales que $N(U|q) = 0$.

El autómata K-ESS suavizado se representa finalmente mediante un array de transiciones (Varona, 2000; Torres and Varona, 2000, 2001). Cada estado q —que suponemos lleva asociada una cadena de longitud $k - 1$, para un cierto $k \in [1, K]$ — se representa mediante $|\Sigma_q| + 1$ posiciones consecutivas del array, las $|\Sigma_q|$ primeras correspondientes a las transiciones observadas en el corpus de entrenamiento, que llevan a estados en el mismo nivel k del autómata, y la última correspondiente a todas las combinaciones no observadas, que llevan al estado de *backoff*, que se encuentra en el nivel $k - 1$ del autómata. Cada posición del array contiene el identificador del estado q , la *historia* asociada al mismo, la palabra que determina la transición, la probabilidad de la misma y el estado destino —en realidad, la primera de las posiciones del array correspondientes al estado destino. El array completo consta de $\sum_{q \in Q^K} (|\Sigma_q| + 1)$ posiciones, por lo que se trata de una representación muy compacta.

Para completar la representación del autómata K-ESS suavizado, es necesario definir una función de búsqueda que opere de manera recursiva en los casos en que aparezcan combinaciones no observadas en el corpus de entrenamiento, pasando por uno o varios estados de *backoff*, incluso hasta llegar el unigrama, si fuera necesario. En cambio, si la combinación ha sido observada en el corpus de entrenamiento, la función de búsqueda simplemente retorna la probabilidad y el estado destino que encuentra en la posición correspondiente del array. El coste computacional de la función de búsqueda depende básicamente del nivel máximo K considerado, que suele ser pequeño —entre 2 y 4—, y del número promedio de transiciones en cada estado del autómata.

4.1.3.5. Evaluación del sistema de reconocimiento integrado.

Para obtener el error de reconocimiento, se aplica el reconocedor integrado a cada muestra del corpus de test. Como resultado se obtiene la secuencia de palabras más probable, que deberá ser alineada con la transcripción correcta con el criterio de minimizar el coste de las operaciones de edición. Tanto la secuencia reconocida como la transcripción correcta están segmentadas en frases, que son las unidades máximas del modelo de lenguaje. Por otra parte, también aparecerán

pseudo-palabras que representan intervalos de silencio. Tanto las marcas de fin de frase como los silencios son eliminados de una y otra secuencia antes de llevar a cabo los alineamientos, ya que sólo nos interesa la precisión del sistema en la conversión voz-texto. De los alineamientos se obtienen el número de palabras acertadas (a), sustituidas (s), borradas (b) insertadas (i). La tasa de error (WER: *Word Error Rate*, en inglés) se calcula según la siguiente expresión:

$$\% WER = \frac{s + b + i}{a + s + b + i} * 100$$

4.2. Los modelos acústicos.

Una parte importante de la experimentación llevada a cabo en este trabajo se dedica a la estimación y evaluación de los modelos acústicos. Se han utilizado MOM discretos, continuos y semicontinuos, con objeto de comparar el rendimiento de todos ellos sobre distintas bases de datos de habla leída y espontánea. El interés de los MOM discretos radica sobre todo en la importantísima reducción del coste computacional que suponen con respecto a los MOM continuos, tanto en lo que se refiere a la estimación de los modelos como —sobre todo— en lo que respecta a la decodificación de las señales de voz. Los modelos discretos encuentran aplicación en tareas muy restringidas, con perplejidades bajas, en las que el reconocimiento está guiado por el modelo de lenguaje, y los modelos acústicos no son determinantes. Cuando se dan las dos circunstancias mencionadas —necesidad de respuesta en tiempo real con recursos computacionales limitados y tareas muy restringidas— los MOM discretos se convierten en seria alternativa a los MOM continuos. Los MOM semicontinuos constituyen una aproximación intermedia entre los MOM discretos y los MOM continuos. En el GRFyTH se incorporaron de forma temprana como alternativa a los MOM discretos (Torres and Casacuberta, 1993a,b; Torres et~al., 1994). Sin embargo, su rendimiento es inferior al de los MOM continuos, y además el procedimiento de estimación tiene un mayor coste computacional. En este trabajo los MOM semicontinuos se han definido como un caso particular de los MOM continuos en el que todos los estados de todos los modelos comparten las mismas gaussianas, de manera que se ha desarrollado un nuevo y único software de entrenamiento y test para ambos tipos de MOM. No sucede lo mismo con los MOM discretos, ya que existen diferencias formales importantes con respecto a los MOM continuos, por lo que se ha desarrollado un software específico.

Como se ha indicado en el apartado 4.1.1, la información acústica se divide en cuatro representaciones que suponemos estadísticamente independientes. Por tanto, la probabilidad de emisión de los MOM se obtiene del producto de las probabilidades de emisión para estas cuatro representaciones acústicas, todas ellas con el mismo peso. En base a la experiencia con distintas topologías de MOM, se ha establecido una topología izquierda-derecha muy sencilla con 3 estados, de los cuales sólo uno —el primero— es inicial y sólo uno —el último— final (véase la Figura 4.2). El tamaño de los diccionarios CV en el caso de MOM discretos, y el número de gaussianas en el caso de MOM continuos, son iguales para todas las representaciones acústicas.

4.2.1. Estimación de los modelos acústicos.

Los MOM se han estimado de acuerdo al criterio de máxima similitud, es decir, modificando sus parámetros de manera que la probabilidad de las muestras del corpus de entrenamiento sea máxima. El método más ortodoxo para realizar dicha estimación es el *algoritmo EM* (*Expectation-Maximization*) (Dempster et~al., 1977), que en su versión para MOM se conoce

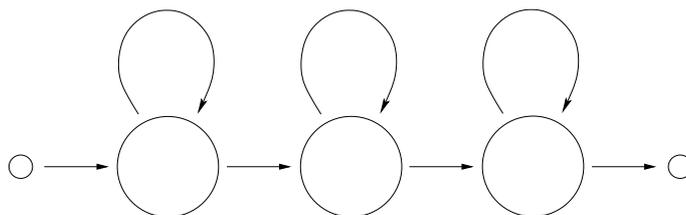


Figura 4.2: Topología de los MOM de unidades subléxicas.

como *algoritmo de Baum-Welch* (Huang et al., 1990). Es posible aproximar el criterio de máxima similitud cambiando la función objetivo: en lugar de maximizar la probabilidad de la muestra dado el modelo, se maximiza la probabilidad del camino óptimo a través del modelo, es decir, el operador suma es sustituido por el operador máximo. Este criterio conduce a lo que se conoce como *entrenamiento por Viterbi* (Juang and Rabiner, 1990; Merhav and Ephraim, 1991). No debe confundirse este algoritmo con el que se utiliza para obtener la decodificación óptima en reconocimiento. En ambos casos se utiliza el mismo algoritmo de programación dinámica, conocido como *algoritmo de Viterbi* (Forney, 1973) —lineal con la longitud de la muestra y cuadrático con el número de estados del modelo—, sólo que en entrenamiento, conocida la transcripción fonética correcta, lo que se obtiene es la segmentación más probable, la única que va a contribuir a la estimación de los modelos.

En los siguientes apartados se discuten las ventajas e inconvenientes de los algoritmos de Baum-Welch y Viterbi en entrenamiento, y se propone un algoritmo híbrido que trata de ser un balance entre ambas aproximaciones.

4.2.1.1. El algoritmo de Baum-Welch.

Existen dos versiones de este algoritmo, una basada en una segmentación explícita de las muestras de entrenamiento, que llamaremos *Baum-Welch restringido*, y otra que maneja todas las posibles segmentaciones, que llamaremos *Baum-Welch general*.

En el *algoritmo de Baum-Welch restringido* se tiene un conjunto de segmentos para cada unidad subléxica, y el MOM correspondiente (inicializado con distribuciones uniformes si se trata de un MOM discreto) se entrena de manera independiente a partir de esos segmentos. Si se utilizan MOM de topologías izquierda-derecha similares a la de la Figura 4.2, la complejidad temporal del algoritmo está aproximadamente en $O(ETC)$ en el caso de MOM discretos, y en $O(ETGD)$ en el caso de MOM continuos (Rodríguez and Torres, 2003), donde E : número de estados de los MOM, T : tamaño (número de tramos) del corpus de entrenamiento, C : tamaño total de los diccionarios CV, G : número de componentes en las mezclas de gaussianas y D : número total de componentes en los vectores acústicos.

El *algoritmo de Baum-Welch general* parte únicamente de las observaciones acústicas y de las transcripciones fonéticas de las muestras, es decir, no utiliza segmentaciones. Para cada muestra se construye un gran MOM concatenando los MOM correspondientes a las unidades de la transcripción fonética. Todas las posibles segmentaciones de cada muestra contribuyen a la estimación de los MOM, con unos pesos que dependen de la probabilidad de cada segmentación. Si se utilizan MOM de topologías izquierda-derecha similares a la de la Figura 4.2, la complejidad temporal del algoritmo está aproximadamente en $O(ETCF)$ en el caso de MOM discretos,

y en $O(ETGDF)$ en el caso de MOM continuos (Rodríguez and Torres, 2003), donde:

$$F = \frac{1}{T} \sum_{l=1}^L T_l F_l$$

es una medida de la longitud promedio de las transcripciones fonéticas, con T_l : número de vectores/etiquetas de la muestra l , F_l : longitud de la transcripción fonética de la muestra l y L : número de muestras en el corpus de entrenamiento.

Por tanto, el coste computacional del algoritmo de Baum-Welch general es aproximadamente F veces mayor que el de la versión restringida. Sin embargo, presenta algunas propiedades interesantes: (1) en el caso de MOM discretos no necesita ningún tipo de inicialización, salvo valores no nulos que verifiquen las restricciones probabilísticas —en la práctica, se suelen utilizar distribuciones uniformes—; y (2) hace un uso exhaustivo de las muestras de entrenamiento, ya que todas las observaciones acústicas sirven en alguna medida para estimar todos los parámetros, de ahí que se obtengan estimaciones óptimas y robustas.

4.2.1.2. El algoritmo de entrenamiento por Viterbi.

Si se utilizan MOM de topologías izquierda-derecha similares a la de la Figura 4.2, el coste computacional del algoritmo de entrenamiento por Viterbi es $O(E^2TF + TC)$ en el caso de MOM discretos, y $O(E^2TF + TGD)$ en el caso de MOM continuos (Rodríguez and Torres, 2003). En la práctica, dependiendo del valor de F , estas complejidades son entre uno y dos órdenes de magnitud inferiores a las del algoritmo de Baum-Welch general. Además, en la mayor parte de los casos el entrenamiento por Viterbi proporciona el mismo rendimiento que el algoritmo de Baum-Welch general. De ahí que sea el más utilizado en la estimación de MOM. Sin embargo, presenta algunos inconvenientes:

- ▷ Incluso en el caso de MOM discretos requiere una inicialización razonable, basada en datos fiables, lo cual en la práctica se consigue bien a través de modelos entrenados sobre otras bases de datos, bien aplicando el algoritmo de Baum-Welch restringido sobre una pequeña fracción del corpus de entrenamiento segmentada manualmente por expertos.
- ▷ Hace un uso muy limitado de las muestras de entrenamiento, ya que los parámetros de un cierto estado del MOM se reestiman utilizando sólo las observaciones acústicas que se hallen en los segmentos correspondientes a dicho estado, lo cual produce estimaciones posiblemente más ajustadas a los datos de entrenamiento pero menos robustas, ya que dependen fuertemente de la calidad de las segmentaciones. Cuando la cantidad de información acústica disminuye, las segmentaciones forzosamente empeoran, y por tanto también las estimaciones que de ellas se extraen. De hecho, aunque se disponga de habla limpia e información acústica suficiente, todavía se producen errores de segmentación. No obstante, en lo que respecta a las estimaciones, si el corpus de entrenamiento es lo bastante grande, los efectos de los errores de segmentación se cancelan unos con otros y prevalecen las observaciones acústicas relevantes, las que identifican cada estado de los MOM. Con habla leída a 16 kHz y el conjunto de unidades de la Tabla 4.2, se ha demostrado que las segmentaciones que implícitamente genera el algoritmo de Viterbi durante el entrenamiento de MOM discretos y semicontinuos coinciden en gran medida con las de un experto, obteniéndose unos buenos resultados en reconocimiento (Torres et-al., 1994).

- ▷ En el caso de MOM discretos, lo ajustado de las estimaciones hace que las distribuciones de emisión puedan contener probabilidades nulas, lo cual puede llegar a abortar la búsqueda de hipótesis durante el reconocimiento. Así pues, es necesario *suavizar* las distribuciones de emisión. En este trabajo se aplica la aproximación más sencilla, que consiste en definir un valor mínimo τ para las probabilidades de emisión, de manera que para todo estado q de un MOM, toda representación acústica k y todo símbolo s , la probabilidad de emisión suavizada $\hat{b}_{qk}(s)$ se calcula como sigue:

$$\hat{b}_{qk}(s) = \begin{cases} \tau & \text{si } b_{qk}(s) < \tau \\ b_{qk}(s) \frac{1-n\tau}{1-P_{res}} & \text{en caso contrario} \end{cases}$$

donde n es el número de símbolos para los que la probabilidad de emisión es menor que τ , y P_{res} es lo que llamamos *probabilidad residual*: la suma de las probabilidades de emisión menores que τ . Este tipo de suavizado sólo garantiza que la búsqueda no va a romperse debido a probabilidades de emisión nulas. Se han propuesto en la literatura técnicas de suavizado más sofisticadas, basadas en coocurrencias de símbolos, en proximidad acústica, etc.

4.2.1.3. Entrenamiento híbrido.

El algoritmo híbrido consiste en:

1. Los MOM se inicializan a partir de un pequeño conjunto de segmentos obtenidos manualmente, aplicando el algoritmo de Baum-Welch restringido.
2. Se construye el MOM de cada muestra del corpus de entrenamiento concatenando los MOM correspondientes a la transcripción fonética. A continuación, mediante el algoritmo de Viterbi, se obtiene la secuencia de estados óptima en el MOM de la muestra, de donde se extrae la segmentación óptima.
3. Los MOM se reestiman a partir de los segmentos obtenidos en el paso (2), aplicando el algoritmo de Baum-Welch restringido.
4. Los pasos (2) y (3) se repiten o bien un número fijo de veces, o bien hasta la convergencia.

Este algoritmo puede verse como una variación del entrenamiento por Viterbi. Los pasos (1) y (2) son idénticos. La diferencia está en el paso (3): en lugar de aplicar las fórmulas de reestimación propias del entrenamiento por Viterbi, se aplican las fórmulas de reestimación de Baum-Welch en su versión restringida. De hecho, también puede verse como una variación del algoritmo de Baum-Welch restringido, en la que se utiliza una segmentación automática en lugar de una segmentación manual.

La complejidad temporal de este algoritmo está en $O(E^2TF + ETC)$ en el caso de MOM discretos, y en $O(E^2TF + ETGD)$ en el caso de MOM continuos, lo cual significa que es ligeramente más costoso que el entrenamiento por Viterbi. Como contrapartida, se incrementa la robustez de las estimaciones con respecto a las obtenidas por Viterbi, ya que en la reestimación por Baum-Welch —versión restringida— se consideran todas las posibles secuencias de estados a través de cada segmento de entrenamiento, mientras que en la reestimación por Viterbi tan sólo se considera la secuencia óptima de estados.

4.2.1.4. Fórmulas de reestimación.

No es fácil encontrar en la literatura las fórmulas de reestimación de los parámetros de los MOM con observaciones múltiples, debido a que en los últimos años la aproximación dominante reúne en un único vector todas las informaciones. Hay que retroceder hasta los primeros años 90 para encontrar referencias al respecto. En (Huang et-al., 1990, páginas 216-217) tan sólo se dan las indicaciones necesarias para obtener las fórmulas de reestimación con observaciones múltiples. En (Segura, 1991) se entrenan MOM con observaciones múltiples pero sólo aparecen las fórmulas de reestimación para el caso de observaciones simples. Sólo en (Casacuberta, 1992) se ha encontrado una formalización explícita de las fórmulas de reestimación cuando se utilizan observaciones múltiples, para MOM discretos, continuos y semicontinuos, aunque se echan en falta las fórmulas correspondientes al algoritmo de Baum-Welch general. En el apéndice C se presentan las fórmulas de reestimación correspondientes a los algoritmos de Baum-Welch y Viterbi, para MOM discretos, continuos y semicontinuos. En dicho apéndice se hace tratamiento explícito de los MOM con observaciones múltiples, se tratan de forma unificada los MOM continuos y semicontinuos —en términos de MOM continuos con ligaduras de gaussianas— y se presentan explícitamente las fórmulas de reestimación para el algoritmo de Baum-Welch general.

4.2.2. Inicialización de los modelos continuos y semicontinuos.

Los MOM continuos y semicontinuos se inicializan a partir de los MOM discretos. Esto significa que para poder estimar y utilizar en reconocimiento MOM continuos y semicontinuos es necesario haber entrenado previamente MOM discretos. Las probabilidades de transición se copian tal cual del MOM discreto. En cada estado de un MOM continuo o semicontinuo se tendrán también cuatro representaciones acústicas, y cada una de ellas se modelará mediante una mezcla ponderada de gaussianas.

En el caso de los MOM semicontinuos los pesos de las componentes de la mezcla se copian de las probabilidades de emisión de los MOM discretos, y los parámetros de la gaussianas (medias y varianzas, si suponemos matrices de covarianza diagonales), de los centroides y varianzas internas obtenidos mediante cuantificación vectorial, de manera que el tamaño de las mezclas coincidirá precisamente con el tamaño de los diccionarios CV.

En el caso de los MOM continuos es necesario conocer de antemano el tamaño g de las mezclas. A continuación, en cada representación acústica de cada estado de cada MOM continuo, la mezcla se compone con las g probabilidades de emisión más altas del MOM discreto, renormalizando sus valores para que sumen 1 y copiando los correspondientes centroides y varianzas internas del diccionario CV como parámetros de las gaussianas.

4.2.3. Evaluación de los modelos acústicos.

La evaluación de los modelos acústicos se lleva a cabo mediante una aplicación que se conoce como *decodificación acústico-fonética* (DAF), que consiste en obtener la secuencia óptima de unidades subléxicas, dada una señal de entrada. Para ello se construye un modelo de reconocimiento muy simple, en el que son posibles todas las transiciones entre modelos acústicos, salvo las que se producen entre alófonos de una misma unidad, que es la única restricción de tipo fonológico. La probabilidad asignada a la transición entre dos modelos acústicos i y j es:

$$prob(i \rightarrow j) = \begin{cases} 0 & unidad(j) = unidad(i) \\ 1/n(i) & unidad(j) \neq unidad(i) \end{cases}$$

donde $n(i)$ es el número de modelos tales que $unidad(j) \neq unidad(i)$. Nótese que $\sum_{j=1}^M prob(i \rightarrow j) = 1$, con M : número de unidades subléxicas. La probabilidad de transición entre modelos acústicos podría ajustarse para mejorar el rendimiento en DAF. Sin embargo, no se van a efectuar ese tipo de ajustes, ya que con esta experimentación sólo se pretende comparar el rendimiento de distintos modelos acústicos, y escoger los mejores para integrarlos en el reconocedor.

Una vez construido el modelo, para obtener la secuencia óptima de estados —y por extensión, de unidades subléxicas— se aplica el algoritmo de Viterbi. Finalmente, la secuencia más probable de unidades subléxicas se alinea con la transcripción fonética correcta, con el criterio de minimizar el coste de las operaciones de edición. Antes de hacerlo se eliminan los silencios, ya que la secuencia reconocida contiene muchos silencios breves que no deben ser tomados como errores. De ese conjunto de alineamientos se extraen cuatro cantidades: el número de aciertos (a), el número de sustituciones (s), el número de borrados (b) y el número de inserciones (i). La tasa de aciertos se calcula según la siguiente expresión:

$$\%DAF = \frac{a}{a + s + b + i} * 100$$

4.3. Experimentos de decodificación acústico-fonética.

Uno de los principales objetivos de este trabajo es comparar el rendimiento en DAF de los modelos acústicos sobre habla leída y sobre habla espontánea. Se ha verificado el rendimiento de MOM discretos, continuos y semicontinuos, estimados mediante los algoritmos de entrenamiento descritos en el apartado 4.2.1. Precisamente, estos experimentos servirán también para decidir cuál de los algoritmos de estimación es el más adecuado —más robusto o menos costoso— en cada caso.

4.3.1. Bases de datos.

Se han realizado experimentos de DAF sobre cuatro bases de datos, dos de habla leída (SENGLAR16 y SENGLAR08) y dos de habla espontánea (INFOTREN y CORLEC-EHU-1). SENGLAR16 y SENGLAR08 son bases de datos fonéticas, diseñadas para entrenar modelos acústicos. SENGLAR16 está adquirida a 16 kHz en condiciones de laboratorio. SENGLAR08 se ha obtenido submuestreando SENGLAR16 a 8 kHz y simulando las características de un canal telefónico, precisamente con objeto de entrenar modelos acústicos aptos para aplicaciones sobre línea telefónica.

INFOTREN y CORLEC-EHU-1 ya han sido descritas en el capítulo 3. CORLEC-EHU-1 se ha dividido en tres bloques de 14 entrevistas, que llamaremos C1, C2 y C3, con una duración y un número de hablantes similares. A continuación se han definido tres particiones distintas, utilizando dos de los bloques como corpus de entrenamiento y el bloque restante como corpus de test. De esta forma se hace un uso óptimo de los datos y se obtienen resultados promediados sobre las tres particiones, lo que aumenta la fiabilidad de los mismos.

Tabla 4.3: Bases de datos de habla leída.

	SENGLAR16		SENGLAR08	
Frecuencia de muestreo (kHz)	16		8	
Entorno de grabación	Laboratorio		Laboratorio	
Canal	Micrófono		Línea telefónica (simulada)	
Diseño	Fonética		Fonética	
	Entrenamiento	Test	Entrenamiento	Test
Número de hablantes	57	33	57	26
Número de frases	1529	700	1529	493
Número de tramos	469620	244026	469626	179762
Número de fonemas	60399	32034	60399	23607
<i>F</i>	47.93	–	47.93	–

Tabla 4.4: Bases de datos de habla espontánea.

	INFOTREN		CORLEC-EHU-1		
Frecuencia de muestreo (kHz)	8		16		
Entorno de grabación	Oficina		Todos (sin restricciones)		
Canal	Línea telefónica		Grabadora analógica		
Diseño	Tarea de acceso a información		Entrevistas en radio y TV		
	Entrenamiento	Test	C1	C2	C3
Número de hablantes	63	12	37	35	33
Número de turnos	1349	308	700	690	690
Número de tramos	703719	182722	680538	683142	672884
Número de fonemas	61611	13394	89851	91685	96028
			C2+C3	C1+C3	C1+C2
<i>F</i>	81.41	–	321.43	336.27	332.98

Las principales características de estas bases de datos se presentan en las Tablas 4.3 y 4.4. En estas tablas el tamaño de cada corpus se expresa en términos del número de frases/turnos, el número de tramos —a razón de 10 milisegundos por tramo— y el número de fonemas —donde se incluyen los silencios. El parámetro F es la longitud promedio de las transcripciones fonéticas, tal como se ha definido en el apartado 4.2.1.1. En el caso de CORLEC-EHU-1, la Tabla 4.4 muestra los datos correspondientes a los tres bloques C1, C2 y C3. Los datos para las tres particiones entrenamiento/test se obtienen sumando los de dos de los bloques como corpus de entrenamiento y tomando los del tercer bloque como datos del corpus de test: $C2+C3/C1$, $C1+C3/C2$ y $C1+C2/C3$. Sólo en el caso del parámetro F se muestran explícitamente los valores correspondientes a los tres corpus de entrenamiento. Nótese que el valor de F es mucho mayor para CORLEC-EHU-1 que para las otras bases de datos. Esto tendrá como consecuencia que el coste del algoritmo de Baum-Welch general sea muy superior al del algoritmo de Viterbi, sobre todo en el caso de MOM continuos.

4.3.1.1. Distribución de fonemas.

En la Tabla 4.5 se muestra la distribución de unidades subléxicas asimilables a fonemas —es decir, todas salvo el silencio—, en porcentaje y en números absolutos, obtenida para el corpus de entrenamiento de SENGLAR —la misma a 16 y 8 kHz—, así como las distribuciones obtenidas para los corpus de entrenamiento de las dos bases de datos de habla espontánea. No hay grandes diferencias en cuanto a las cifras relativas. Las vocales suman en todos los casos algo más del 47 % de los fonemas, las consonantes oclusivas alrededor del 19 %, las nasales el 10 %, las fricativas otro 10 % y las líquidas alrededor del 12 %. De unas bases de datos a otras sólo se observan en general pequeños movimientos en la distribución interna de los fonemas dentro de cada categoría. Hay un grupo de fonemas con frecuencias inferiores al 1 % (tS , f , x , J , L , rr), que pueden ser problemáticos a la hora de estimar sus correspondientes modelos acústicos, debido a la falta de muestras, especialmente en los casos de SENGLAR e INFOTREN. No obstante, aún suponiendo que los modelos estén infraentrenados, dado que en conjunto suman menos del 3 % de los fonemas, su efecto sobre las tasas de reconocimiento no puede ser importante.

Lo relevante, desde el punto de vista metodológico, es que parece posible entrenar MOM directamente a partir de un corpus de entrenamiento de habla espontánea —utilizando el algoritmo de Baum-welch general— ya que se tienen muestras suficientes como para hacerlo, al menos tantas como en bases de datos fonéticas, y con la misma distribución. Por otra parte, mientras que SENGLAR16 está adquirida en condiciones de laboratorio y SENGLAR08 sólo presenta un ligero ruido de canal, las bases de datos de habla espontánea —especialmente CORLEC-EHU-1— contienen numerosos ruidos de fondo y de canal. Por tanto, al tratarse de bases de datos con numerosas distorsiones acústicas, podría argumentarse que los modelos resultantes no serían tan precisos como los obtenidos sobre bases de datos acústicamente *limpias*. Sin embargo, los modelos entrenados sobre SENGLAR16 o SENGLAR08 no pueden aplicarse directamente sobre el habla espontánea de CORLEC-EHU-1 o INFOTREN. Es necesario un proceso de adaptación, ya que las condiciones ambientales y de canal son muy distintas. En este caso, la alternativa consiste en recurrir a modelos entrenados previamente sobre habla leída *únicamente* para inicializar los modelos de habla espontánea. Después, en un segundo paso, seguiría siendo necesario reestimar éstos sobre el corpus de entrenamiento de habla espontánea. Se ha demostrado que ambos caminos —inicialización uniforme e inicialización con modelos entrenados sobre habla leída— conducen a resultados similares (Jeanrenaud et al., 1995).

Tabla 4.5: Distribución de fonemas en los corpus de entrenamiento.

	CORLEC-EHU-1				
	SENGLAR	INFOTREN	C2+C3	C1+C3	C1+C2
p	1514 (2.67 %)	1331 (2.18 %)	5631 (3.02 %)	5615 (3.04 %)	5520 (3.06 %)
b	1567 (2.76 %)	2002 (3.28 %)	4394 (2.35 %)	4308 (2.33 %)	4106 (2.28 %)
t	2481 (4.37 %)	2432 (3.98 %)	8985 (4.81 %)	8816 (4.77 %)	8669 (4.81 %)
d	2565 (4.52 %)	3307 (5.42 %)	8842 (4.73 %)	8764 (4.74 %)	8586 (4.76 %)
k	2147 (3.78 %)	1638 (2.68 %)	8190 (4.39 %)	8098 (4.38 %)	7862 (4.36 %)
g	710 (1.25 %)	916 (1.50 %)	1669 (0.89 %)	1737 (0.94 %)	1672 (0.93 %)
tS	250 (0.44 %)	207 (0.34 %)	494 (0.26 %)	478 (0.26 %)	474 (0.26 %)
f	454 (0.80 %)	178 (0.29 %)	1046 (0.56 %)	1082 (0.59 %)	1062 (0.59 %)
T	960 (1.69 %)	1557 (2.55 %)	3178 (1.70 %)	3151 (1.70 %)	3069 (1.70 %)
s	3823 (6.74 %)	4284 (7.02 %)	15160 (8.12 %)	14716 (7.96 %)	14304 (7.93 %)
x	315 (0.56 %)	260 (0.43 %)	1068 (0.57 %)	1068 (0.58 %)	1018 (0.56 %)
m	1966 (3.46 %)	1596 (2.61 %)	6266 (3.36 %)	6117 (3.31 %)	5997 (3.32 %)
n	3816 (6.72 %)	3304 (5.41 %)	13810 (7.39 %)	13790 (7.46 %)	13280 (7.36 %)
J	153 (0.27 %)	155 (0.25 %)	489 (0.26 %)	480 (0.26 %)	401 (0.22 %)
l	2971 (5.24 %)	3065 (5.02 %)	8121 (4.35 %)	7988 (4.32 %)	8023 (4.45 %)
L	439 (0.77 %)	338 (0.55 %)	1158 (0.62 %)	1171 (0.63 %)	1013 (0.56 %)
r	3416 (6.02 %)	5200 (8.52 %)	10004 (5.36 %)	10032 (5.43 %)	10040 (5.57 %)
rr	402 (0.71 %)	161 (0.26 %)	801 (0.43 %)	778 (0.42 %)	815 (0.45 %)
i	3730 (6.57 %)	5318 (8.71 %)	13311 (7.13 %)	13325 (7.21 %)	13012 (7.21 %)
e	7781 (13.71 %)	9566 (15.67 %)	27203 (14.57 %)	26962 (14.59 %)	26301 (14.58 %)
a	7859 (13.85 %)	8215 (13.46 %)	21378 (11.45 %)	21102 (11.42 %)	20618 (11.43 %)
o	5450 (9.60 %)	4688 (7.68 %)	19164 (10.26 %)	18953 (10.25 %)	18309 (10.15 %)
u	1977 (3.48 %)	1334 (2.19 %)	6393 (3.42 %)	6295 (3.41 %)	6222 (3.45 %)

Tabla 4.6: Distorsión final de los diccionarios CV correspondientes a las bases de datos adquiridas a 8 kHz. Entre paréntesis se indica el número de iteraciones empleadas para construirlos.

	SENGLAR08	INFOTREN
Cepstrales	30.1899 (295)	32.4161 (269)
Δ Cepstrales	2.9886 (285)	2.4902 (311)
Δ^2 Cepstrales	1.1029 (276)	$9.6564 \cdot 10^{-1}$ (245)
E+ Δ E	$1.8389 \cdot 10^{-4}$ (321)	$1.7878 \cdot 10^{-4}$ (292)

Tabla 4.7: Distorsión final de los diccionarios CV correspondientes a las bases de datos adquiridas a 16 kHz. Entre paréntesis se indica el número de iteraciones empleadas para construirlos.

	SENGLAR16	CORLEC-EHU-1		
		C2+C3	C1+C3	C1+C2
Cepstrales	71.2613 (263)	49.4462 (258)	49.5196 (300)	48.8704 (250)
Δ Cepstrales	6.6172 (299)	4.3050 (290)	4.2883 (352)	4.1785 (291)
Δ^2 Cepstrales	2.3865 (272)	1.6460 (276)	1.6321 (316)	1.5963 (304)
E+ Δ E	$1.9914 \cdot 10^{-4}$ (322)	$1.1058 \cdot 10^{-4}$ (303)	$1.1516 \cdot 10^{-4}$ (284)	$1.1348 \cdot 10^{-4}$ (336)

4.3.1.2. Cuantificación vectorial.

Al utilizar MOM discretos, la información acústica sufre una distorsión importante debido al proceso de cuantificación vectorial, ya que cada vector de parámetros es sustituido por el centroide más cercano de un diccionario CV. Obviamente, cuanto mayor sea este diccionario, más *cerrada* será la cuadrícula definida por los centroides y menor la distorsión introducida al etiquetar los vectores, aunque también será más costoso el proceso de etiquetado y menos robusta la estimación de las probabilidades de emisión de los MOM. Así pues, el mejor rendimiento de los MOM discretos se obtendrá de un compromiso entre la precisión del etiquetado y la robustez de las estimaciones. En base a la experiencia se ha fijado el tamaño de los diccionarios a 256 centroides.

Todos los diccionarios CV —para todas las representaciones acústicas y para todas las bases de datos— se han construido con el mismo método de duplicación de los centroides, el mismo criterio de convergencia y el mismo número final de centroides. Lo que cambia de unas bases de datos a otras es la dimensión de los vectores de cepstrales y el número de vectores del corpus de entrenamiento (véanse las Tablas 4.3 y 4.4). En las Tablas 4.6 y 4.7 se presenta la distorsión final de los diccionarios CV obtenidos en cada representación acústica para cada base de datos, así como el número total de iteraciones empleadas para construirlos.

La distorsión final de un diccionario —definida como la distancia promedio de un vector a su centroide, calculada para el conjunto de vectores utilizado en la construcción del mismo— permite predecir hasta cierto punto el rendimiento de los MOM discretos, ya que determina la pérdida de información acústica. Pero la calidad de un diccionario —y la de los modelos basados en el mismo— no puede evaluarse únicamente en términos de la distorsión que introduce. Hay que tener en cuenta las propiedades de los vectores de parámetros sobre los que está construido. Así, unas condiciones de grabación adversas podrían *ecualizar* los vectores de parámetros y disminuir su capacidad de distinguir entre los sonidos del habla. Como consecuencia, la distorsión

final del diccionario CV sería menor, ya que los vectores del corpus de entrenamiento estarían *más cerca* unos de otros, pero ello sólo indicaría una pérdida de capacidad discriminativa. Este efecto se pone de manifiesto especialmente en el caso de CORLEC-EHU-1, si lo comparamos con SENGLAR16. Más adelante en este capítulo se comprobará cómo a pesar de que los diccionarios CV muestran distorsiones notablemente inferiores, las tasas de reconocimiento con MOM discretos serán mucho peores para CORLEC-EHU-1 que para SENGLAR16. En estos casos es necesario mejorar el preproceso de la señal, con objeto de obtener parámetros más inmunes a las condiciones ambientales. En el caso de INFOTREN, las condiciones ambientales no difieren demasiado de las de SENGLAR08. Son sólo ligeramente más ruidosas, de ahí que se obtengan distorsiones muy similares. La presencia de fenómenos extralingüísticos — ruidos aislados, saturaciones— puede explicar la mayor distorsión final del diccionario de los cepstrales.

4.3.1.3. Matrices de covarianza.

Otro de los aspectos importantes de las bases de datos es comprobar la hipótesis de que las matrices de covarianza son diagonales para cada una de las representaciones acústicas. Esta hipótesis se basa en la presunción de independencia estadística entre las distintas componentes de los vectores acústicos, y de ella depende la corrección de los MOM continuos que se utilizan en este trabajo.

En cada representación acústica de cada base de datos se han considerado todos los vectores del corpus de entrenamiento y se ha calculado la matriz de correlación. Cada elemento ρ_{xy} de dicha matriz es el coeficiente de correlación entre las componentes x e y , y se obtiene del elemento correspondiente de la matriz de covarianza c_{xy} dividido por las desviaciones típicas de las respectivas componentes: $\rho_{xy} = \frac{c_{xy}}{\sigma_x \sigma_y}$. Así, en la diagonal se tiene que $\rho_{xx} = 1$ y fuera de la diagonal se obtienen valores en el rango $[-1, 1]$. La hipótesis de independencia estadística implica que los coeficientes ρ_{xy} son todos nulos para $x \neq y$. En la práctica se obtienen coeficientes no nulos pero relativamente pequeños en valor absoluto (véanse las Tablas 4.8 y 4.9 y la Figura 4.3).

Las correlaciones más altas se obtienen entre los coeficientes cepstrales, con máximos que en los peores casos llegan a 0.5, pero con medias entre 0.11 y 0.16. Esto significa que en el caso de los cepstrales las covarianzas de la diagonal son en promedio entre 6 y 9 veces más grandes que las covarianzas de fuera de la diagonal. Por otra parte, las correlaciones entre las derivadas primeras son más pequeñas que entre los cepstrales, y las correlaciones entre las derivadas segundas son, a su vez, más pequeñas que entre las derivadas primeras. Pero las correlaciones más pequeñas —casi nulas— son las que se obtienen entre la energía y su derivada.

Es destacable el hecho de que en todas las representaciones acústicas las correlaciones más altas se obtienen sobre SENGLAR08. A priori deberían ser las bases de datos *más ruidosas* las que mostrarán una peor adecuación de los parámetros. Sin embargo, las correlaciones más bajas se dan para INFOTREN en el caso de los cepstrales y sus derivadas, y para CORLEC-EHU-1 en el caso de $E + \Delta E$. En todos los casos, los resultados obtenidos indican que la hipótesis de independencia estadística entre las componentes de los vectores acústicos —y la simplificación de las distribuciones normales que ello implica— es *aproximadamente correcta*.

Tabla 4.8: Valor medio, desviación típica, máximo y mínimo del valor absoluto de los elementos fuera de la diagonal en las matrices de correlación obtenidas para los coeficientes cepstrales y sus derivadas primera y segunda.

	Cepstrales			
	Media	Desviación	Máximo	Mínimo
SENGLAR16	0.1175	0.0872	0.3262	0.0007
SENGLAR08	0.1626	0.1357	0.4939	0.0018
INFOTREN	0.1154	0.0743	0.3672	0.0117
C2+C3	0.1319	0.0982	0.4636	0.0003
C1+C3	0.1388	0.1066	0.5037	0.0004
C1+C2	0.1333	0.1089	0.4935	0.0046

	Δ Cepstrales			
	Media	Desviación	Máximo	Mínimo
SENGLAR16	0.1044	0.0826	0.3611	0.0032
SENGLAR08	0.1240	0.1164	0.4240	0.0001
INFOTREN	0.0689	0.0500	0.2291	0.0030
C2+C3	0.1065	0.0830	0.3609	0.0004
C1+C3	0.1151	0.0895	0.4140	0.0001
C1+C2	0.1142	0.0880	0.3983	0.0032

	Δ^2 Cepstrales			
	Media	Desviación	Máximo	Mínimo
SENGLAR16	0.0935	0.0776	0.3231	0.0002
SENGLAR08	0.1082	0.1018	0.3831	0.0029
INFOTREN	0.0624	0.0539	0.1945	0.0004
C2+C3	0.0894	0.0631	0.2549	0.0014
C1+C3	0.0953	0.0687	0.3134	0.0027
C1+C2	0.0936	0.0702	0.2914	0.0024

Tabla 4.9: Coeficiente de correlación entre energía y derivada de la energía.

	$\rho_{E\Delta E}$
SENGLAR16	0.0013
SENGLAR08	0.0034
INFOTREN	0.0009
C2+C3	0.0003
C1+C3	0.0002
C1+C2	0.0001

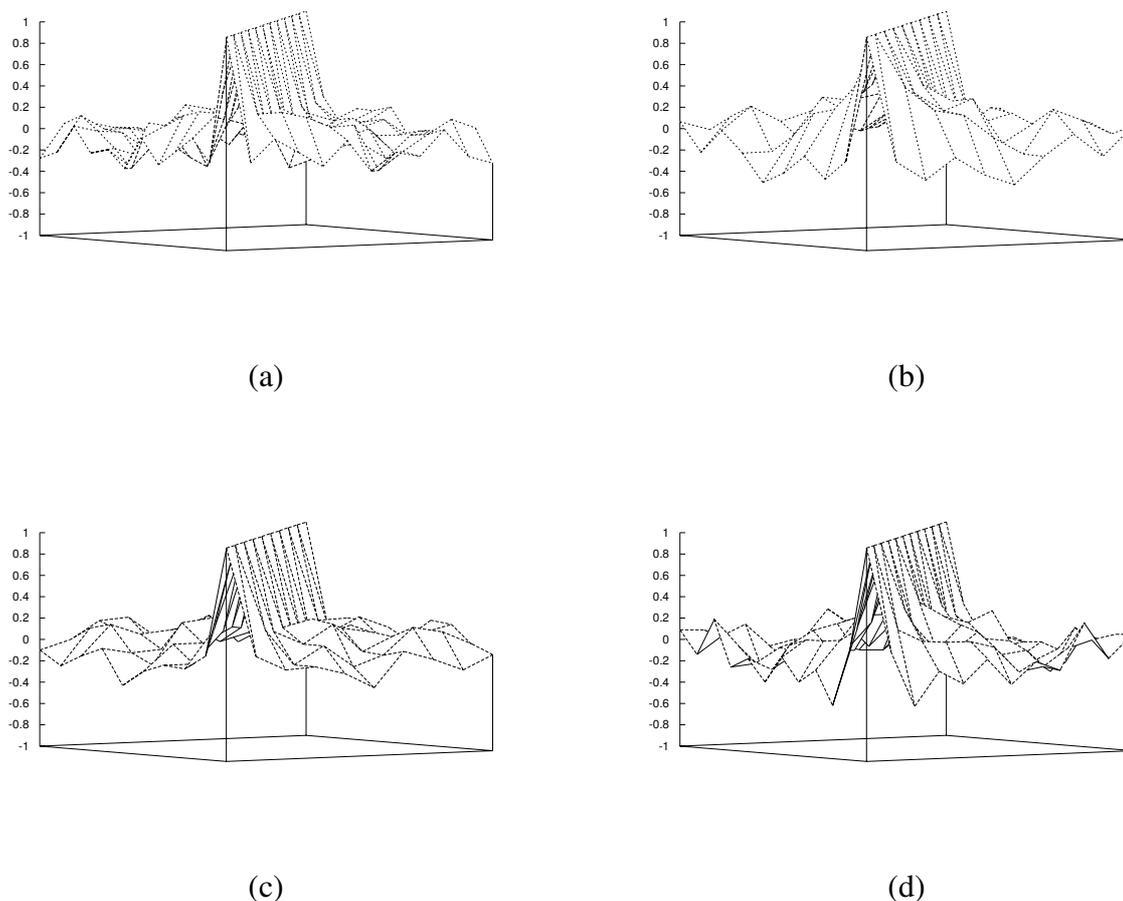


Figura 4.3: Matrices de correlación de los coeficientes cepstrales, calculadas sobre los conjuntos de entrenamiento de SENGLAR16 (a), SENGLAR08 (b), INFOTREN (c) y la partición 2 de CORLEC-EHU-1 (d).

4.3.2. Resultados con MOM discretos.

En el caso de MOM discretos, se han comparado dos metodologías de entrenamiento. La primera de ellas consiste en utilizar una fracción del corpus de entrenamiento segmentada manualmente —con suficientes muestras de cada unidad— para generar unos modelos iniciales, típicamente mediante 6 iteraciones del algoritmo de Baum-Welch restringido (en adelante, BWr); a continuación, se utiliza el corpus de entrenamiento completo para generar los modelos definitivos, aplicando el algoritmo de entrenamiento de Viterbi o, alternativamente, el algoritmo híbrido descrito en el apartado 4.2.1.3, a partir de los modelos iniciales obtenidos previamente. La segunda metodología consiste en inicializar los MOM con distribuciones uniformes y aplicar sobre el corpus de entrenamiento completo el algoritmo de Baum-Welch general (en adelante, BWg). En realidad, el algoritmo BWg puede aplicarse también sobre modelos iniciales aleatorios o sobre modelos entrenados mediante el algoritmo BWr. Se utilizan modelos equiprobables porque es la aproximación más sencilla. En ambos casos, se ejecuta un número fijo de iteraciones, típicamente entre 12 y 20, que suelen ser suficientes para alcanzar la convergencia, y se obtiene la tasa de DAF de los modelos resultantes de cada iteración frente a un conjunto de

Tabla 4.10: Resultados de DAF sobre SENGLAR16 y SENGLAR08, utilizando MOM discretos obtenidos mediante los algoritmos de Baum-Welch general, Viterbi e híbrido.

	Baum-Welch general	Viterbi	Híbrido
SENGLAR16	65.28	65.36	65.35
SENGLAR08	62.70	57.70	59.15

muestras independiente del corpus de entrenamiento, que llamamos *corpus de test*. Finalmente, para los experimentos de reconocimiento se escogen los modelos acústicos que proporcionen la tasa de DAF más alta sobre el corpus de test.

4.3.2.1. MOM discretos sobre habla leída.

En la Tabla 4.10 se muestran las tasas óptimas de DAF obtenidas con MOM discretos sobre habla leída. En el caso de SENGLAR16, los tres algoritmos de entrenamiento proporcionan el mismo rendimiento. El algoritmo híbrido necesita la misma inicialización que el algoritmo de Viterbi pero no aporta mejoras, y tiene un coste temporal entre dos y tres veces mayor. El algoritmo BWg parte de modelos equiprobables, pero tiene en la práctica un coste 15 veces mayor que el de Viterbi —en este caso, teniendo en cuenta la topología del modelo, el tamaño de los diccionarios CV y el valor de F , el coste del algoritmo BWg se había estimado unas 50 veces mayor que el de Viterbi.

En el caso de SENGLAR08, los modelos obtenidos mediante los tres métodos de entrenamiento proporcionan tasas de DAF significativamente distintas. Cabe suponer que la sensibilidad de las estimaciones de Viterbi a la calidad de las segmentaciones está teniendo efecto, ya que los modelos obtenidos mediante el algoritmo híbrido —que suaviza hasta cierto punto los efectos de la segmentación— dan un rendimiento notablemente mejor que los de Viterbi, aunque lejos de los mejores, que se obtienen mediante el algoritmo BWg. La diferencia entre el algoritmo de Viterbi y éste último es de 5 puntos en términos absolutos. En cuanto a la complejidad temporal, en este caso los algoritmos BWg e híbrido resultan 16 y 1.5 veces más costosos que el algoritmo de Viterbi, respectivamente.

¿Cuánto han empeorado las segmentaciones para que el rendimiento del algoritmo de Viterbi disminuya tan notablemente? Para responder a esta pregunta se han comparado las segmentaciones que se producen implícitamente durante el entrenamiento de los modelos —algoritmos de Viterbi e híbrido— con las generadas manualmente por expertos. Sólo 162 de las 1529 frases que componen el corpus de entrenamiento disponen de segmentación manual, pero creemos que son suficientes desde un punto de vista estadístico. En realidad, lo que se hace es segmentar este subconjunto aplicando una iteración del algoritmo de entrenamiento de Viterbi a partir de unos ciertos modelos, lo que se conoce como *alineamiento forzado*. Sea n_b el número de tramos que han sido asignados en alineamiento forzado al mismo segmento que los expertos, y n_m el número de tramos asignados a un segmento *equivocado*. Entonces:

$$S = \frac{n_b}{n_b + n_m} * 100$$

es el porcentaje de tramos asignados al segmento *correcto*. Es decir, en el caso de los algoritmos de Viterbi e híbrido, S puede interpretarse como el porcentaje de información acústica que sería *correctamente utilizada* en la estimación de los parámetros.

Tabla 4.11: *Calidad* de las segmentaciones automáticas obtenidas a partir de MOM discretos en SENGLAR16 y SENGLAR08.

	Viterbi	Híbrido	Baum-Welch general
SENGLAR16	89.72	89.64	87.98
SENGLAR08	58.86	55.03	41.67

Tabla 4.12: Tasas óptimas de DAF obtenidas sobre INFOTREN y las tres particiones de CORLEC-EHU-1, utilizando MOM discretos y aplicando los algoritmos Baum-Welch general, Viterbi e híbrido.

	Baum-Welch general	Viterbi	Híbrido
INFOTREN	51.98	51.92	51.85
CORLEC-EHU-1: C2+C3/C1	44.73	44.61	44.77
CORLEC-EHU-1: C1+C3/C2	45.76	45.32	45.57
CORLEC-EHU-1: C1+C2/C3	46.30	46.19	46.40

Se han aplicado los modelos óptimos obtenidos por Viterbi y por el algoritmo híbrido, y también, para completar el cuadro, los modelos óptimos obtenidos mediante el algoritmo BWg. En la Tabla 4.11 se pone de manifiesto que la calidad de las segmentaciones generadas implícitamente por los algoritmos de Viterbi e híbrido pasa de casi el 90 % con habla leída a 16 kHz a menos del 60 % con habla leída a 8 kHz. Es destacable el hecho de que a 16 kHz los tres tipos de modelos generan segmentaciones de calidad similar, mientras que a 8 kHz se observan notables diferencias, especialmente entre los modelos entrenados por Viterbi ($S = 58,86$) y los entrenados mediante el algoritmo BWg ($S = 41,67$). Lo más interesante es que éstos últimos, sin ser capaces de segmentar correctamente, sí ofrecen el mejor rendimiento.

4.3.2.2. MOM discretos sobre habla espontánea.

En los experimentos de DAF sobre habla espontánea los MOM discretos se inicializan con modelos entrenados sobre habla leída. En el caso de INFOTREN, para obtener los modelos iniciales, en primer lugar se ha etiquetado el corpus de entrenamiento de SENGLAR08 con el diccionario CV de INFOTREN, a continuación se han ejecutado 6 iteraciones del algoritmo BWg —sobre la parte del corpus de entrenamiento de SENGLAR08 segmentada a mano—, y por último, 12 iteraciones del algoritmo híbrido —sobre el corpus de entrenamiento de SENGLAR08. Una vez obtenidos los modelos iniciales, se comprueba el rendimiento de los tres algoritmos de entrenamiento: BWg, Viterbi e híbrido, utilizando los corpus de entrenamiento y test de INFOTREN. El mismo procedimiento se aplica sobre las tres particiones de CORLEC-EHU-1, con modelos iniciales entrenados sobre SENGLAR16 —lo que implica tres nuevos etiquetados basados en los diccionarios CV de CORLEC-EHU-1. En la Tabla 4.12 se muestran las tasas óptimas de DAF obtenidas para INFOTREN y CORLEC-EHU-1.

En el caso de INFOTREN, los tres algoritmos proporcionan tasas de DAF cercanas al 52 %, sólo que para esta base de datos cada iteración de BWg resulta 14.4 veces más costosa que una iteración de Viterbi —teóricamente, teniendo en cuenta el valor de F , debería ser unas 60 veces más costosa— y cada iteración del algoritmo híbrido, 2.3 veces más costosa que una de Viterbi. Nótese que las tasas caen más de 10 puntos con respecto a las obtenidas sobre SENGLAR08. Ello se debe por un lado a la presencia de fenómenos acústicos, como alargamientos, pausas

habladas, ruidos, etc. que no se han modelado, y por otro, a la mayor variabilidad acústica intrínseca del habla espontánea, que pone de manifiesto o acentúa efectos de coarticulación, características propias del hablante, etc.

También en el caso de CORLEC-EHU-1 los tres algoritmos tienen rendimientos similares, aunque las tasas de DAF varían de una partición a otra. El promedio se sitúa en 45.6 %, que corresponde aproximadamente a la tasa obtenida para la segunda partición. Al igual que en SENGLAR16, las señales de CORLEC-EHU-1 están muestreadas a 16 kHz, pero las tasas se sitúan casi 20 puntos por debajo. Ello pone de manifiesto, al igual que en INFOTREN, la mayor variabilidad del habla espontánea y la presencia de fenómenos acústicos no modelados, pero sobre todo las condiciones ambientales y de canal, que son particularmente difíciles para esta base de datos. Por último, en lo que se refiere al coste temporal, teniendo en cuenta los valores de F , el algoritmo BWg debería ser alrededor de 80 veces más costoso que el de Viterbi. Sin embargo, en la práctica sólo es 12 veces más costoso. Por su parte, el algoritmo híbrido resulta en promedio 1.3 veces más costoso que el algoritmo de Viterbi.

4.3.2.3. Algunas conclusiones sobre MOM discretos.

Teniendo en cuenta la mayor robustez mostrada por el algoritmo BWg en todas las situaciones, el hecho de que no requiere ningún tipo especial de inicialización, que proporciona probabilidades de emisión suavizadas de manera natural, y que cada iteración de entrenamiento no lleva en la práctica tanto tiempo como cabía esperar de la complejidad asintótica —dependiendo de la CPU y del tamaño del corpus de entrenamiento, cada iteración puede llevar entre 10 minutos y 2 horas de tiempo real—, creemos que es la opción más adecuada para entrenar MOM discretos. Por contra, el algoritmo de Viterbi, aunque requiere tiempos de entrenamiento muy inferiores, puede mostrar un comportamiento subóptimo en determinadas condiciones, como muestran los resultados sobre SENGLAR08. Además, necesita una buena inicialización y requiere aplicar un método de suavizado de las probabilidades de emisión. De hecho, los malos resultados obtenidos sobre SENGLAR08 podrían estar motivados en parte por una inicialización no lo bastante robusta.

En lo que se refiere al tiempo que lleva obtener las decodificaciones, sobre una CPU AMD Athlon a 1 GHz los tiempos están comprendidos entre 0.02 y 0.05 veces tiempo real. Esto hace de los MOM discretos una buena alternativa en ciertas aplicaciones no demasiado complejas, especialmente cuando se necesita operar en tiempo real. Por ello, en los capítulos siguientes se mantendrá una línea de mejora de los MOM discretos.

4.3.3. Resultados con MOM continuos y semicontinuos.

4.3.3.1. MOM continuos y semicontinuos sobre habla leída.

Los MOM continuos y semicontinuos se inicializan a partir de los MOM discretos óptimos, tal como se explica en el apartado 4.2.2. A continuación se aplica un número fijo de iteraciones del algoritmo de estimación. Finalmente, se evalúan los modelos resultantes de cada iteración sobre el corpus de test y se escogen como modelos óptimos aquéllos que dan una tasa de DAF más alta. Se ha comparado de nuevo el rendimiento de los distintos métodos de estimación de parámetros, Baum-Welch general, Viterbi e híbrido, con objeto de determinar cuál de ellos muestra un mejor comportamiento con MOM continuos, sobre habla leída y sobre habla espontánea.

Tabla 4.13: Tasas de DAF óptimas obtenidas sobre SENGLAR16, utilizando MOM semicontinuos y continuos de 8, 16, 32 y 64 gaussianas.

	BW general	Viterbi	Híbrido
MOM semicontinuos	69.19	69.27	69.17
MOM continuos 8g	69.79	70.01	69.83
MOM continuos 16g	71.37	71.56	71.46
MOM continuos 32g	72.47	72.38	72.42
MOM continuos 64g	72.88	72.85	72.90

Tabla 4.14: Tasas de DAF óptimas obtenidas sobre SENGLAR08, utilizando MOM semicontinuos y continuos de 8, 16, 32 y 64 gaussianas.

	BW general	Viterbi	Híbrido
MOM semicontinuos	–	63.60	63.59
MOM continuos 8g	65.71	65.86	65.77
MOM continuos 16g	67.12	67.10	67.13
MOM continuos 32g	68.30	68.27	68.21
MOM continuos 64g	68.67	68.60	68.60

En las Tablas 4.13 y 4.14 se muestran los resultados para SENGLAR16 y SENGLAR08, respectivamente, con MOM semicontinuos de 256 gaussianas y MOM continuos de 8, 16, 32 y 64 gaussianas por estado y representación acústica. Las tasas de DAF con MOM continuos de 64 gaussianas alcanzan valores cercanos al 73 % en el caso de SENGLAR16, y se sitúan claramente por encima del 68 % en el caso de SENGLAR08. Los MOM semicontinuos, cuya estimación supone el mayor coste temporal de todos los casos (aproximadamente 4 veces superior al coste de estimación de los MOM continuos con mezclas de 64 gaussianas), no superan el rendimiento de los MOM continuos, debido, probablemente, a que no es posible definir un gran número de distribuciones de probabilidad, todas ellas con la suficiente precisión, con tan sólo 256 gaussianas compartidas por todos los estados.

Por otra parte, no se observan diferencias significativas en el rendimiento de los tres algoritmos de entrenamiento. Sí las hay en lo que se refiere al coste temporal: en promedio, los algoritmos BWg e híbrido resultan 15 y 2 veces más costosos que el algoritmo de Viterbi, respectivamente. Nuevamente, es significativo que, teóricamente, teniendo en cuenta el valor de F y dependiendo del número de gaussianas, el algoritmo BWg debería ser entre 60 y 100 veces más costoso que el de Viterbi. Por último, aunque las mejores tasas se obtienen con MOM de 64 gaussianas, su coste duplica tanto en entrenamiento como en test al de los MOM de 32 gaussianas. Como muestra, en una CPU AMD Athlon a 1 GHz el coste de decodificación con MOM de 32 y 64 gaussianas es 0.71 y 1.34 veces tiempo real, respectivamente. Teniendo en cuenta estos resultados, en adelante se optará por los MOM continuos de 32 gaussianas como un buen compromiso entre coste computacional y rendimiento en DAF.

4.3.3.2. MOM continuos sobre habla espontánea.

Teniendo en cuenta que el rendimiento de los MOM semicontinuos sobre habla leída es muy inferior al de los MOM continuos, en los experimentos sobre habla espontánea sólo se han

Tabla 4.15: Tasas de DAF óptimas obtenidas sobre INFOTREN y CORLEC-EHU-1, utilizando MOM continuos de 8, 16, 32 y 64 gaussianas, y aplicando el algoritmo de entrenamiento de Viterbi.

	CORLEC-EHU-1			
	INFOTREN	C2+C3/C1	C1+C3/C2	C1+C2/C3
MOM continuos 8g	53.87	47.84	49.21	49.98
MOM continuos 16g	55.55	48.97	50.14	50.96
MOM continuos 32g	56.86	49.66	51.05	51.78
MOM continuos 64g	58.12	50.29	51.73	52.54

aplicado éstos últimos. Por otra parte, los parámetros de los modelos se han estimado mediante el algoritmo de Viterbi, ya que, como se ha comprobado sobre habla leída, los modelos resultantes se muestran tan robustos como los obtenidos mediante el algoritmo BWg, y su coste es notablemente inferior.

En el caso de INFOTREN, la utilización de MOM continuos sitúa las tasas de DAF ligeramente por debajo del 54 % con mezclas de 8 gaussianas, y por encima del 58 % con mezclas de 64 gaussianas (véase la Tabla 4.15). En el caso de CORLEC-EHU-1, la tasa de DAF promedio es del 49 % con mezclas de 8 gaussianas, y del 51.5 % con mezclas de 64 gaussianas. En lo que respecta al coste de decodificación, en el caso de INFOTREN, sobre una CPU AMD Athlon a 1 GHz, los MOM continuos con 8, 16, 32 y 64 gaussianas operan en 0.37, 0.62, 1.14 y 2.16 veces tiempo real, respectivamente. En el caso de CORLEC-EHU-1, el coste de decodificación sobre una CPU AMD Athlon a 1.6 GHz, utilizando MOM continuos con 8, 16, 32 y 64 gaussianas, está en 0.22, 0.39, 0.73 y 1.40 veces tiempo real, respectivamente. Así pues, dependiendo de la capacidad del procesador, del grado de precisión requerido y de la necesidad o no de operar en tiempo real, se utilizarán más o menos gaussianas por mezcla. En los capítulos siguientes, con objeto de reducir la experimentación, se ha optado por utilizar MOM continuos de 32 gaussianas, que dan un rendimiento razonable y permiten que los prototipos actuales del GRFyTH operen en tiempo real.

A pesar de estas mejoras, las tasas de INFOTREN siguen estando 10 puntos por debajo de las de SENGLAR08, y las de CORLEC-EHU-1 alrededor de 20 puntos por debajo de las de SENGLAR16 (véase la Tabla 4.16, en la que se resumen los resultados de DAF). Estas diferencias están motivadas en parte por las condiciones ambientales y de canal, que reducen la capacidad discriminativa de los parámetros acústicos, en parte por la presencia de fenómenos de habla espontánea de tipo acústico y léxico que no han sido modelados, y en parte por la mayor variabilidad acústica, de pronunciaciones y de hablantes. En los capítulos 5 y 6 se describirán las estrategias adoptadas para mejorar la calidad de los modelos acústicos, atendiendo a algunas de las fuentes de variabilidad mencionadas.

4.4. Experimentos de reconocimiento.

Para verificar el rendimiento de los modelos acústicos en reconocimiento de habla leída continua, se dispone de un corpus de aplicación para el acceso a una supuesta base de datos geográfica, llamado BDGEO. Este corpus forma parte de la base de datos de habla leída en castellano Albayzin (Casacuberta et-al., 1991; Moreno et-al., 1993; Díaz et-al., 1998). Para generar el modelo de lenguaje se utiliza un conjunto de 8262 frases con 81620 palabras. El

Tabla 4.16: Resumen de las tasas de DAF obtenidas sobre habla leída y habla espontánea, utilizando MOM discretos, semicontinuos y continuos de 8, 16, 32 y 64 gaussianas. En el caso de los MOM discretos, se han escogido las tasas obtenidas de las estimaciones por BWg. En el resto de casos, las tasas corresponden a las estimaciones por Viterbi. En el caso de CORLEC-EHU-1, se muestra la tasa promedio sobre las tres particiones.

	SENGLAR16	SENGLAR08	INFOTREN	CORLEC-EHU-1
MOM discretos	65.28	62.70	51.98	45.60
MOM semicontinuos	69.27	63.60	–	–
MOM continuos 8g	70.01	65.86	53.87	49.01
MOM continuos 16g	71.56	67.10	55.55	50.02
MOM continuos 32g	72.38	68.27	56.86	50.83
MOM continuos 64g	72.85	68.60	58.12	51.52

Tabla 4.17: Tasas de DAF sobre BDGEO-16 y BDGEO-8, utilizando MOM discretos y MOM continuos de 32 gaussianas.

	BDGEO-16	BDGEO-8
MOM discretos	64.44	62.93
MOM continuos 32g	71.66	67.73

corpus de test consta de 600 frases con 5655 palabras. El vocabulario, que cubre ambos corpus, contiene 1212 palabras. Se dispone de dos versiones de este corpus, a 16 y 8 kHz —en las mismas condiciones que SENGLAR16 y SENGLAR08—, que denominaremos BDGEO-16 y BDGEO-8, respectivamente. En el caso de BDGEO-8, sólo se dispone de un subconjunto de 522 frases del corpus de test, que contiene 4812 palabras. Como referencia, en la Tabla 4.17 se muestran las tasas de DAF obtenidas sobre BDGEO-16 y BDGEO-8, utilizando MOM discretos y MOM continuos de 32 gaussianas —estimados a partir de los corpus de entrenamiento de SENGLAR16 y SENGLAR08.

En el caso de INFOTREN, para entrenar el modelo de lenguaje y evaluar el sistema de reconocimiento se han utilizado los mismos conjuntos de entrenamiento y test utilizados en DAF. El corpus de entrenamiento consta de 191 diálogos, con un total de 1349 turnos, que contienen 15367 palabras y 571 silencios. El corpus de test consta de 36 diálogos, con 308 turnos que contienen 3311 palabras y 182 silencios. El vocabulario consta de 1909 palabras.

Como ya se ha indicado, CORLEC-EHU-1 consta de entrevistas tomadas de radio y televisión, de temática dispar, por lo que los tres bloques definidos, C1, C2 y C3, no comparten más que una parte del vocabulario y de la sintaxis. Desde luego, no pueden considerarse como tres instancias distintas de una misma *tarea*. Esto significa que la estrategia utilizada en DAF —entrenar los modelos sobre dos de los bloques y comprobar su rendimiento sobre el otro bloque— no tiene sentido en este caso. No obstante, se han puesto a prueba tres aproximaciones: la primera es precisamente la utilizada en DAF; la segunda consiste en entrenar un único modelo de lenguaje con los tres bloques, con lo cual el corpus de test forma parte del corpus de entrenamiento; y la tercera, la más extrema, consiste en estimar el modelo de lenguaje a partir del propio corpus de test. Como veremos más adelante, la primera de las aproximaciones, en la que los corpus de entrenamiento y test son absolutamente independientes y en gran medida *disjuntos*, resulta en una perplejidad exageradamente alta. Esto significa que los dos bloques a

Tabla 4.18: Características de los corpus de texto utilizados para estimar y evaluar los modelos de lenguaje en BDGEO-16, BDGEO-8, INFOTREN y CORLEC-EHU-1.

	BDGEO-16	BDGEO-8	INFOTREN	CORLEC-EHU-1
Diálogos entrenamiento	–	–	191	42
Frases/Turnos entrenamiento	8262	8262	1349	2123
Palabras entrenamiento	81620	81620	15367	72647
Diálogos test	–	–	36	14, 14, 14
Frases/Turnos test	600	522	308	700, 690, 690
Palabras test	5655	4812	3311	20951, 21068, 22718
Palabras vocabulario	1212	1212	1909	8242

partir de los cuales se estima el modelo de lenguaje apenas aportan información sobre lo que sucede en el tercer bloque. Por otra parte, la tercera aproximación está demasiado *inclinada* hacia el corpus de test. En el caso ideal, se tendría un corpus muy grande de transcripciones de la lengua hablada, que permitiría modelar ésta de forma robusta. Dicha base de datos *contendría* también al corpus de test, en el sentido de que una gran parte del léxico y la sintaxis —es decir, las combinaciones de palabras— del corpus de test aparecerían literalmente en el corpus de entrenamiento. La mejor aproximación al corpus ideal que tenemos a nuestro alcance es precisamente el corpus CORLEC-EHU-1 completo, y por ello será la opción elegida. Dicho corpus consta de 2123 turnos, con 72647 palabras y 4963 silencios. Se han considerado tres corpus de test —los mismos C1, C2 y C3 utilizados en DAF—, que contienen 20951, 21068 y 22718 palabras, y 629, 536 y 424 silencios, respectivamente. El vocabulario del corpus completo consta de 8242 palabras. En la Tabla 4.18 se resumen las características de los corpus de texto utilizados para estimar y evaluar los modelos de lenguaje.

Las transcripciones ortográficas de los corpus de entrenamiento de INFOTREN y CORLEC-EHU-1 —con las que se han estimado los modelos de lenguaje— se han generado a partir de las anotaciones en formato EHU simplificado, aplicando el siguiente conjunto de reglas:

1. Los segmentos solapados o ruidosos se mantienen, es decir, el texto afectado por dichas marcas se trata como texto *limpio*.
2. Una marca de continuación implica que el texto del turno en cuestión se une con el del siguiente turno del mismo hablante. De esta forma se mantiene la continuidad de la intervención, interrumpida generalmente por un turno de asentimiento del interlocutor.
3. Los ruidos aislados de todo tipo, las pausas habladas, las afirmaciones/negaciones guturales y las palabras cortadas se eliminan, es decir, no producen ningún símbolo ortográfico. Esto hace que algunos turnos desaparezcan.
4. Las pausas de silencio se transcriben con el símbolo ortográfico *P*.
5. Las marcas de alargamientos y de palabras mal pronunciadas no se tienen en cuenta, es decir, producen la misma secuencia de símbolos ortográficos que obtendríamos en su ausencia.
6. Las palabras extranjeras y las siglas se transcriben ortográficamente con la secuencia *AAAA=BBBB*, donde *AAAA* es la versión ortográfica usual y *BBBB* son los grafemas

correspondientes a su pronunciación en castellano. El resultado, por ejemplo *light=lait*, se trata como una palabra más del vocabulario. El modelo acústico de estas palabras se construye sólo con la parte que sigue al símbolo =.

7. Las marcas de fenómenos sintácticos y pragmáticos se eliminan.

Las transcripciones ortográficas de los corpus de test de INFOTREN y CORLEC-EHU-1 se obtienen aplicando el mismo conjunto de reglas, salvo las dos primeras. Los segmentos solapados o ruidosos se eliminan por completo, ya que sólo se dispone de las señales correspondientes a los segmentos *limpios*. Si el solapamiento o el ruido de fondo afectan al turno completo, entonces no se genera transcripción ortográfica y el turno como tal desaparece. Por otra parte, no se mantiene la continuidad de las intervenciones, ya que las señales están ubicadas en ficheros distintos y deben ser reconocidas por separado.

En todos los casos se han estimado modelos de lenguaje k -explorables en sentido estricto (k -ESS), con $k = 2$, $k = 3$ y $k = 4$. En las bases de datos de habla espontánea el símbolo P (correspondiente al modelo acústico de silencio) puede ser tratado de dos formas: (a) integrarlo en el modelo de lenguaje como una palabra más; o (b) eliminarlo del corpus de entrenamiento antes de estimar el modelo de lenguaje. En cualquier caso, el símbolo P formará parte del vocabulario, y podrá aparecer en la secuencia reconocida de palabras, sólo que en el caso (b) se le asignará la misma probabilidad que al resto de palabras no vistas en el corpus de entrenamiento. El símbolo P no aparece en el corpus de entrenamiento de BDGEO, por lo que en este caso forzosamente habrá de integrarse en el modelo de lenguaje como palabra no vista.

En la Tabla 4.19 se muestran las perplejidades para $k = 2$, $k = 3$ y $k = 4$, sobre los corpus de test de BDGEO, INFOTREN y CORLEC-EHU-1. Para estas dos últimas bases de datos se consideran dos modelos de lenguaje, según como sea tratada la unidad P , bien como palabra no vista, bien integrada en el modelo como una palabra más. En el caso de CORLEC-EHU-1, además del modelo de lenguaje estimado sobre el corpus completo (que, resumidamente, denominaremos ML123), se han considerado las otras dos aproximaciones, aunque sólo para el caso en que la unidad P está integrada en el modelo: el modelo de lenguaje estimado a partir de los bloques *complementarios* (ML23, ML13 y ML12), y el modelo de lenguaje estimado a partir del propio bloque (ML1, ML2 y ML3).

Atendiendo únicamente a los valores de la perplejidad mostrados en la Tabla 4.19, parece conveniente incluir la unidad P como una palabra más en el modelo de lenguaje, ya que la perplejidad de éste se reduce de manera notable, entre un 25 % y un 40 %. Esto significa que las pausas de silencio se comportan como elementos lingüísticamente significativos. Por otra parte, tanto en BDGEO como en INFOTREN —que poseen conjuntos de entrenamiento y de test independientes—, el modelo de lenguaje con $k = 3$ es el que proporciona una perplejidad más baja sobre el conjunto de test, y, por tanto, la opción óptima *a priori* para el reconocedor. En el caso de CORLEC-EHU-1, hay que analizar las tres aproximaciones. En los casos en que el conjunto de test está incluido en el conjunto de entrenamiento, la perplejidad decrece monótonamente al aumentar el valor de k . Como es lógico, las perplejidades más bajas se obtienen siempre con el modelo de lenguaje obtenido a partir del propio bloque de test. Los modelos de lenguaje obtenidos a partir de los bloques complementarios producen perplejidades muy grandes —del orden de 300—, que aumentan monótonamente con el valor de k . Esto se explica, como se ha dicho, por la absoluta independencia entre los conjuntos de entrenamiento y test.

Tabla 4.19: Perplejidades de los modelos de lenguaje ($k = 2$, $k = 3$ y $k = 4$), calculadas sobre los corpus de test de BDGEO, INFOTREN y CORLEC-EHU-1.

		Perplejidad		
		$k = 2$	$k = 3$	$k = 4$
Unidad P palabra no vista	BDGEO	13.10	7.53	7.17
	INFOTREN	44.78	41.01	41.69
	CORLEC-EHU-1: C1 (ML123)	60.60	13.76	8.15
	CORLEC-EHU-1: C2 (ML123)	61.61	14.34	8.55
	CORLEC-EHU-1: C3 (ML123)	60.87	12.99	7.55
Unidad P integrada en el modelo	INFOTREN	31.74	29.09	29.82
	CORLEC-EHU-1: C1 (ML123)	45.37	7.76	3.68
	CORLEC-EHU-1: C2 (ML123)	45.35	7.96	3.81
	CORLEC-EHU-1: C3 (ML123)	46.71	7.79	3.73
	CORLEC-EHU-1: C1 (ML23)	304.54	328.52	340.90
	CORLEC-EHU-1: C2 (ML13)	292.90	318.15	330.14
	CORLEC-EHU-1: C3 (ML12)	290.89	312.48	325.25
	CORLEC-EHU-1: C1 (ML1)	31.35	5.67	3.26
	CORLEC-EHU-1: C2 (ML2)	31.79	5.86	3.39
CORLEC-EHU-1: C3 (ML3)	33.12	5.82	3.33	

Por otra parte, a medida que crece el valor de k , los modelos se hacen más complejos —es decir, más grandes— y también más precisos en la descripción de las combinaciones de palabras encontradas en el corpus de entrenamiento. Por muy grande que sea éste, hay un valor de k a partir del cual se produce un ajuste excesivo del modelo al corpus de entrenamiento. En el caso de BDGEO e INFOTREN, dicho valor parece ser $k = 3$. Con respecto a la complejidad de los modelos, en la Tabla 4.20 se muestra el número de transiciones definidas para cada modelo de lenguaje. Los modelos obtenidos con $k = 4$ son aproximadamente dos veces más grandes que los obtenidos con $k = 3$, y éstos, a su vez, alrededor de tres veces más grandes que los obtenidos con $k = 2$. Por otra parte, los modelos de lenguaje de INFOTREN son los más *pequeños*; los de BDGEO son entre 1.5 y 2 veces más grandes; y los de CORLEC-EHU-1 —los más grandes— tienen tamaños entre 8 y 10 veces los de INFOTREN.

Con objeto de validar las conclusiones que pueden extraerse de las perplejidades, se han llevado a cabo experimentos de reconocimiento sobre BDGEO-8, con MOM discretos y continuos de 32 gaussianas, utilizando modelos de lenguaje k -ESS con $k = 2$, $k = 3$ y $k = 4$. Para cada valor de k se ha realizado un barrido del factor α que determina el peso relativo de la probabilidad del modelo de lenguaje con respecto a las probabilidades acústicas en el proceso de optimización que dirige la búsqueda. En este caso, los resultados de reconocimiento confirman lo observado en las perplejidades, ya que, efectivamente, los errores más pequeños se obtienen para $k = 3$ (véase la Tabla 4.21). Así pues, el resto de experimentos se ha llevado a cabo con los modelos de lenguaje obtenidos para $k = 3$. En la Tabla 4.22 se muestra el error de reconocimiento (%WER) obtenido sobre las bases de datos de habla leída y habla espontánea. Para éstas últimas se muestra el error con los dos modelos de lenguaje considerados —la unidad P introducida por *backoff* o integrada en el modelo como una palabra más. En todos los casos se ha efectuado un barrido del factor α , cuyo valor óptimo se muestra entre paréntesis junto al error de reconocimiento.

Tabla 4.20: Número de transiciones de que constan los modelos de lenguaje definidos para BDGEO, INFOTREN y CORLEC-EHU-1 ($k = 2$, $k = 3$ y $k = 4$).

		Número de transiciones			
		$k = 2$	$k = 3$	$k = 4$	
		BDGEO	9303	31016	68942
Unidad P palabra no vista		INFOTREN	5540	17149	34049
		CORLEC-EHU-1 (ML123)	54672	151135	272163
Unidad P integrada en el modelo		INFOTREN	5874	18496	36891
		CORLEC-EHU-1 (ML123)	54898	154490	282651
		CORLEC-EHU-1 (ML23)	40688	111012	199086
		CORLEC-EHU-1 (ML13)	40079	109589	196828
		CORLEC-EHU-1 (ML12)	39516	107227	191982
		CORLEC-EHU-1 (ML1)	22416	58018	100866
		CORLEC-EHU-1 (ML2)	23331	59909	103604
		CORLEC-EHU-1 (ML3)	23809	62165	108373

Tabla 4.21: Error de reconocimiento (%WER) sobre BDGEO-8, para $k = 2$, $k = 3$ y $k = 4$, utilizando MOM discretos y MOM continuos de 32 gaussianas. Entre paréntesis se muestra el valor óptimo de α .

	%WER		
	$k = 2$	$k = 3$	$k = 4$
MOMd	16.81 ($\alpha = 6$)	12.12 ($\alpha = 5$)	12.20 ($\alpha = 5$)
MOMc32g	11.17 ($\alpha = 9$)	7.12 ($\alpha = 9$)	7.90 ($\alpha = 9$)

Tabla 4.22: Error de reconocimiento (%WER) sobre las bases de datos de habla leída y habla espontánea, utilizando MOM discretos y MOM continuos de 32 gaussianas, y aplicando modelos de lenguaje k-ESS con $k = 3$. Entre paréntesis se muestra el valor óptimo de α .

		%WER		
		MOM-d	MOM-c32g	
		BDGEO-16	10.66 ($\alpha = 5$)	6.41 ($\alpha = 9$)
		BDGEO-8	12.12 ($\alpha = 5$)	7.12 ($\alpha = 9$)
Unidad P palabra no vista		INFOTREN	43.46 ($\alpha = 5$)	37.05 ($\alpha = 7$)
		CORLEC-EHU-1: C1 (ML123, MOM: C2+C3)	66.13 ($\alpha = 8$)	–
		CORLEC-EHU-1: C2 (ML123, MOM: C1+C3)	61.13 ($\alpha = 8$)	–
		CORLEC-EHU-1: C3 (ML123, MOM: C1+C2)	61.40 ($\alpha = 8$)	–
Unidad P integrada en el modelo		INFOTREN	42.98 ($\alpha = 5$)	37.36 ($\alpha = 7$)
		CORLEC-EHU-1: C1 (ML123, MOM: C2+C3)	65.55 ($\alpha = 8$)	–
		CORLEC-EHU-1: C2 (ML13, MOM: C1+C3)	87.68 ($\alpha = 8$)	–
		CORLEC-EHU-1: C2 (ML123, MOM: C1+C3)	60.07 ($\alpha = 8$)	42.34 ($\alpha = 10$)
		CORLEC-EHU-1: C2 (ML2, MOM: C1+C3)	47.13 ($\alpha = 8$)	–
		CORLEC-EHU-1: C3 (ML123, MOM: C1+C2)	60.71 ($\alpha = 8$)	–

En primer lugar, obsérvese que el valor óptimo de α depende del corpus de test considerado, pero también de los modelos acústicos —discretos o continuos—, ya que, mientras que las probabilidades del modelo de lenguaje son las mismas en ambos casos, el rango de variación de las probabilidades acústicas cambia radicalmente, y, por tanto, también deberá hacerlo el factor que establece el balance óptimo entre ambas. De hecho, sólo puede hablarse propiamente de *probabilidades* acústicas cuando se utilizan MOM discretos. Con MOM continuos lo que se tiene son *funciones densidad de probabilidad*, que pueden alcanzar valores arbitrariamente grandes.

En segundo lugar, mientras que los resultados de reconocimiento obtenidos con MOM discretos sobre BDGEO-16 y BDGEO-8 son coherentes con los obtenidos en DAF (el error es 1.5 puntos menor sobre BDGEO-16 que sobre BDGEO-8), los resultados con MOM continuos no parecen encajar con los obtenidos en DAF, ya que el error sobre BDGEO-16 es sólo 0.7 puntos menor que sobre BDGEO-8 —en DAF las diferencias se sitúan en torno a los 4 puntos: véase la Tabla 4.17. Esto pone de manifiesto la importancia del modelo de lenguaje en el proceso de búsqueda, sobre todo en tareas como BDGEO, con vocabularios medios y perplejidades bajas, en las que la información acústica, aunque imprescindible, no es determinante —en realidad, es el factor α el que, al enfatizar las probabilidades del modelo de lenguaje, reserva un papel secundario a los modelos acústicos.

En tercer lugar, el error de reconocimiento es mucho mayor en habla espontánea que en habla leída. BDGEO-8 e INFOTREN están adquiridas ambas a 8 kHz sobre canal telefónico real y simulado. Sin embargo, con MOM discretos el error de reconocimiento para BDGEO-8 se sitúa en torno al 12 %, mientras que para INFOTREN está en el 43 %. Con MOM continuos el error es aproximadamente del 7 % para BDGEO-8, y del 37 % para INFOTREN. Esto pone de manifiesto por un lado una mayor dificultad de la tarea acometida en el caso de INFOTREN —véanse las perplejidades en la Tabla 4.19—, y por otro la dificultad intrínseca del habla espontánea a nivel acústico: presencia de fenómenos acústicos no modelados —pausas habladas, alargamientos, ruidos producidos por los hablantes, etc.—, mayor variabilidad en la velocidad y el volumen de elocución, efectos de coarticulación mucho más acusados, etc. En el caso de INFOTREN, no parece que tenga especial relevancia la degradación producida por el ruido ambiental, ya que fue adquirida en entorno de oficina.

En cuarto lugar, el tratamiento de la unidad de silencio P en la estimación del modelo de lenguaje no parece tener una repercusión significativa en el error de reconocimiento. Si se utilizan MOM discretos, la inclusión de los silencios como pseudo-palabras adicionales en el corpus de entrenamiento produce una pequeña reducción del error, tanto en INFOTREN como en las tres particiones de CORLEC-EHU-1 —con el modelo de lenguaje del corpus completo: ML123. Por el contrario, si se utilizan MOM continuos, al menos en el caso de INFOTREN, se observa un ligero incremento del error. Teniendo en cuenta estos resultados y la gran disminución de la perplejidad que supone, en adelante se optará por integrar los silencios en el modelo de lenguaje.

Por último, en el caso de CORLEC-EHU-1 —siempre considerando el modelo de lenguaje del corpus completo, ML123—, el error de reconocimiento con MOM discretos se sitúa en alrededor del 61 % para las particiones 2 y 3, y en alrededor del 66 % para la partición 1, que parece la más *difícil*. Recordemos que se trata de una base de datos de diálogos naturales entre personas —es decir, no ligada a una tarea—, adquirida en condiciones ambientales adversas, con equipo muy rudimentario, que posteriormente ha sido transferida a formato digital a 16 kHz. Sólo es posible compararla con INFOTREN en lo que se refiere al estudio de las características

del habla espontánea, tarea llevada a cabo en el capítulo anterior.

También con MOM discretos y únicamente sobre la partición 2, se han llevado a cabo sendos experimentos de reconocimiento utilizando los modelos de lenguaje alternativos, ML13 y ML2, obteniendo errores del 87.68 % y del 47.13 %, respectivamente. Estos datos confirman la conveniencia de adoptar un único modelo basado en el corpus completo, como punto intermedio y razonable entre ambos extremos.

La gran duración de muchas de las intervenciones de que consta CORLEC-EHU-1 hace que los experimentos de reconocimiento lleven mucho más tiempo que para el resto de bases de datos. Esto ha obligado a restringir la experimentación sobre CORLEC-EHU-1 con MOM continuos; de ahí que aparezcan huecos en la Tabla 4.22. También por esa razón, en los capítulos siguientes sólo se presentarán resultados para la partición 2. Precisamente, el error de reconocimiento con MOM continuos de 32 gaussianas sobre la partición 2 de CORLEC-EHU-1 es tan sólo del 42.34 %, casi 18 puntos inferior al que se obtiene con MOM discretos, y sólo 5 puntos superior al obtenido sobre INFOTREN —aunque, no lo olvidemos, el modelo de lenguaje utilizado en los experimentos se ha estimado a partir del corpus completo, incluyendo el corpus de test.

Resumen del capítulo.

En este capítulo se han descrito, en primer lugar, las características más importantes del sistema de reconocimiento del GRFyTH: los parámetros acústicos, la cuantificación vectorial, el conjunto básico de unidades subléxicas, el modelo de lenguaje y, con mayor extensión, los modelos acústicos. Puesto que una parte de la experimentación tiene como objetivo establecer la metodología de estimación más adecuada —más robusta y menos costosa— tanto sobre habla leída como sobre habla espontánea, se han repasado los algoritmos clásicos de estimación de modelos ocultos de Markov (Baum-Welch y Viterbi) y se ha propuesto, como alternativa, el así llamado *algoritmo híbrido*, que aplica iterativamente el algoritmo de Baum-Welch restringido sobre los segmentos obtenidos mediante alineamiento forzado. En el apéndice C se incluyen las fórmulas de reestimación de MOM discretos, semicontinuos y continuos con representaciones acústicas múltiples. Aunque el rendimiento de los MOM discretos es notablemente inferior al de los MOM continuos, en determinadas tareas no demasiado complejas, con vocabularios medios y perplejidades bajas, proporcionan la precisión acústica suficiente, y su coste computacional es mucho más bajo, lo cual puede ser determinante para el desarrollo de prototipos en tiempo real. Esta es la razón principal por la que se ha realizado experimentación con MOM discretos, pero es que, además, la inicialización de los MOM continuos y semicontinuos se efectúa a partir de MOM discretos, utilizando las probabilidades de emisión normalizadas como pesos de las componentes en las mezclas, y las medias y varianzas de los diccionarios CV como parámetros de las gaussianas.

En el caso de los MOM continuos, el rendimiento obtenido con los tres algoritmos de estimación es prácticamente idéntico. De ahí que se haya optado por el algoritmo de Viterbi, que es el menos costoso. En el caso de los MOM discretos, el algoritmo de Viterbi se muestra muy sensible a la calidad de las segmentaciones. Cuando éstas no alcanzan una calidad mínima, los modelos resultantes ofrecen un rendimiento inferior al que proporcionan los modelos entrenados mediante el algoritmo de Baum-Welch general. Por su parte, los modelos obtenidos mediante el algoritmo híbrido proporcionan un rendimiento similar al que se obtiene con los modelos entrenados por Viterbi, mientras que el coste de estimación es ligeramente superior.

Así pues, en el caso de los MOM discretos se ha optado por el algoritmo de Baum-Welch general.

Cuestiones metodológicas aparte, el objetivo principal de este capítulo es evaluar el rendimiento de los modelos acústicos sobre habla leída y sobre habla espontánea, utilizando la configuración básica del sistema para habla leída. Tanto con MOM discretos como con MOM continuos, las tasas de DAF sobre INFOTREN están 10 puntos por debajo de las obtenidas sobre SENGLAR08, y las de CORLEC-EHU-1 alrededor de 20 puntos por debajo de las obtenidas sobre SENGLAR16. Las causas hay que buscarlas en las condiciones ambientales y de canal, en la presencia de fenómenos acústicos y léxicos no modelados y en la mayor variabilidad acústica, de pronunciaciones y de hablantes. Las tasas de DAF con MOM continuos están entre 6 y 7 puntos por encima de las obtenidas con MOM discretos.

En los experimentos de reconocimiento se han utilizado modelos léxicos lineales y modelos de lenguaje k -explorables en sentido estricto con $k = 3$. Los modelos de lenguaje estimados para las bases de datos de habla espontánea incluyen una pseudo-palabra adicional que representa los fragmentos de silencio. El mantenimiento de estos silencios en el corpus de entrenamiento hace descender la perplejidad entre un 25 % y un 40 % con respecto a su filtrado, aunque no afecta significativamente al error de reconocimiento. En cualquier caso, el descenso en la perplejidad revela que en habla espontánea las pausas de silencio se comportan como elementos lingüísticamente significativos.

Los resultados de reconocimiento sobre habla espontánea son mucho peores que sobre habla leída. Aunque una parte de la dificultad viene dada por la propia tarea (perplejidades más altas), son sobre todo las condiciones ambientales y de canal adversas, y la mayor variabilidad acústica del habla espontánea, los factores que determinan dichas diferencias. Así, con MOM discretos, el error de reconocimiento es del 12 % para BDGEO8 y del 43 % para INFOTREN, del 10.66 % para BDGEO16 y superior al 60 % para CORLEC-EHU-1. Con MOM continuos de 32 gaussianas por mezcla, el error es del 7 % para BDGEO8 y del 37 % para INFOTREN, del 6.41 % para BDGEO16 y del 42.34 % para la partición 2 de CORLEC-EHU-1.

Capítulo 5

Modelización de los fenómenos de habla espontánea.

El habla espontánea presenta una gran cantidad de fenómenos acústicos y léxicos que incrementan la variabilidad en la duración de los sonidos, en la calidad y énfasis de las articulaciones, etc. (Shriberg, 1999). En la literatura se han propuesto distintas estrategias para modelar explícitamente este tipo de fenómenos (Rose and Riccardi, 1999; O’Shaughnessy, 1999; Liu et al., 1998). Además, otras líneas de investigación como los modelos de pronunciación, los modelos duracionales o la inferencia de topologías de MOM, están encaminadas también a resolver la variabilidad introducida por estos fenómenos.

En este capítulo se propone una ampliación del conjunto de unidades subléxicas que cubra todos los fenómenos acústicos que aparecen en habla espontánea. Este tipo de aproximación es muy habitual en otras lenguas (inglés, japonés, alemán), pero no se ha desarrollado aún en castellano. Se obtendrán resultados de DAF y de reconocimiento que revelarán la utilidad de esta aproximación con respecto a la aproximación básica presentada en el capítulo anterior. También se presentará una aproximación intermedia, consistente en definir una única *unidad basura* para modelar todos los fenómenos de habla espontánea, que, como se verá, muestra un rendimiento inferior a la aproximación propuesta. Finalmente, se definirá un conjunto de unidades contextuales sobre el conjunto ampliado de unidades subléxicas, para tener en cuenta las dependencias contextuales en los casos en que el número de muestras disponibles permita modelarlas.

5.1. Ampliación del conjunto de unidades subléxicas.

En la Tabla 5.1 se muestra el conjunto de unidades que conforman la ampliación propuesta. El conjunto está formado por los fenómenos acústicos —salvo los silencios, que ya estaban incluidos— y algunos de los fenómenos léxicos descritos en el capítulo 2, etiquetados con letras mayúsculas. Las pausas habladas y los alargamientos representan para nosotros un mismo fenómeno acústico, en el primer caso como elementos pseudo-léxicos independientes, y en el segundo caso como parte de la realización acústica de una palabra. Así, por ejemplo, las muestras correspondientes a los alargamientos del sonido /m/ pueden juntarse con las correspondientes a las pausas habladas etiquetadas con ese mismo sonido, y entrenar con ellas un único modelo de /m/ *alargada*, etiquetada *M*. Precisamente, los sonidos alargados se etiquetan internamente con la misma letra que identifica el sonido pero en mayúscula. La distribución

Tabla 5.1: Ampliación del conjunto de unidades subléxicas para las bases de datos de habla espontánea: codificación interna del GRFyTH y descripción del sonido.

Codificación GRFyTH	Descripción
W	Ruido externo o de canal aislado
G	Aspiración producida por el hablante
K	Chasquido de labios producido por el hablante
T	Tos producida por el hablante
A	Alargamiento del sonido /a/ o pausa hablada realizada como /a/
E	Alargamiento del sonido /e/ o pausa hablada realizada como /e/
I	Alargamiento del sonido /i/
O	Alargamiento del sonido /o/
U	Alargamiento del sonido /u/
L	Alargamiento del sonido /l/
M	Alargamiento del sonido /m/ o pausa hablada realizada como /m/
N	Alargamiento del sonido /n/
R	Alargamiento del sonido /r/
S	Alargamiento del sonido /s/
B	Pausa hablada de identidad acústica confusa o inclasificable
X	Sonido gutural de afirmación

duracional que se deriva de las probabilidades de transición de los MOM es de tipo exponencial, lo cual puede no ser adecuado. Por ello, en algunos trabajos se modela explícitamente la duración de las unidades, mediante parámetros adicionales o mediante distribuciones que se acoplan a los MOM (Burshtein, 1996; Bonafonte et al., 1996; Chung and Seneff, 1999). La solución propuesta en este trabajo trata de aprovechar la potencia de los MOM para representar la duración de los segmentos, distinguiendo segmentos normales de segmentos largos y entrenando modelos específicos para unos y otros. De esta forma no se incrementa la complejidad de los algoritmos y se mejora la solución obvia que consiste en definir un único modelo, que absorbería en sus parámetros la variabilidad duracional de las muestras de entrenamiento. Tal es la aproximación que implícitamente se maneja con el conjunto básico de unidades.

En cuanto a los fenómenos léxicos, inicialmente se habían previsto dos unidades, etiquetadas *X* y *Z*, correspondientes a los sonidos guturales de afirmación y negación, respectivamente. En la práctica, al no haberse encontrado sonidos guturales de negación, la unidad *Z* ha sido descartada.

No siempre se tendrán muestras suficientes para entrenar los MOM de estas unidades. En tal caso, si se trata de alargamientos, las pocas muestras existentes son asignadas al sonido normal, y si se trata de otro tipo de unidades, reciben un tratamiento específico. Así, por ejemplo, en el caso de la unidad *T*, si se tuvieran muy pocas muestras, éstas pasarían a engrosar el conjunto de muestras de la unidad *W*.

Las transcripciones fonéticas en términos del conjunto ampliado de unidades subléxicas se obtienen en dos pasos: (1) a partir de las anotaciones en formato EHU simplificado se extraen transcripciones ortográficas enriquecidas con silencios y con las unidades de la Tabla 5.1; y (2) a estas transcripciones ortográficas se les aplica el conversor grafema-fonema que genera transcripciones fonéticas *canónicas* en castellano, el cual ha sido modificado para mantener los

silencios y las unidades acústicas adicionales. Para llevar a cabo el paso (1) se aplica el siguiente conjunto de reglas:

1. Los segmentos solapados o ruidosos se eliminan por completo, ya que sólo se dispone de las señales correspondientes a los segmentos *limpios*. Si el solapamiento o el ruido de fondo afectan al turno completo, entonces no se genera transcripción ortográfica.
2. Los ruidos externos aislados se transcriben como *W*, las aspiraciones producidas por el hablante como *G*, los chasquidos de labios como *K* y las toses como *T*.
3. Las pausas de silencio se transcriben como *P*.
4. Las pausas habladas realizadas como los sonidos /a/, /e/ y /m/ se transcriben como *A*, *E* y *M*, respectivamente. Las pausas habladas no identificables se transcriben como *B*.
5. Los símbolos ortográficos afectados por la marca de alargamiento se transcriben en mayúsculas. Sólo se han podido identificar los 10 tipos de alargamientos enumerados en la Tabla 5.1.
6. Las palabras mal pronunciadas y cortadas se transcriben tal cual han sido pronunciadas. Es decir, si una palabra aparece cortada, se transcribe únicamente el fragmento pronunciado.
7. Las afirmaciones guturales de afirmación se transcriben como *X*.
8. Las palabras extranjeras y las siglas se transcriben ortográficamente tal como han sido pronunciadas, es decir, con los grafemas correspondientes a su pronunciación en castellano.
9. Las marcas de fenómenos sintácticos y pragmáticos se eliminan, ya que no influyen en la transcripción fonética.

Este procedimiento se aplica literalmente sobre CORLEC-EHU-1, pero no sobre INFOTREN. Esta última no contiene solapamientos ni afirmaciones guturales, ni siglas, ni palabras extranjeras, y muy pocos fragmentos ruidosos. Estos últimos consisten normalmente en ruidos de fondo puntuales o saturaciones del canal que afectan a una o dos palabras y no las distorsionan gravemente. Por ello, a costa de introducir un cierto grado de distorsión en los modelos, se ha optado por eliminar esas marcas, dejando las palabras afectadas como si estuvieran libres de ruido.

En la Tabla 5.2 se muestra el número de unidades que resultan en las transcripciones fonéticas de INFOTREN y CORLEC-EHU-1. Se ha exigido un mínimo de 30 muestras de entrenamiento para que una unidad pase a formar parte del inventario definitivo. Las unidades *T*, *U* y *R* no han alcanzado esa cifra ni en INFOTREN ni en CORLEC-EHU-1, por lo que han sido reducidas —*T* sustituida por *W*, *U* por *u* y *R* por @— y no aparecen en la tabla. La unidad *X* sólo aparece en CORLEC-EHU-1. Aunque en uno de los bloques sólo aparecen 29 muestras, lo que se tiene en cuenta para validar la unidad es la suma para dos bloques, que da 69, 74 y 63 muestras de entrenamiento en cada una de las particiones, respectivamente. En algunos casos el número de muestras de entrenamiento es quizá demasiado pequeño como para entrenar modelos robustos, pero a cambio esto permite disponer de modelos acústicos especializados.

Tabla 5.2: Número de muestras de las unidades acústicas adicionales en los corpus de entrenamiento y test de INFOTREN, y en los tres bloques de CORLEC-EHU-1. Entre paréntesis se indica el porcentaje sobre el total de unidades.

	INFOTREN		CORLEC-EHU-1		
	Entrenamiento	Test	C1	C2	C3
W	312 (0.48)	79 (0.56)	90 (0.10)	85 (0.09)	54 (0.06)
G	1093 (1.69)	311 (2.20)	826 (0.90)	1004 (1.07)	984 (1.01)
K	507 (0.79)	93 (0.66)	73 (0.08)	59 (0.06)	42 (0.04)
A	262 (0.41)	62 (0.44)	253 (0.28)	227 (0.24)	189 (0.19)
E	715 (1.11)	193 (1.37)	594 (0.65)	515 (0.55)	501 (0.51)
I	95 (0.15)	23 (0.16)	134 (0.15)	92 (0.10)	94 (0.10)
O	86 (0.13)	14 (0.10)	231 (0.25)	206 (0.22)	229 (0.23)
L	88 (0.14)	17 (0.12)	41 (0.04)	29 (0.03)	31 (0.03)
M	132 (0.20)	55 (0.39)	82 (0.09)	116 (0.12)	131 (0.13)
N	35 (0.05)	7 (0.05)	86 (0.09)	85 (0.09)	96 (0.10)
S	34 (0.05)	12 (0.08)	98 (0.11)	76 (0.08)	64 (0.07)
B	169 (0.26)	41 (0.29)	187 (0.20)	192 (0.21)	189 (0.19)
X	–	–	34 (0.04)	29 (0.03)	40 (0.04)

Los porcentajes de fenómenos con respecto al número total de unidades son muy similares en los corpus de entrenamiento y test de INFOTREN, y también en los tres bloques de CORLEC-EHU-1. Esto indica que los fenómenos tienen una distribución bastante homogénea dentro de cada base de datos. Sin embargo, si comparamos ambas bases de datos, se comprueba que la proporción de fenómenos acústicos es mucho mayor en INFOTREN que en CORLEC-EHU-1. Los fenómenos añadidos suponen un 5.46 % del corpus de entrenamiento y un 6.42 % del corpus de test en INFOTREN, mientras que representan tan sólo un 2.98 %, un 2.89 % y un 2.70 % de los tres bloques de CORLEC-EHU-1, respectivamente. En ambos casos, la cantidad de fenómenos es muy pequeña como para que su modelización suponga en sí misma un incremento importante de las tasas de DAF. Sin embargo, su inserción en el entrenamiento hará que mejoren los modelos de fonemas y de silencio —que no se verán *contaminados* con muestras extrañas—, y por tanto también los modelos de palabra que con ellos se construyan, con lo que el rendimiento global del sistema podría mejorar notablemente.

5.1.1. Un conjunto alternativo de unidades subléxicas.

Se plantea una aproximación muy simple: definir una única unidad subléxica adicional, denominada *unidad basura*, representada con el símbolo F , para modelar todos los fenómenos acústicos *extraños*, en concreto ruidos externos, pausas habladas, palabras cortadas y sonidos guturales de afirmación. Esta unidad no es homogénea internamente, ya que modela fenómenos acústicos muy distintos, pero tiene la ventaja de contar con muchas muestras, con lo que la estimación del modelo acústico correspondiente será más robusta. Durante el reconocimiento la unidad basura debería activarse con observaciones acústicas que no se acomoden a ninguna las unidades básicas.

Para estimar los modelos acústicos es necesario disponer de transcripciones acústicas en términos del conjunto formado por las 24 unidades básicas más la unidad basura. Para ello,

Tabla 5.3: Número de muestras de la *unidad basura* en los corpus de entrenamiento y test de INFOTREN, y en los tres bloques de CORLEC-EHU-1. Entre paréntesis se indica el porcentaje sobre el total de unidades.

	INFOTREN		CORLEC-EHU-1		
	Entrenamiento	Test	C1	C2	C3
F	1085 (1.73)	280 (2.05)	704 (0.78)	647 (0.70)	660 (0.68)

en primer lugar se generan las transcripciones ortográficas a partir de las anotaciones, según el siguiente conjunto de reglas:

1. Los segmentos solapados o ruidosos se eliminan por completo. Si el solapamiento o el ruido de fondo afectan al turno completo, entonces no se genera transcripción ortográfica.
2. Los ruidos aislados externos, las toses, las pausas habladas, las afirmaciones guturales y las palabras cortadas producen el símbolo ortográfico *F*.
3. Las aspiraciones y los chasquidos de labios producidos por el hablante se transcriben como *P*, ya que por lo general muestran un nivel de energía muy bajo.
4. Las pausas de silencio se transcriben con el símbolo ortográfico *P*.
5. Las marcas de alargamientos y de palabras mal pronunciadas no se tienen en cuenta, es decir, producen la misma secuencia de símbolos ortográficos que obtendríamos en su ausencia.
6. Las palabras extranjeras y las siglas se transcriben ortográficamente tal como han sido pronunciadas, es decir, con los grafemas correspondientes a su pronunciación en castellano.
7. Las marcas de fenómenos sintácticos y pragmáticos se eliminan, ya que no influyen en la transcripción fonética.

Finalmente, la transcripción fonética de cada turno se obtiene aplicando a la transcripción ortográfica el conversor grafema-fonema habitual, que mantiene los símbolos *F* y *P*. En la Tabla 5.3 se indica el número de muestras de la unidad *F* que resultan para las dos bases de datos de habla espontánea. De nuevo se comprueba que la distribución de fenómenos es bastante uniforme dentro de cada base de datos, y que la frecuencia relativa de los mismos en INFOTREN es más de dos veces la observada en CORLEC-EHU-1.

5.2. Resultados de DAF.

Para evaluar los modelos acústicos correspondientes al conjunto de unidades ampliado, la secuencia reconocida y la transcripción correcta se pueden alinear sin más, es decir, sin eliminar de ellas silencios, ruidos, alargamientos, etc. La tasa resultante no sería exactamente la de DAF sino la del reconocimiento conjunto de fonemas y fenómenos acústicos. Al haber incrementado el número de unidades subléxicas, la probabilidad de confundir unas con otras es mayor, y es posible, incluso, que se obtengan tasas peores que en la aproximación básica. Alternativamente,

Tabla 5.4: Tasas de DAF sobre INFOTREN, utilizando el conjunto de unidades subléxicas ampliado, el conjunto alternativo formado por los fonemas, el silencio y la unidad basura, y, como referencia, el conjunto básico de unidades subléxicas, con MOM discretos y MOM continuos de 8, 16, 32 y 64 gaussianas. Se indica también qué unidades intervienen en los alineamientos.

	%DAF						
	Conjunto ampliado			Conjunto alternativo		Conjunto básico	
	Todas	Fonemas+PH+A	Fonemas	Todas	Fonemas	Todas	Fonemas
MOMd	51.50	54.16	55.81	50.84	53.49	49.76	51.98
MOMc8	52.70	56.86	58.41	53.83	56.54	51.63	53.87
MOMc16	54.76	58.83	60.43	55.44	58.20	53.29	55.55
MOMc32	56.21	60.03	61.34	56.49	59.35	54.48	56.86
MOMc64	57.15	60.52	61.74	57.11	60.07	55.72	58.12

antes de alinear las secuencias correcta y reconocida, se pueden filtrar los fenómenos acústicos, eliminando ruidos, pausas habladas no identificadas, silencios y afirmaciones guturales, y sustituyendo las unidades alargadas por unidades normales. Esto produciría dos secuencias de fonemas, y de su alineamiento obtendríamos la tasa de DAF. En este trabajo se ha considerado también una tercera posibilidad, en la que se eliminan los silencios, los ruidos, las pausas habladas sin identificar y las afirmaciones guturales, pero se mantienen los alargamientos y el resto de pausas habladas. Con ello se pretende evaluar la capacidad de los MOM —sin ningún tipo de adaptación o mejora— para identificar pausas habladas y alargamientos. En la aproximación basada en una única unidad basura sólo se consideran dos alineamientos, uno sin filtrar nada y otro eliminando las unidades *P* y *F* de las secuencias reconocida y correcta.

Basándonos en las conclusiones del capítulo 4, se han llevado a cabo dos líneas de experimentación. La primera de ellas utiliza MOM discretos, inicializados a partir de modelos básicos obtenidos sobre habla leída y aplicando 20 iteraciones del algoritmo BWg. Los MOM se inicializan a partir de los modelos óptimos obtenidos en el capítulo 4. Los modelos de alargamientos se inicializan con los modelos correspondientes a los fonemas sin alargar, y el resto de fenómenos recibe una inicialización equiprobable. En el caso del conjunto de unidades alternativo, la unidad basura también se inicializa con distribuciones uniformes. La segunda línea de experimentación utiliza MOM continuos de 8, 16, 32 y 64 gaussianas, inicializados con los mejores modelos discretos obtenidos anteriormente y estimados mediante 20 iteraciones del algoritmo de entrenamiento por Viterbi. En ambos casos los modelos obtenidos en cada iteración son evaluados frente a un corpus de test y aquéllos que dan una tasa de DAF más alta son elegidos como óptimos.

5.2.1. Resultados de DAF sobre INFOTREN.

En la Tabla 5.4 se muestran los resultados de DAF sobre INFOTREN, utilizando el conjunto de unidades subléxicas ampliado y el conjunto alternativo con la unidad basura. Si atendemos únicamente a las columnas 3 y 5 —correspondientes a alineamientos entre secuencias de fonemas, es decir, habiendo filtrado el resto de unidades— vemos que las tasas que proporciona el conjunto ampliado están alrededor de 2 puntos absolutos por encima de las que proporciona el conjunto alternativo, que a su vez superan en alrededor de 2 puntos las del conjunto básico (fonemas+silencio) presentadas en el capítulo anterior (véanse las Tablas 4.12 y 4.15). El

coste temporal del reconocimiento en DAF se incrementa con respecto al conjunto básico de unidades, ya que aumenta el número de modelos, y por tanto el número de estados, y como se sabe, el algoritmo de Viterbi es cuadrático con éste último. No sucederá lo mismo en el sistema de reconocimiento integrado, ya que, como veremos, el tamaño del vocabulario aumentará en unas pocas entradas, y los modelos de palabras se mantendrán inalterados. En todo caso, si los modelos acústicos son más precisos, la búsqueda debería ser más eficiente y, por tanto, más rápida.

Para evaluar las tasas de las columnas 1 y 4, correspondientes a los alineamientos sin filtrados previos, es necesario compararlas con las que se obtienen con el conjunto básico sin filtrar los silencios, que aparecen en la columna 6. Como puede verse, tanto el conjunto ampliado como el conjunto alternativo mejoran el rendimiento del conjunto básico. Por otra parte, en el caso del conjunto ampliado, el tener más unidades —posiblemente peor entrenadas que las del conjunto alternativo— hace que se confundan más fácilmente unas con otras. De ahí que la columna 4, correspondiente al conjunto alternativo, muestre en algunos casos una tasa más alta que la columna 1. Por último, la columna 2, que presenta la tasa obtenida con los alineamientos entre fonemas, pausas habladas y alargamientos, demuestra que en esas confusiones intervienen sobre todo los ruidos, pausas habladas sin identificar y los silencios, ya que la mejora que se observa entre las columnas 1 y 2 es notablemente superior a la que se produce entre las columnas 2 y 3.

5.2.1.1. Análisis de la matriz de confusión.

Para analizar con más detalle los errores de DAF, se ha procesado la matriz de confusión obtenida en el alineamiento de la secuencias correcta y la secuencia reconocida con MOM continuos de 32 gaussianas, utilizando el conjunto ampliado de unidades, sin filtrar nada. En la Tabla 5.5 se muestran los resultados de DAF desglosados por unidades, así como las sustituciones más frecuentes para cada unidad.

En el caso de los silencios, ruidos y pausas habladas sin identificar se observa un gran número de inserciones (63.17 %) y un escasísimo número de borrados (0.68 %). Ello quiere decir que se reconocen muchos segmentos pequeños de este tipo de unidades que no están presentes en las anotaciones del corpus de test, y que dichos segmentos se identifican con facilidad —no siempre correctamente. En cuanto a las sustituciones, se producen sobre todo entre unidades de esta misma categoría: unos ruidos se confunden con otros o con silencios.

Con respecto a los alargamientos (o pausas habladas, en el caso de *A*, *E* y *M*), se siguen observando muy pocos borrados (1.08 %) y menos inserciones que en el caso anterior, aunque su número es todavía relativamente elevado (30.99 %). De hecho, es muy habitual encontrar en las secuencias reconocidas una unidad normal seguida de la misma unidad alargada (o viceversa), casos éstos que quizá no deberían contarse como errores. Por eso, aunque el número de sustituciones es aceptable (17.12 %), la tasa de aciertos es de sólo el 50.81 %. La mayor parte de las sustituciones que tienen lugar en esta categoría se producen con otros alargamientos, con el mismo sonido sin alargar, con ruidos o con silencios.

En lo que respecta al conjunto de fonemas, una parte importante de las sustituciones se produce entre sonidos del mismo tipo (entre oclusivas, entre vocales, etc.). Es destacable el buen reconocimiento de las vocales, en particular la /i/ (75.37 %) y la /a/ (74.68 %). La tasa de aciertos entre los fonemas es del 60.10 %.

Tabla 5.5: Resultados de DAF sobre INFOTREN, utilizando el conjunto ampliado de unidades subléxicas y MOM continuos de 32 gaussianas, desglosados por unidades: número de aciertos (A), sustituciones (S), borrados (B) e inserciones (I), tasa de DAF y las 3 sustituciones más frecuentes —entre paréntesis se indica el porcentaje sobre el total de sustituciones.

Unidad	A	S	B	I	%DAF	Sustituciones más frecuentes		
P	180	2	0	628	22.22	b(50.00)	e(50.00)	
W	44	31	4	144	19.73	G(16.13)	r(16.13)	K(12.90)
G	231	76	4	216	43.83	P(34.21)	K(14.47)	p(7.89)
K	72	21	0	178	26.57	P(28.57)	s(19.05)	G(9.52)
B	6	30	5	45	6.98	P(16.67)	n(16.67)	e(10.00)
A	55	7	0	50	49.11	E(28.57)	P(14.29)	g(14.29)
E	155	35	3	59	61.51	e(25.71)	n(11.43)	a(11.43)
I	16	7	0	5	57.14	e(28.57)	G(14.29)	P(14.29)
O	9	5	0	5	47.37	o(80.00)	W(20.00)	
L	6	10	1	3	30.00	l(40.00)	d(20.00)	W(10.00)
M	39	15	1	48	37.86	n(26.67)	p(13.33)	G(6.67)
N	0	7	0	1	0.00	m(28.57)	K(14.29)	M(14.29)
S	2	9	1	1	15.38	P(22.22)	s(22.22)	p(11.11)
p	245	41	12	106	60.64	t(43.90)	k(14.63)	W(7.32)
b	250	175	63	53	46.21	d(30.86)	g(12.57)	p(11.43)
t	436	82	16	77	71.36	k(37.80)	p(26.83)	z(13.41)
d	343	221	69	118	45.67	t(22.62)	r(22.62)	b(14.03)
k	276	55	16	87	63.59	t(23.64)	P(14.55)	z(14.55)
g	58	100	18	73	23.29	b(18.00)	r(15.00)	e(10.00)
c	40	9	4	7	66.67	t(22.22)	z(22.22)	r(22.22)
f	25	27	6	0	43.10	z(55.56)	x(11.11)	p(7.41)
z	248	35	13	145	56.24	t(31.43)	x(11.43)	r(11.43)
s	638	219	55	123	61.64	z(57.53)	r(5.94)	e(4.11)
x	41	9	0	11	67.21	k(22.22)	z(22.22)	i(22.22)
m	244	74	21	33	65.59	n(37.84)	b(21.62)	d(12.16)
n	504	168	67	169	55.51	m(35.12)	r(10.71)	l(7.74)
h	34	6	0	1	82.93	n(33.33)	i(33.33)	t(16.67)
l	409	164	85	65	56.57	r(26.22)	b(17.68)	o(9.15)
y	51	38	9	9	47.66	d(23.68)	i(15.79)	l(13.16)
r	593	318	226	110	47.55	d(21.38)	l(12.26)	s(9.43)
@	11	18	5	6	27.50	r(50.00)	d(33.33)	b(11.11)
i	866	111	87	85	75.37	e(18.92)	y(15.32)	z(9.91)
e	1314	440	267	145	60.66	i(34.77)	E(12.05)	o(7.50)
a	1392	236	137	99	74.68	e(33.47)	o(18.22)	A(10.59)
o	611	297	95	145	53.22	a(34.34)	u(27.27)	e(15.49)
u	242	42	17	49	69.14	o(21.43)	g(11.90)	l(11.90)
Total	9686	3140	1307	3099	56.21			

5.2.1.2. Un ejemplo.

La utilización de unidades adicionales para *rellenar* los espacios que quedan entre los fonemas tiene una gran importancia tanto en entrenamiento como en test. En entrenamiento permite estimar las unidades de fonemas con las muestras adecuadas. En test permite ajustar mejor los segmentos, evitando inserciones indeseadas, y tras el filtrado de las unidades adicionales se obtienen secuencias fonéticas más cercanas a las reales. Es difícil medir estos efectos de forma cuantitativa. Para esta base de datos no se dispone de segmentaciones manuales, por lo que no se puede definir nada parecido a la calidad de las segmentaciones del apartado 4.3.2.1. Sí es posible, en cambio, ilustrar lo que sucede mediante un ejemplo.

En la Figura 5.1 se muestra un fragmento de señal con su espectrograma, correspondiente a la frase "*quisiera sabEr WA M*". Nótese el alargamiento interno de la /e/ en "*saber*", el ruido externo y las dos pausas habladas consecutivas. En la figura aparecen 6 segmentaciones. La primera de ellas (empezando por arriba) corresponde a la que se obtiene mediante alineamiento forzado utilizando el conjunto de unidades ampliado y MOM continuos de 32 gaussianas. La segunda es la segmentación reconocida con esos mismos modelos. La tercera y la cuarta son las segmentaciones forzada y reconocida utilizando el conjunto de unidades alternativo (fonemas + silencio + unidad basura) y MOM continuos de 32 gaussianas. La quinta y la sexta son las segmentaciones forzada y reconocida utilizando el conjunto básico de unidades (fonemas + silencio) y MOM continuos de 32 gaussianas.

Nótese cómo se alinean las muestras en las segmentaciones forzadas: con los conjuntos ampliado y alternativo los alineamientos son muy buenos, especialmente en el primer caso, que aparece impreso sobre la señal y el espectrograma; sin embargo, el alineamiento óptimo obtenido con el conjunto básico de unidades, al no disponer de la transcripción completa, asigna tramos erróneos a muchos de los fonemas, sobre todo alrededor de los alargamientos y los ruidos.

Por otra parte, la secuencia de fonemas reconocida con el conjunto de unidades ampliado es muy próxima a la correcta, como también lo es la segmentación. Tan sólo se insertan dos breves segmentos tras la unidad *W* (*t* indica la presencia de un brevísimo silencio seguido de una explosión, y *e* indica la indefinición inicial de la pausa hablada) y un pequeño silencio tras la unidad *M*, que *realmente está presente* en la señal, aunque no se incluye en las anotaciones por ser demasiado breve. Con el conjunto alternativo el segmento conflictivo *WAM* se transcribe como *F*, pero en la secuencia reconocida se identifica como *FamnF*, que reproduce más o menos lo que acústicamente sucede dentro del mismo. Esto supone 4 inserciones con respecto a la transcripción correcta. Algo similar sucede cuando se utiliza el conjunto básico de unidades, ya que el segmento *WAM* —que en este caso se transcribe como la cadena vacía— se interpreta como *etanp*, que supone 5 inserciones.

5.2.1.3. Correlación entre la tasa de DAF y el número de fenómenos.

Los resultados anteriores parecen indicar que una cantidad importante de errores se concentra alrededor de los ruidos, alargamientos y pausas habladas, particularmente cuando se utiliza el conjunto básico de unidades o cuando éste se amplía únicamente con una unidad *basura*. Para comprobarlo se ha considerado por un lado la *densidad de fenómenos* de cada muestra, calculada como el número de fenómenos dividido por la longitud de la transcripción en términos del conjunto ampliado de unidades, y por otro la tasa de DAF obtenida después de filtrar todas las unidades salvo los fonemas. En la Tabla 5.6 se presentan los coeficientes de correlación

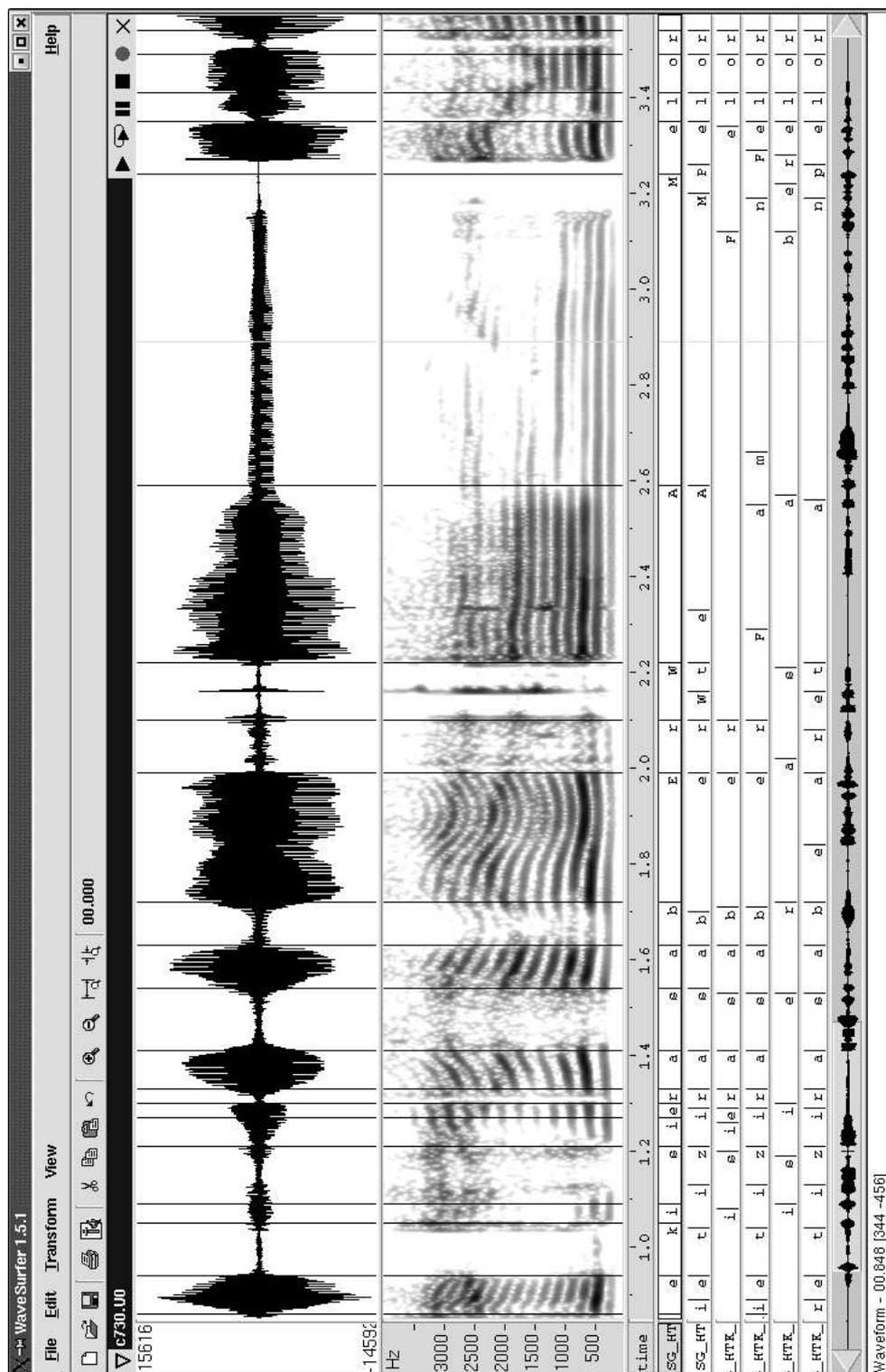


Figura 5.1: Segmentaciones forzada y reconocida obtenidas con el conjunto ampliado (las dos primeras, empezando por arriba), el conjunto alternativo (la tercera y cuarta) y el conjunto básico de unidades (las dos últimas), para la frase "quisiera saber W A M".

Tabla 5.6: Correlación entre la tasa de DAF y la densidad de fenómenos en cada muestra del corpus de test de INFOTREN, utilizando los conjuntos ampliado, alternativo y básico de unidades con MOM continuos de 32 gaussianas. Se indica también la significancia estadística del resultado, con respecto a la hipótesis nula.

	ρ	Significancia
Conjunto ampliado	-0.3829	$t(306) = -7,25, p \leq 3,622 \cdot 10^{-12}$
Conjunto alternativo	-0.3527	$t(306) = -6,594, p \leq 1,949 \cdot 10^{-10}$
Conjunto básico	-0.4152	$t(306) = -7,983, p \leq 3,194 \cdot 10^{-14}$

obtenidos para el corpus de test de INFOTREN utilizando el conjunto ampliado, el conjunto alternativo y el conjunto básico de unidades con MOM continuos de 32 gaussianas. Como era de esperar, se obtienen correlaciones negativas, aunque bastante bajas. Además, no se observan diferencias significativas entre los tres conjuntos de unidades.

5.2.2. Resultados de DAF sobre CORLEC-EHU-1.

En la Tabla 5.7 se presentan los resultados de DAF sobre CORLEC-EHU-1 utilizando los conjuntos de unidades ampliado y alternativo. Las tasas entre fonemas (columnas 3 y 5) son muy inferiores a las obtenidas sobre INFOTREN, lo cual pone de manifiesto una vez más las difíciles condiciones de esta base de datos. También son más pequeñas que en INFOTREN las mejoras que introducen las unidades del conjunto ampliado. En general, las tasas que se obtienen con el conjunto ampliado apenas son un punto más altas que las del conjunto alternativo, y éstas, a su vez, sólo mejoran unas décimas las del conjunto básico de unidades (véase la Tabla 5.8). Esto puede ser debido, por una parte, a que los fenómenos del conjunto ampliado se producen en CORLEC-EHU-1 en una proporción mucho menor que en INFOTREN, lo cual limita la mejora que potencialmente puede suponer la introducción de modelos explícitos para este tipo de fenómenos. Por otra parte, la gran variabilidad en las condiciones ambientales y de canal hace que los fenómenos de habla espontánea, al igual que los propios fonemas, no se modelen adecuadamente, limitando aún más su potencial aportación. Por último, una diferencia importante con respecto INFOTREN es que parece haber un mayor número de errores relacionados con los alargamientos y pausas habladas que con el resto de unidades adicionales, ya que el incremento que se produce entre las columnas 1 y 2 en la Tabla 5.7 es más pequeño que el que se produce entre las columnas 2 y 3.

5.2.2.1. Análisis de la matriz de confusión.

En la Tabla 5.9 se muestran los resultados de DAF con MOM continuos de 32 gaussianas, desglosados por unidades, así como las sustituciones más frecuentes para cada unidad —obtenidas después de procesar la matriz de confusión. En los alineamientos se ha utilizado el conjunto ampliado de unidades.

Es destacable el gran número de borrados y sustituciones que se da en el conjunto de fonemas (16.33 % y 26.44 %, respectivamente), mientras que el número de inserciones es relativamente bajo (6.25 %). Esto puede estar poniendo de manifiesto una articulación deficiente de los sonidos —mayor velocidad de elocución, pronunciaciones más relajadas que en habla leída—, que en ocasiones no llegan a identificarse o se confunden con otros similares —de hecho, las sustituciones más frecuentes son las que se producen entre sonidos del mismo tipo. La tasa de

Tabla 5.7: Tasas de DAF sobre las tres particiones de CORLEC-EHU-1, utilizando el conjunto de unidades subléxicas ampliado y el conjunto alternativo formado por los fonemas, el silencio y la unidad basura, con MOM discretos y MOM continuos de 8, 16, 32 y 64 gaussianas. Se indica también qué unidades intervienen en los alineamientos.

	%DAF - C2+C3/C1				
	Conjunto ampliado			Conjunto alternativo	
	Todas	Fonemas+PH+A	Fonemas	Todas	Fonemas
MOMd	42.04	42.43	45.67	44.34	44.80
MOMc8	46.08	46.86	49.23	47.87	48.48
MOMc16	47.35	47.93	50.11	48.66	49.21
MOMc32	48.19	48.71	50.63	49.45	50.01
MOMc64	49.13	49.59	51.14	50.24	50.81

	%DAF - C1+C3/C2				
	Conjunto ampliado			Conjunto alternativo	
	Todas	Fonemas+PH+A	Fonemas	Todas	Fonemas
MOMd	44.06	44.36	47.04	45.55	46.02
MOMc8	48.35	48.93	50.96	49.00	49.63
MOMc16	49.42	50.00	51.76	49.74	50.43
MOMc32	50.44	50.98	52.42	50.68	51.38
MOMc64	51.32	51.79	53.05	51.30	51.99

	%DAF - C1+C2/C3				
	Conjunto ampliado			Conjunto alternativo	
	Todas	Fonemas+PH+A	Fonemas	Todas	Fonemas
MOMd	45.01	45.14	47.64	46.16	46.60
MOMc8	49.47	50.01	51.52	49.98	50.50
MOMc16	50.72	51.14	52.47	50.88	51.39
MOMc32	51.64	52.01	53.13	51.62	52.13
MOMc64	52.45	52.73	53.66	52.40	52.91

Tabla 5.8: Tasas de DAF promedio para el conjunto ampliado, el conjunto alternativo y el conjunto básico de unidades subléxicas, filtrando todo salvo los fonemas, con MOM discretos y MOM continuos de 8, 16, 32 y 64 gaussianas.

	%DAF - Promedio		
	Conjunto ampliado	Conjunto alternativo	Conjunto básico
MOMd	46.78	45.81	45.60
MOMc8	50.57	49.54	49.01
MOMc16	51.45	50.34	50.02
MOMc32	52.06	51.17	50.83
MOMc64	52.62	51.90	51.52

Tabla 5.9: Resultados de DAF sobre el bloque C2 de CORLEC-EHU-1, utilizando el conjunto ampliado de unidades subléxicas y MOM continuos de 32 gaussianas, desglosados por unidades: número de aciertos (A), sustituciones (S), borrados (B) e inserciones (I), tasa da DAF y las 3 sustituciones más frecuentes —entre paréntesis se indica el porcentaje sobre el total de sustituciones.

Unidad	A	S	B	I	%DAF	Sustituciones más frecuentes		
P	384	126	26	736	30.19	G(26.19)	p(7.14)	t(6.35)
W	11	70	4	96	6.08	G(18.57)	P(14.29)	f(8.57)
G	793	185	26	233	64.11	P(49.73)	p(4.86)	k(3.78)
K	15	42	2	93	9.87	P(28.57)	G(14.29)	s(9.52)
X	11	18	0	7	30.56	B(16.67)	K(11.11)	t(11.11)
B	33	132	27	107	11.04	n(12.12)	P(9.85)	p(5.30)
A	161	59	7	97	49.69	a(44.07)	E(5.08)	O(5.08)
E	371	126	18	124	58.06	e(31.75)	i(11.11)	n(7.94)
I	56	30	6	69	34.78	i(20.00)	P(10.00)	E(10.00)
O	138	57	11	107	44.09	o(36.84)	m(7.02)	a(7.02)
L	5	22	2	18	10.64	l(27.27)	t(18.18)	n(9.09)
M	47	58	11	81	23.86	N(20.69)	n(18.97)	P(6.90)
N	27	47	11	70	17.42	n(34.04)	M(19.15)	E(8.51)
S	23	49	4	36	20.54	s(36.73)	P(10.20)	E(6.12)
p	1403	1018	355	364	44.68	t(37.62)	f(13.85)	k(10.81)
b	407	1035	657	167	17.96	n(10.43)	m(9.57)	d(8.89)
t	2870	1051	500	355	60.09	p(21.31)	z(17.98)	k(17.13)
d	935	1823	1493	188	21.06	l(12.07)	t(11.63)	r(8.94)
k	2005	1406	568	320	46.64	t(23.47)	p(16.36)	f(6.83)
g	128	374	299	127	13.79	r(12.30)	b(8.56)	m(8.29)
c	163	69	13	59	53.62	s(33.33)	t(21.74)	z(15.94)
f	254	195	66	196	35.72	z(37.44)	t(14.87)	s(10.26)
z	923	485	141	285	50.33	s(20.62)	f(19.59)	x(17.11)
s	5677	1011	617	258	75.06	z(14.84)	f(7.02)	r(6.82)
x	323	144	44	117	51.43	z(30.56)	k(11.81)	s(11.11)
m	1369	1167	533	248	41.27	n(33.93)	b(11.31)	l(9.43)
n	3282	2056	1236	524	46.24	l(15.03)	m(14.79)	r(11.09)
h	123	59	24	64	45.56	n(18.64)	m(10.17)	l(10.17)
l	1515	1568	966	339	34.53	r(17.60)	n(12.82)	m(7.72)
y	205	204	91	118	33.17	i(27.45)	d(11.27)	l(6.37)
r	1715	1747	1533	371	31.96	l(11.79)	d(7.27)	n(7.10)
@	144	214	67	58	29.81	r(26.64)	b(7.48)	d(7.01)
i	3961	1471	989	373	58.30	e(24.07)	y(9.04)	u(8.43)
e	7175	3350	2462	480	53.28	o(21.88)	i(15.97)	a(11.64)
a	7478	1642	1127	318	70.78	o(29.66)	e(20.16)	A(7.61)
o	5077	2669	1320	352	53.91	a(27.24)	u(18.85)	e(13.45)
u	1894	670	609	335	53.99	o(26.12)	l(6.72)	m(6.57)
Total	51101	26449	15865	7890	50.44			

DAF para los fonemas es tan sólo del 50.97 %. Sin duda, también están pesando las condiciones ambientales y de canal, que parecen influir más sobre las bajas frecuencias. Así, es significativo que sea la /s/ el único fonema con una tasa de DAF mayor en CORLEC-EHU-1 que en INFOTREN (75.06 % frente a 61.64 %). Como se sabe, la /s/ se caracteriza por una fuerte componente espectral a altas frecuencias, que se modela mejor a 16 kHz que a 8 kHz, y que no parece estar contaminada por ruido en el caso de CORLEC-EHU-1. También es significativo, con respecto a INFOTREN, el descenso en las tasas de las vocales. En particular, la tasa de DAF de la /i/ pasa del 75.37 % al 58.30 %, y la de la /u/, del 69.14 % al 53.99 %.

La tasa de DAF para los alargamientos es del 42.51 %, 8 puntos inferior a la obtenida para INFOTREN. Esta diferencia proviene del mayor número de borrados (3.59 % frente al 1.08 % de INFOTREN) y especialmente de sustituciones (23.00 % frente al 17.12 % de INFOTREN), ya que la cantidad de inserciones, aún siendo importante, es proporcionalmente idéntica a la obtenida para INFOTREN (30.90 % frente al 30.99 %). Entre las sustituciones más frecuentes se cuentan los propios sonidos sin alargar y otros alargamientos. Este incremento en la proporción de sustituciones es atribuible nuevamente a la mayor probabilidad de confundir unos sonidos con otros cuando las condiciones ambientales o de canal son adversas.

En cuanto al resto de fenómenos, es destacable el gran número de inserciones, sobre todo de silencios, que, sin embargo, es proporcionalmente inferior al obtenido sobre INFOTREN (40.04 % frente a 63.17 %). Por otra parte, crece el porcentaje de borrados, que sigue siendo bajo (2.68 % frente a 0.68 %) y crece también el porcentaje de sustituciones (18.04 % frente a 8.35 %). Como resultado neto de todo esto, la tasa de DAF de este tipo de fenómenos es más alta que la observada en INFOTREN (39.25 % frente a 27.8 %). A este crecimiento de la tasa de DAF contribuye especialmente el hecho de que las aspiraciones producidas por el usuario (unidad *G*) se reconocen mejor en CORLEC-EHU-1 que en INFOTREN (64.11 % frente a 43.83 %). Esto se explica con los mismos argumentos aportados en el caso del fonema /s/.

5.2.2.2. Un ejemplo.

Para completar el análisis de los resultados de DAF sobre CORLEC-EHU-1 —al menos de forma cualitativa—, en la Figura 5.2 se muestra un fragmento de señal con su espectrograma, correspondiente a la frase "*G y vAn WM observando*". La figura muestra asimismo tres pares de segmentaciones, que corresponden al alineamiento forzado y la secuencia reconocida con MOM continuos de 32 gaussianas, utilizando los conjuntos de unidades ampliado (segmentaciones primera y segunda, empezando por arriba), alternativo (tercera y cuarta) y básico (quinta y sexta), respectivamente.

Los alineamientos forzados que se obtienen con los conjuntos ampliado y alternativo son aceptables. Con el conjunto básico el alargamiento interno del fonema /a/ en la palabra *van* se alinea perfectamente con el modelo de la /a/, pero el segmento *WM* no aparece anotado, con lo cual sus muestras son erróneamente asignadas a las unidades que le rodean, en concreto a la /n/ anterior y a los fonemas /o/ y /b/ que le siguen. Curiosamente, el alineamiento que proporciona el conjunto alternativo alrededor del ruido y la pausa hablada es más ajustado que el que proporciona el conjunto ampliado. La unidad *F* engloba justamente los segmentos correspondientes al ruido externo, la pausa hablada nasal y a la pequeña distorsión que sigue a ésta —que debería haberse anotado como *B*. La unidad *F* ha sido entrenada con un conjunto grande y heterogéneo de ruidos y pausas habladas, de ahí que se adecúe perfectamente a esta combinación de sonidos. Sin embargo, con el conjunto ampliado la unidad *W* recoge sólo la parte final del ruido y la unidad *M* recoge estrictamente la pausa hablada nasal, lo cual hace que

Tabla 5.10: Correlación entre la tasa de DAF y la densidad de fenómenos en cada muestra de test de CORLEC-EHU-1, utilizando los conjuntos ampliado, alternativo y básico de unidades con MOM continuos de 32 gaussianas. Se indica también la significancia estadística del resultado, con respecto a la hipótesis nula.

	ρ	Significancia
Conjunto ampliado	-0.4114	$t(2135) = -20,86, p \leq 8,401 \cdot 10^{-88}$
Conjunto alternativo	-0.5064	$t(2135) = -27,14, p \leq 1,502 \cdot 10^{-138}$
Conjunto básico	-0.1065	$t(2078) = -4,88, p \leq 1,123 \cdot 10^{-6}$

la pequeña distorsión que le sigue sea erróneamente asignada en parte al fonema /o/ (que no recibe una sola de sus muestras) y en parte al fonema /b/, que también recoge una parte de las muestras del fonema /o/, pero ninguna de las suyas, que son asignadas al fonema /s/.

Las secuencias reconocidas presentan todas ellas numerosos errores, especialmente la que se obtiene con el conjunto básico de unidades. La aspiración inicial es reconocida como *Gt* con el conjunto ampliado y como *Ft* con el conjunto alternativo, es decir, en ambos casos se reconoce correctamente el fenómeno pero después, en el tramo final de la aspiración, que contiene un pequeño silencio, se inserta el fonema /t/. Con el conjunto básico de unidades el fenómeno se identifica como *dt*. Por otra parte, el alargamiento interno del fonema /a/ en la palabra *van* es reconocido correctamente en todos los casos, mientras que el ruido externo provoca numerosas inserciones. La pausa hablada que sigue al ruido se reconoce como *N* con el conjunto ampliado, como *n* con el conjunto alternativo y como *l* con el conjunto básico. Los tres casos se cuentan como errores, aunque sólo el último debería considerarse como tal, ya que confunde un sonido nasal con un sonido líquido.

Con este ejemplo podemos comprender por qué el conjunto ampliado apenas mejora el rendimiento en DAF de los otros dos conjuntos de unidades sobre CORLEC-EHU-1: debido a las condiciones adversas y variables de ruido, no es posible modelar adecuadamente ruidos, alargamientos, o incluso los propios fonemas, y el conjunto ampliado de unidades no puede evitar los errores que sistemáticamente aparecen alrededor de los fenómenos de habla espontánea.

5.2.2.3. Correlación entre la tasa de DAF y la densidad de fenómenos.

Con objeto de comprobar cuantitativamente hasta qué punto la tasa de DAF se ve afectada por la presencia de fenómenos de habla espontánea, se ha calculado la correlación entre la tasa de DAF y la densidad de fenómenos en las muestras de test de CORLEC-EHU-1, utilizando MOM continuos de 32 gaussianas y los tres conjuntos de unidades (véase la Tabla 5.10). Los coeficientes de correlación se han calculado juntando los datos correspondientes a las tres particiones. Para los tres conjuntos de unidades se obtienen correlaciones negativas, estadísticamente significativas. Sin embargo, para el conjunto básico la correlación es muy pequeña, lo que parece indicar que la tasa de DAF es independiente de la densidad de fenómenos. Para los conjuntos ampliado y alternativo la correlación es más elevada, pero no lo bastante como para establecer una dependencia clara entre la tasa de DAF y la densidad de fenómenos.

5.2.3. Resultados de DAF con unidades contextuales.

En los apartados 5.2.1 y 5.2.2 se ha comprobado que el conjunto de unidades subléxicas ampliado —que incluye unidades específicas para los fenómenos de habla espontánea de tipo

acústico— proporciona un mejor rendimiento que el conjunto básico (fonemas + silencio) y que el conjunto alternativo (fonemas + silencio + unidad *basura*). Las diferencias son claras en el caso de INFOTREN y más pequeñas en el caso de CORLEC-EHU-1, debido en parte a la menor proporción de fenómenos y en parte a las condiciones adversas de ruido de esta base de datos. En cualquier caso, en adelante se utilizará el conjunto ampliado como conjunto de unidades incontextuales para habla espontánea.

Sin embargo, no se ha modelado la variabilidad acústica interna de las unidades subléxicas, en particular la que proviene de la coarticulación de los sonidos. Es de esperar que esta variabilidad sea mucho mayor en habla espontánea que en habla leída. Por tanto, la mejora que potencialmente puede introducir el uso de unidades subléxicas contextuales debería ser mayor en habla espontánea que en habla leída. Con objeto de comparar el grado en que la variabilidad debida a la coarticulación afecta al reconocimiento en uno y otro caso, se han definido conjuntos de unidades contextuales mediante una técnica muy sencilla: se examinan las unidades a izquierda y derecha —una de ellas o ambas— para clasificar las muestras de cada unidad en el corpus de entrenamiento, y se utilizan sólo aquellos subconjuntos de muestras suficientemente grandes como para estimar los modelos correspondientes de manera robusta. Esta metodología se ha aplicado sobre el conjunto básico de unidades en habla leída y sobre el conjunto ampliado en habla espontánea. A las unidades que resultan las denominaremos *unidades contextuales por conteo* (Bonafonte et al., 1995), y aunque no todas son asimilables a fonemas, abusando de la terminología llamaremos *bifonemas a izquierda* a las unidades definidas teniendo en cuenta la unidad precedente, *bifonemas a derecha* a las unidades definidas teniendo en cuenta la unidad siguiente y *trifonemas* a las unidades definidas teniendo en cuenta ambas.

Se ha elegido esta metodología debido a su simplicidad. Evidentemente, en la literatura pueden encontrarse aproximaciones de todo tipo para modelar la variabilidad contextual, como unidades específicas para las transiciones (Varona et al., 1995; Nedel et al., 2000), unidades alternativas al fonema (Mariño et al., 2000) y, sobre todo, lo que se conoce como *trifonemas generalizados* (Lee et al., 1990), obtenidos normalmente mediante árboles de decisión que combinan la información sobre el contexto con medidas de discriminación acústica (Bahl et al., 1991; Young et al., 1994; Kuhn et al., 1995). Probablemente una metodología basada en esta última aproximación deberá ser aplicada en el futuro para optimizar el rendimiento de los modelos acústicos en habla espontánea.

Tanto en habla leída como en habla espontánea se ha exigido un mínimo de 300 muestras para poder definir y estimar un MOM contextual. Este umbral se ha establecido heurísticamente en base a experiencias previas (López de Ipiña et al., 1999, 2000). En otros trabajos que aplican el mismo criterio de *entrenabilidad* se han manejado umbrales más bajos, entre 100 y 200 muestras (Bonafonte et al., 1995; Mariño et al., 2000). Utilizar un umbral demasiado bajo podría desembocar en modelos poco robustos. Exigir un mayor número de muestras proporciona estimaciones más robustas, pero posiblemente no el rendimiento óptimo del sistema de reconocimiento. Para alcanzarlo habría que aplicar distintos umbrales y elegir aquél que proporcionase las mejores tasas. En este caso no se busca optimizar el sistema de reconocimiento sino tan sólo caracterizar el habla espontánea, de ahí que se haya optado por un umbral *robusto*. De hecho, hay unidades incontextuales que no alcanzan el umbral de 300 muestras (véanse las tablas 4.5 y 5.2), por lo que no se podrán generar para ellas variedades contextuales.

Los tres tipos de unidades se definen, entrenan y evalúan separadamente, lo cual quiere decir que comparten muestras de entrenamiento. Así, por ejemplo, si el trifonema *e-s+t* —el fonema /s/ con el fonema /e/ a la izquierda y el fonema /t/ a la derecha— cuenta con 1224

Tabla 5.11: Número de bifonemas a izquierda, bifonemas a derecha y trifonemas definidos en las bases de datos de habla leída y habla espontánea, aplicando un umbral de 300 muestras. Entre paréntesis se indica el porcentaje de unidades del corpus de entrenamiento y del corpus de test cubiertas por cada conjunto. En la última columna se indica el número total de unidades contextuales por conteo, así como la cobertura que proporcionan.

	Bifonemas-I	Bifonemas-D	Trifonemas	Total
SEGLAR16	66 (61.36 %, 60.98 %)	66 (61.36 %, 62.22 %)	6 (3.62 %, 1.54 %)	138 (82.92 %, 83.07 %)
SEGLAR08	66 (61.36 %, 60.60 %)	66 (61.36 %, 61.84 %)	6 (3.62 %, 1.49 %)	138 (82.92 %, 82.79 %)
INFOTREN	64 (62.10 %, 60.32 %)	63 (61.41 %, 59.71 %)	16 (9.31 %, 7.83 %)	143 (82.48 %, 80.20 %)
CORLEC-EHU-1	143 (84.18 %, 83.80 %)	146 (84.84 %, 84.52 %)	98 (25.45 %, 25.22 %)	387 (94.14 %, 93.99 %)

muestras en el corpus de entrenamiento, obviamente el bifonema a izquierda $e-s$ y el bifonema a derecha $s+t$ tendrán al menos 1224 muestras, y las tres unidades se entrenarán por separado compartiéndolas. En la Tabla 5.11 se presenta el número de bifonemas a izquierda, bifonemas a derecha y trifonemas definidos para SENGLAR16, SENGLAR08, INFOTREN y la partición 2 de CORLEC-EHU-1, así como la cobertura del corpus de entrenamiento y del corpus de test que proporciona cada conjunto de unidades por separado, y la cobertura global de los tres conjuntos. Las unidades incontextuales proporcionan cobertura para el resto de casos.

La mayor parte de las unidades contextuales definidas son variedades de los fonemas. En INFOTREN se han generado tan sólo dos bifonemas a izquierda de las unidades adicionales ($o-G$ y $s-G$), y otro tanto sucede con CORLEC-EHU-1 ($P-K$ y $P-G$). En SENGLAR16 y SENGLAR08 el hecho de haber anotado todas las frases del corpus de entrenamiento con silencios inicial y final, hace que aparezcan variedades de esta unidad, en concreto 4 bifonemas a izquierda, 2 bifonemas a derecha y 3 trifonemas.

De los datos de cobertura se puede deducir que los trifonemas, por sí mismos, no pueden mejorar significativamente las tasas de DAF. No obstante, en todos los casos la cobertura global de las unidades contextuales está por encima del 80 %. Así pues, se están modelando dependencias contextuales para una gran parte de las muestras de entrenamiento, y el impacto de los modelos resultantes debería ser notable, muy próximo al máximo alcanzable mediante esta metodología.

Los MOM discretos de unidades contextuales se inicializan con los modelos correspondientes a las unidades incontextuales, y se entrenan mediante el algoritmo BWg (12 iteraciones en habla leída, 20 iteraciones en habla espontánea). Para ello se generan transcripciones del corpus de entrenamiento en términos del conjunto de unidades contextuales, completándolas —cuando no hay cobertura— con las unidades incontextuales. Estas últimas participan en el entrenamiento pero no se reestiman. El mismo procedimiento se aplica para entrenar MOM continuos de 32 gaussianas, sólo que inicializándolos a partir de los mejores MOM discretos contextuales e incontextuales, y aplicando el algoritmo de entrenamiento por Viterbi. El modelo de reconocimiento se monta con el conjunto de unidades contextuales y las unidades incontextuales, que prestan la cobertura necesaria. En la construcción de las hipótesis se aplican restricciones fonológicas que tienen que ver con la definición de las unidades. Así, por ejemplo, al bifonema $s+t$ sólo puede seguirle una variedad del fonema $/t/$ que sea compatible a la izquierda con el fonema $/s/$. En la cadena reconocida sólo aparecen las unidades centrales, es decir, no se hace explícita la variedad contextual que ha sido reconocida. Los alineamientos se realizan después de filtrar todas las unidades salvo los fonemas, como se ha explicado al comienzo del

Tabla 5.12: Tasas de DAF para SENGLAR16, SENGLAR08, INFOTREN y la partición 2 de CORLEC-EHU-1 utilizando unidades contextuales, por separado y conjuntamente, con MOM discretos. En todos los casos el modelo de reconocimiento incluye las unidades incontextuales, para asegurar la cobertura.

	Bifonemas-I	Bifonemas-D	Trifonemas	Todos
SENGLAR16	69.54	69.46	65.55	69.93
SENGLAR08	67.09	67.14	62.92	67.36
INFOTREN	61.25	61.71	57.57	61.97
CORLEC-EHU-1	51.79	51.94	51.32	52.92

Tabla 5.13: Tasas de DAF para SENGLAR16, SENGLAR08, INFOTREN y la partición 2 de CORLEC-EHU-1 utilizando unidades contextuales, por separado y conjuntamente, con MOM continuos de 32 gaussianas. En todos los casos el modelo de reconocimiento incluye las unidades incontextuales, para asegurar la cobertura.

	Bifonemas-I	Bifonemas-D	Trifonemas	Todos
SENGLAR16	75.38	75.55	72.50	75.75
SENGLAR08	70.49	72.57	68.51	72.35
INFOTREN	64.76	66.98	62.77	66.71
CORLEC-EHU-1	57.09	56.98	55.95	57.64

apartado 5.2.

En las Tablas 5.12 y 5.13 se muestran los resultados de DAF para habla leída y habla espontánea utilizando unidades contextuales, por separado y conjuntamente, con MOM discretos y continuos, respectivamente. En el caso de CORLEC-EHU-1 se han llevado a cabo experimentos de DAF sólo sobre la partición 2. Los experimentos para las particiones 1 y 3 no se han realizado debido al elevado coste computacional que supone. Se ha elegido precisamente la partición 2 (C1+C3 en entrenamiento, C2 en test) porque, atendiendo a los resultados con unidades incontextuales, las tasas de DAF que se obtienen sobre ella son próximas a la media que se obtiene para las tres particiones. Se está hablando de costes realmente muy elevados. Por ejemplo, cada iteración de entrenamiento de los MOM continuos de bifonemas a izquierda lleva 3.5 veces más tiempo que el entrenamiento de los MOM continuos de unidades incontextuales. Esto se debe al incremento en el número de gaussianas, que corresponde al del incremento en el número de modelos, que pasa de 37 a 180: cada iteración de entrenamiento sobre el corpus C1+C3 de CORLEC-EHU-1 lleva casi 11 horas en una CPU Intel Xeon a 2.4 GHz. El coste computacional crece aún más cuando se trata de reconocer, ya que el número de estados del modelo de reconocimiento se multiplica casi por 5 en el caso de los bifonemas a izquierda. En teoría, el coste podría incrementarse en un factor de 23. En la práctica, el coste computacional del reconocimiento crece tan sólo en un factor de 7, debido a las restricciones fonológicas: cada experimento de reconocimiento de los bifonemas a izquierda sobre el corpus de test C2 de CORLEC-EHU-1 lleva algo más de 15 horas en una CPU Intel Xeon a 2.4 GHz.

En todos los casos, tanto en habla leída como en habla espontánea, se utilicen MOM discretos o MOM continuos de 32 gaussianas, la reducción del error en DAF, con respecto a los resultados que se obtienen con unidades incontextuales, es superior al 10%. No se observan grandes diferencias entre unas bases de datos y otras: INFOTREN es la que presenta mejoras más acusadas, de cerca del 14%, y CORLEC-EHU-1 la que presenta mejoras más modestas,

de alrededor del 11 %. Se puede concluir que la modelización de la variabilidad contextual en habla espontánea tiene un efecto relativo en las tasas de DAF similar al observado en habla leída.

Sin embargo, en términos absolutos, la mejora que introducen las unidades contextuales con respecto a las incontextuales es mayor para habla espontánea que para habla leída. Así, por ejemplo, con MOM discretos la mejora es de unos 4.65 puntos para habla leída y de 6 puntos para habla espontánea. Con MOM continuos se obtienen mejoras de entre 3 y 4 puntos para habla leída y de algo más de 5 puntos para habla espontánea.

Por último, la tasa de DAF obtenida con MOM discretos de unidades contextuales es, en todos los casos, 5 puntos inferior a la obtenida con MOM continuos, si bien el coste computacional de los primeros es muchísimo menor. Así, por ejemplo, reconocer el corpus de test C2 de CORLEC-EHU-1 con MOM continuos de 32 gaussianas de bifonemas a izquierda lleva en la práctica 18 veces más tiempo que hacerlo con MOM discretos.

5.3. Resultados de reconocimiento.

En el modelo de lenguaje del sistema de referencia se añadía al vocabulario una *pseudo-palabra* —el símbolo *P*— que representaba el silencio. En este caso se añadirán 13 *pseudo-palabras* en el caso de INFOTREN, y 14 *pseudo-palabras* en el caso de CORLEC-EHU-1, que representan los distintos fenómenos acústicos que pueden encontrarse en habla espontánea —incluido el silencio—, identificados mediante los mismos símbolos utilizados en el conjunto ampliado de unidades subléxicas. De hecho, los modelos acústicos de estas pseudo-palabras están formados simplemente por los MOM de las unidades subléxicas correspondientes.

Las transcripciones ortográficas de los turnos que componen las bases de datos de habla espontánea consistirán, por tanto, en secuencias de palabras y *pseudo-palabras*, y ambas podrán ser tenidas en cuenta por el modelo de lenguaje. En este trabajo se consideran dos aproximaciones: en la primera, las pseudo-palabras son eliminadas del corpus de entrenamiento y el modelo de lenguaje las trata como palabras no vistas, por medio del suavizado (*backoff*); en la segunda aproximación, las transcripciones ortográficas conservan las pseudo-palabras, y el modelo de lenguaje las trata como al resto de palabras del corpus de entrenamiento. En el cálculo del error de reconocimiento, si no se especifica lo contrario, las pseudo-palabras son filtradas antes de efectuar los alineamientos entre las secuencias correcta y reconocida.

Para estimar el modelo de lenguaje, en primer lugar es necesario generar las transcripciones ortográficas del corpus de entrenamiento. Partiendo de las anotaciones en formato EHU simplificado, se ha aplicado el siguiente conjunto de reglas:

1. Los segmentos solapados o ruidosos se mantienen, es decir, el texto afectado por dichas marcas se trata como texto *limpio*.
2. Una marca de continuación implica que el texto del turno en cuestión se une con el del siguiente turno del mismo hablante. De esta forma se mantiene la continuidad de la intervención, interrumpida generalmente por un turno de asentimiento del interlocutor.
3. Los ruidos externos aislados se transcriben como *W*, las aspiraciones producidas por el hablante como *G*, los chasquidos de labios como *K* y las toses como *T*.
4. Las pausas de silencio se transcriben como *P*.

5. Las pausas habladas realizadas como los sonidos /a/, /e/ y /m/ se transcriben como *A*, *E* y *M*, respectivamente. Las pausas habladas no identificables se transcriben como *B*.
6. Los símbolos ortográficos afectados por la marca de alargamiento que aparecen en el interior de una palabra se procesan como no alargados. Los que aparecen al principio o al final de una palabra generan dos símbolos: el grafema normal en el borde de la palabra y, separado por un espacio blanco, el mismo grafema en mayúscula, como si se tratara de un fenómeno aislado. Esta decisión tiene que ver con la necesidad de definir un único modelo por palabra. Dado el escaso número de alargamientos internos —tal como se ha visto en el capítulo 3, un 12 % en INFOTREN y un 6 % en CORLEC-EHU-1— su eliminación no introduce una gran distorsión en el modelo. En cuanto a los alargamientos de sonidos en los bordes, la adición artificial de una pseudo-palabra no hace sino reproducir una realidad acústica. Así, la palabra *desdE* —nótese el alargamiento final del fonema /e/— generaría la secuencia *desde E*. Por último, las palabras monofonémicas alargadas se transcriben con el grafema correspondiente en mayúscula.
7. Las palabras mal pronunciadas y cortadas se transcriben como si se hubieran pronunciado correctamente.
8. Las afirmaciones guturales de afirmación se transcriben como *X*.
9. Las palabras extranjeras y las siglas se transcriben ortográficamente con la secuencia *AAAA=BBBB*, donde *AAAA* es la versión ortográfica usual y *BBBB* son los grafemas correspondientes a su pronunciación en castellano. El resultado, por ejemplo *light=lait*, se trata como una palabra más del vocabulario. El modelo acústico de estas palabras se construye sólo con la parte que sigue al símbolo =.
10. Las marcas de fenómenos sintácticos y pragmáticos se eliminan.

Las transcripciones ortográficas de los corpus de test de INFOTREN y CORLEC-EHU-1 se obtienen aplicando el mismo conjunto de reglas, salvo las dos primeras: los segmentos solapados o ruidosos se eliminan por completo y no se mantiene la continuidad de las intervenciones. Esto es necesario porque las señales correspondientes a cada turno se ubican en ficheros distintos, y se han eliminado de ellas los fragmentos ruidosos o solapados.

El corpus de entrenamiento del modelo de lenguaje de INFOTREN contiene 15341 palabras y 3997 fenómenos acústicos, y el corpus de test 3309 palabras y 1063 pseudo-palabras. El vocabulario consta de 1909 palabras y 13 pseudo-palabras. En el caso de CORLEC-EHU-1, se ha estimado un único modelo de lenguaje (ML123) a partir del corpus completo, que consta de 2385 turnos —en los que se incluyen los segmentos ruidosos o solapados—, 72462 palabras y 10768 pseudo-palabras. Se han considerado tres corpus de test, correspondientes a los bloques C1, C2 y C3, esta vez sólo con los segmentos considerados *limpios*, que contienen 20899, 21011 y 22671 palabras, y 3295, 3189 y 3010 pseudo-palabras, respectivamente. El vocabulario consta de 8287 palabras y 14 pseudo-palabras (véase la Tabla 5.14).

Se han entrenado modelos de lenguaje *k*-explorables en sentido estricto, para $k = 2$, $k = 3$ y $k = 4$, con las pseudo-palabras integradas en el modelo o introducidas por *backoff* como palabras no vistas. Los modelos resultantes son sólo ligeramente más grandes que los generados en el capítulo 4. En la Tabla 5.15 se muestra el número de transiciones de que consta cada uno de ellos (compárense con los datos de la Tabla 4.20). Las perplejidades de estos modelos —calculadas sobre los correspondientes corpus de test— se muestran en la Tabla 5.16.

Tabla 5.14: Características de los corpus de texto —fenómenos de habla espontánea incluidos— utilizados para estimar y evaluar los modelos de lenguaje de INFOTREN y CORLEC-EHU-1.

nº palabras / nº FHE	INFOTREN	CORLEC-EHU-1
Corpus de entrenamiento	15341 / 3997	72462 / 10768
Corpus de test	3309 / 1063	20899 / 3295 21011 / 3189 22671 / 3010
Vocabulario	1909 / 13	8247 / 14

Tabla 5.15: Número de transiciones de los modelos de lenguaje de INFOTREN y CORLEC-EHU-1 ($k = 2$, $k = 3$ y $k = 4$), estimados a partir de transcripciones enriquecidas con pseudo-palabras que representan fenómenos de habla espontánea.

	Número de transiciones					
	FHE por <i>backoff</i>			FHE integrados		
	$k = 2$	$k = 3$	$k = 4$	$k = 2$	$k = 3$	$k = 4$
INFOTREN	5537	16920	33367	6570	21832	44660
CORLEC-EHU-1 (ML123)	55176	153059	276188	57548	165835	306829

Tabla 5.16: Perplejidades de los modelos de lenguaje de INFOTREN y CORLEC-EHU-1 ($k = 2$, $k = 3$ y $k = 4$), calculadas sobre los correspondientes corpus de test. Los modelos han sido estimados a partir de transcripciones enriquecidas con pseudo-palabras que representan fenómenos de habla espontánea.

	Perplejidad					
	FHE por <i>backoff</i>			FHE integrados		
	$k = 2$	$k = 3$	$k = 4$	$k = 2$	$k = 3$	$k = 4$
INFOTREN	196.93	186.89	190.21	36.84	34.88	36.37
CORLEC-EHU-1: C1 (ML123)	89.64	26.78	18.13	48.93	8.29	3.84
CORLEC-EHU-1: C2 (ML123)	94.23	27.50	18.27	51.15	8.09	3.66
CORLEC-EHU-1: C3 (ML123)	89.11	23.60	15.41	52.23	7.82	3.58

Tabla 5.17: Error de reconocimiento (%WER) sobre INFOTREN, utilizando MOM discretos y MOM continuos de 32 gaussianas, y modelos de lenguaje k-ESS con $k = 2$, $k = 3$ y $k = 4$ y los fenómenos de habla espontánea integrados. Entre paréntesis se muestran los valores óptimos de α .

	%WER		
	$k = 2$	$k = 3$	$k = 4$
MOM discretos	36.28 ($\alpha = 5$)	35.05 ($\alpha = 5$)	34.81 ($\alpha = 5$)
MOM continuos 32g	30.01 ($\alpha = 8$)	29.32 ($\alpha = 8$)	29.10 ($\alpha = 8$)

La presencia de fenómenos de habla espontánea en los corpus de entrenamiento y test incrementa las perplejidades con respecto a las obtenidas en el capítulo 4, especialmente cuando los fenómenos se introducen en el modelo de lenguaje por *backoff*.

En el caso de INFOTREN, si los fenómenos se integran en el modelo de lenguaje como palabras normales, se obtienen perplejidades en torno a 36. Sin embargo, si se introducen por *backoff*, las perplejidades suben hasta 190. Este incremento es debido a la alta frecuencia de los fenómenos, y pone de manifiesto que dichos fenómenos *realmente* se producen con regularidad y operan como elementos lingüísticamente significativos. En el caso de CORLEC-EHU-1 se observa un comportamiento similar, aunque las diferencias no son tan grandes.

En el caso de CORLEC-EHU-1, las perplejidades descienden monótonamente al aumentar el valor de k , debido, como ya sabemos, a que el corpus de test forma parte del corpus de entrenamiento, decisión ésta que ya ha sido discutida en el capítulo 4. Con $k = 3$ y los fenómenos integrados en el modelo, la perplejidad se sitúa en alrededor de 8 —sólo ligeramente superior a la obtenida para BDGEO. Siendo ésta una perplejidad relativamente baja, las adversas condiciones de grabación y la elevada variabilidad acústica de esta base de datos harán que el error de reconocimiento se sitúe por encima de lo que dicha perplejidad parece indicar —y, en particular, muy por encima del error obtenido sobre BDGEO-16. En general, los valores de perplejidad por sí mismos *sólo* pueden ser indicadores del error de reconocimiento si el resto de condiciones se mantienen inalteradas. En la práctica, esto significa que las perplejidades sirven como indicadores del error sólo si se considera una misma base de datos. Por ejemplo, en (Chelba, 2000), utilizando trigramas, se reportan perplejidades de 68.56 para *Switchboard* y de 217.8 para una base de datos de noticias de radio y televisión —ambas bases de datos contienen habla espontánea genérica en inglés—, mientras que los errores de reconocimiento son del 40.5 % y el 33 %, respectivamente.

Se han ejecutado experimentos de reconocimiento sobre INFOTREN utilizando los modelos acústicos óptimos obtenidos en DAF y los modelos de lenguaje obtenidos para $k = 2$, $k = 3$ y $k = 4$, con los fenómenos de habla espontánea integrados. Tal como se muestra en la Tabla 5.17, el error de reconocimiento disminuye a medida que crece el valor de k , situándose en alrededor del 35 % con MOM discretos y ligeramente por encima del 29 % con MOM continuos de 32 gaussianas. Estos datos representan una disminución del error de alrededor del 20 % con respecto al error de reconocimiento obtenido con el conjunto básico de unidades subléxicas. Por otra parte, el modelo de lenguaje obtenido con $k = 3$ ofrece un balance adecuado entre rendimiento y coste computacional.

Fijado $k = 3$, se ha ejecutado otra serie de experimentos sobre INFOTREN para comprobar cuál de las dos formas de incorporar los fenómenos de habla espontánea al modelo de lenguaje proporciona un mejor rendimiento. Los resultados de reconocimiento están en consonancia con las perplejidades: es preferible mantener los fenómenos en el corpus de entrenamiento que

Tabla 5.18: Error de reconocimiento (%WER) sobre INFOTREN, utilizando MOM discretos y MOM continuos de 32 gaussianas, y modelos de lenguaje k-ESS con $k = 3$, con los fenómenos integrados e insertados por *backoff*. Se incluye también, como referencia, el error obtenido sin modelar los FHE. Entre paréntesis se muestran los valores óptimos de α .

	%WER		
	FHE integrados	FHE por <i>backoff</i>	Sin FHE
MOM discretos	35.05 ($\alpha = 5$)	41.29 ($\alpha = 5$)	42.98 ($\alpha = 5$)
MOM continuos 32g	29.32 ($\alpha = 8$)	34.35 ($\alpha = 9$)	37.36 ($\alpha = 7$)

filtrarlos e insertarlos por *backoff* (véase la Tabla 5.18).

En el caso de CORLEC-EHU-1, se ha ejecutado un único experimento de reconocimiento, sobre la partición 2, utilizando MOM continuos de 32 gaussianas —los óptimos de DAF— y el modelo de lenguaje k-ESS obtenido para $k = 3$, con los fenómenos de habla espontánea integrados. Si bien el modelo de lenguaje se ha estimado a partir del corpus completo, los modelos acústicos se han estimado únicamente a partir de los bloques C1 y C3, y son, por tanto, independientes del corpus de test C2. En estas condiciones, el error de reconocimiento es del 36.36 % (con $\alpha = 10$), 6 puntos por debajo del obtenido sin modelar los fenómenos de habla espontánea (42.34 %, también con $\alpha = 10$), lo cual representa un 14 % de disminución del error. Como vemos, incorporar explícitamente los fenómenos de habla espontánea al vocabulario de la tarea mejora el rendimiento del sistema de reconocimiento, aunque, al igual que en DAF, dicha mejora es menos significativa en CORLEC-EHU-1 que en INFOTREN.

En las tablas anteriores, el error de reconocimiento se ha calculado alineando las secuencias *correcta* y reconocida después de filtrar los fenómenos de habla espontánea. Sin embargo, una inspección de las secuencias reconocidas pone de manifiesto que las pseudo-palabras —es decir, los fenómenos de habla espontánea acústicos, y en particular, las pausas de silencio, las pausas habladas y los alargamientos— se reconocen con bastante fiabilidad. A continuación se muestran sendos ejemplos —transcripciones ortográficas con símbolos de fenómenos acústicos (véase la Tabla 5.1)—, tomados de INFOTREN y CORLEC-EHU-1, en los que puede apreciarse cómo pausas y alargamientos podrían utilizarse como *anclas* o referencias en la secuencia acústica, para, por ejemplo, dividirla en trozos y convertir la búsqueda de hipótesis en una secuencia de pequeñas búsquedas:

Ejemplo 6: Intervención tomada de INFOTREN: (a) transcripción correcta, (b) secuencia reconocida incluyendo los FHE en el modelo de lenguaje, y (c) secuencia reconocida sin incluir FHE en el modelo de lenguaje.

(a) G hola buenos días G E mire M quisiera información sobre el horario de trenes para el próximo martes G E a ver salida desde barcelona_sants E M con destino madrid G me interesaría por favor E por la mañana temprano P pero no muy temprano no excesivamente temprano

(b) K G hola buenos días G E mira E M quisiera información sobre el horario de trenes para A el próximo martes G K E sabe salida desde barcelona hasta G tren E de E M con destino madrid G me interesaría por favor desde E por la mañana G en cama P para la vuelta para este solamente tendrá

(c) cuál a buenos días y en y en tercera información sobre el horario de trenes para el próximo martes treinta de salida desde barcelona_sants tren con destino a madrid me interesaría por favor éste por la mañana en granada P P pero no muy tendrá a las que salen de tendrá no P P

Ejemplo 7: Intervención tomada de CORLEC-EHU-1: (a) transcripción correcta, (b) secuencia reconocida incluyendo los FHE en el modelo de lenguaje, y (c) secuencia reconocida sin incluir FHE en el modelo de lenguaje.

(a) G y van W M observando cada A cada panel G y luego ellos mismos hacen unos M A sus sus apuntes

(b) y al público se llamaba cava A catalán G y luego ellos mismos hacen unos M A sus sus apuntes

(c) y mar junto sábado para acá principio ellos mismos hacen unos días sus apuntes

La introducción de modelos acústicos de los fenómenos de habla espontánea, por un lado evita que éstos sean reconocidos erróneamente como palabras, y por otro favorece el reconocimiento correcto de las palabras de su entorno. Esto se observa especialmente en el ejemplo tomado de INFOTREN. En el ejemplo tomado de CORLEC-EHU-1, la secuencia de palabras obtenida en el caso (b) no mejora demasiado con respecto a la obtenida en el caso (c), pero aún así, muchos de los fenómenos se reconocen correctamente. Esto puede deberse por un lado a que estos fenómenos tienen una duración prolongada, y por otro a que disponen de un gran número de muestras en el corpus de entrenamiento, por lo que los modelos resultantes son bastante robustos.

5.3.1. Resultados de reconocimiento con unidades contextuales.

La utilización de los conjuntos de unidades subléxicas contextuales definidos en el apartado 5.2.3 no requiere modificar los modelos de lenguaje, sino tan sólo redefinir los modelos léxicos. Cada palabra del vocabulario deberá transcribirse de forma tan precisa como sea posible utilizando las unidades de estos nuevos conjuntos. Partiendo de las transcripciones fonéticas de las palabras y pseudo-palabras, se han construido modelos léxicos lineales, aplicando las siguientes reglas:

1. Para el primer fonema de la transcripción f_1 , se consulta el fonema siguiente f_2 . Como primera opción se elige el bifonema a derecha f_1+f_2 , si es que pertenece al conjunto de unidades. En caso contrario se elige la unidad incontextual.
2. Para los fonemas intermedios f_i , la primera opción es el trifonema $f_{i-1}-f_i+f_{i+1}$. Si no pertenece al conjunto, se prueba con el bifonema a izquierda $f_{i-1}-f_i$, a continuación con el bifonema a derecha f_i+f_{i+1} , y por último, si ninguno de ellos pertenece al conjunto de unidades, se elige la unidad incontextual.
3. Para el último fonema de la transcripción f_L , se consulta el fonema anterior f_{L-1} . Como primera opción se elige el bifonema a izquierda $f_{L-1}-f_L$, si es que pertenece al conjunto de unidades. En caso contrario se elige la unidad incontextual.
4. Las palabras monofonémicas y las pseudo-palabras se transcriben con la unidad incontextual.

Esta forma de construir los modelos de palabra no es más que una aproximación razonable a modelos más generales —en forma de grafo— basados en unidades contextuales, que incluso podrían representar las variantes de pronunciación más habituales. No cabe duda de que

Tabla 5.19: Error de reconocimiento (%WER) sobre INFOTREN y CORLEC-EHU-1 con MOM de unidades contextuales y modelos de lenguaje k-ESS con $k = 3$ y los fenómenos de habla espontánea integrados. Con objeto de comparar la aportación de las unidades contextuales, se presentan también resultados de reconocimiento sobre BDGEO-8 y BDGEO-16, utilizando modelos de lenguaje k-ESS con $k = 3$, así como los resultados obtenidos con MOM de unidades incontextuales. Entre paréntesis se muestran los valores óptimos de α .

	%WER			
	Unidades contextuales		Unidades incontextuales	
	MOMd	MOMc32g	MOMd	MOMc32g
INFOTREN	32.65 ($\alpha = 5$)	27.22 ($\alpha = 8$)	35.05 ($\alpha = 5$)	29.32 ($\alpha = 8$)
BDGEO-8	11.82 ($\alpha = 5$)	6.96 ($\alpha = 9$)	12.12 ($\alpha = 5$)	7.12 ($\alpha = 9$)
CORLEC-EHU-1: C2 (ML123)	–	28.64 ($\alpha = 10$)	–	36.36 ($\alpha = 10$)
BDGEO-16	–	6.21 ($\alpha = 9$)	–	6.41 ($\alpha = 9$)

la modelización de la variabilidad contextual *dentro* de las palabras mejorará el rendimiento del sistema. Sin embargo, debería tenerse en cuenta la variabilidad contextual no sólo *dentro de* sino también *entre* las palabras —véanse, por ejemplo, (Odell, 1995; Beyerlein et al., 1997). La utilización de unidades contextuales entre palabras da lugar a modelos léxicos mucho más complejos. Como consecuencia, la búsqueda de hipótesis habría de tener en cuenta las restricciones impuestas por la elección de unidades en los bordes, lo que en primera instancia eleva el coste computacional considerablemente. Para moderar el crecimiento del coste computacional se han propuesto diversas técnicas de poda (Beulen et al., 1999; Aubert, 1999) o de minimización del autómata de búsqueda (Mohri et al., 1998). Estos desarrollos sobrepasan los objetivos de este trabajo. Por ello, se ha optado por modelos de palabra con *bordes incontextuales hacia fuera*.

En la Tabla 5.19 se muestran los resultados de reconocimiento obtenidos sobre INFOTREN y CORLEC-EHU-1 utilizando modelos acústicos de unidades contextuales y modelos de lenguaje k-ESS con $k = 3$ y los fenómenos de habla espontánea integrados. Para comparar la aportación de las unidades contextuales sobre habla espontánea con su aportación sobre habla leída, se presentan también los resultados de reconocimiento utilizando modelos acústicos de unidades contextuales para BDGEO-8 (MOM discretos y continuos) y BDGEO-16 (sólo MOM continuos), y modelos de lenguaje k-ESS con $k = 3$. Una observación con respecto a esta comparación: sobre habla leída se maneja el conjunto básico de unidades incontextuales (23 fonemas + unidad de silencio), mientras que sobre habla espontánea se maneja el conjunto ampliado (23 fonemas + 13/14 FHE). Sin embargo, como se ha señalado más arriba, la mayor parte de las unidades correspondientes a FHE no produce variedades contextuales, de modo que, tanto sobre habla leída como sobre habla espontánea, las mejoras deben atribuirse a la modelización de la variabilidad contextual de los fonemas. En la Tabla 5.19 también se presentan, como referencia, los resultados obtenidos con MOM de unidades incontextuales.

La utilización de modelos acústicos contextuales en la construcción de los modelos de palabra tiene como consecuencia, en todos los casos, una reducción del error de reconocimiento. Dicha reducción es tan sólo de entre el 2 % y el 3 % para las bases de datos de habla leída, pero alcanza valores mucho más significativos para las bases de datos de habla espontánea: en torno al 7 % para INFOTREN y superior al 20 % para la partición 2 de CORLEC-EHU-1. Por otra parte, las mejoras relativas son similares independientemente de que se utilicen MOM discretos o MOM continuos: en el caso de INFOTREN, por ejemplo, las reducciones son del 6.85 % y

del 7.16 %, respectivamente. Por último, teniendo en cuenta las adversas condiciones acústicas de CORLEC-EHU-1 —y con todas las reservas que puedan derivarse del hecho de utilizar un modelo de lenguaje *cerrado*— es destacable el relativamente bajo error de reconocimiento (28.64 %), muy próximo al obtenido sobre INFOTREN (27.22 %), siendo ésta una base de datos mucho menos ruidosa. Recordemos, por ejemplo, que en el caso de *Switchboard*, aceptada como una de las bases de datos de habla espontánea más *difíciles* en lengua inglesa, la literatura reporta errores comprendidos entre el 25 % y el 40 %, dependiendo de la complejidad de los sistemas de reconocimiento aplicados (Hain et al., 2000; Ljolje et al., 2000; Colthurst et al., 2000; Stolcke et al., 2000; Sundaram et al., 2000).

Resumen del capítulo

La modelización explícita de los fenómenos de habla espontánea de tipo acústico y su inclusión en el léxico y en el modelo de lenguaje como pseudo-palabras —13 en el caso de INFOTREN, 14 en el caso de CORLEC-EHU-1— mejora significativamente el rendimiento del sistema de reconocimiento. La inclusión de las pseudo-palabras en el corpus de entrenamiento y su integración plena en el modelo de lenguaje desde el principio ha mostrado un mejor rendimiento que si simplemente se añaden mediante suavizado como palabras no vistas. Las mejoras son muy claras en INFOTREN, pero no tanto en CORLEC-EHU-1. Esto puede ser debido en parte al nivel de ruido de las grabaciones de CORLEC-EHU-1, que ecualiza los parámetros acústicos e impide que los modelos acústicos se definan con la suficiente precisión durante el entrenamiento. También hay que tener en cuenta que los fenómenos añadidos suponen algo menos del 3 % en CORLEC-EHU-1, mientras que en INFOTREN suponen alrededor del 6 %. En ambos casos, la cantidad de fenómenos es muy pequeña como para que su modelización implique en sí misma reducciones importantes del error. Sin embargo, su utilización hace que mejore la definición del resto de los modelos, y por tanto también la de los modelos de palabra que con ellos se construyen.

Se ha modelado la variabilidad acústica interna de las unidades subléxicas, en particular la que proviene de la coarticulación de los sonidos. Para ello se han definido conjuntos de unidades contextuales, que incluyen bifonemas a izquierda, bifonemas a derecha y trifonemas con suficientes muestras —más de 300— en el corpus de entrenamiento. Los conjuntos definidos dan cobertura para más del 80 % de las muestras, tanto en el corpus de entrenamiento como en el corpus de test. Se ha comprobado que el uso de estas unidades —junto con las unidades incontextuales, que aseguran la cobertura— incrementa las tasas de reconocimiento, especialmente en DAF. La aportación de los trifonemas es muy poco significativa en las bases de datos más pequeñas (SENGLAR16, SENGLAR08 e INFOTREN), ya que hay muy pocas unidades que superen el número de muestras establecido, pero es mucho más importante en el caso de CORLEC-EHU-1. El incremento absoluto de la tasa de DAF obtenida con unidades contextuales, con respecto al uso de unidades incontextuales, es mayor sobre las bases de datos de habla espontánea, pero en términos relativos las mejoras sobre habla espontánea son similares a las obtenidas sobre habla leída: la reducción del error en DAF está entre el 10 % y el 13 % en todos los casos.

Por último, las mejoras que el uso de unidades contextuales introduce en la tasa de reconocimiento son mucho más significativas en habla espontánea que en habla leída. Sobre BDGEO-16 y BDGEO-8 el error se reduce en tan sólo un 2 % o un 3 %, mientras que en INFOTREN la reducción del error está en torno al 7 %, y en CORLEC-EHU-1 supera el 20 %.

Estos resultados se han obtenido sin llegar a aprovechar del todo el potencial que ofrecen las unidades contextuales en la construcción de modelos léxicos, ya que sólo se modelan los efectos de coarticulación *dentro* de las palabras, y los modelos léxicos lineales sólo recogen una de las posibles realizaciones de cada palabra.

Capítulo 6

Modelización del hablante.

En este capítulo se describe una técnica de *clustering* del conjunto de hablantes que permitirá definir modelos específicos para cada uno de los subconjuntos obtenidos. Con ello se pretende eliminar la variabilidad debida al hablante, de manera que los modelos acústicos den cuenta sólo de las diferencias entre los distintos sonidos de la lengua. Esto debería mejorar la discriminación acústica y evitar, por ejemplo, que fonemas distintos pronunciados por dos hablantes sean tenidos por realizaciones del mismo fonema, o que, por el contrario, dos realizaciones del mismo fonema, producidas por hablantes distintos, sean percibidas como dos fonemas diferentes. Este tipo de confusiones son más probables en habla espontánea, debido a la mayor relajación de los hablantes en la articulación de los sonidos. Por otra parte, la definición de distintos grupos de hablantes podría permitir una caracterización diferenciada del habla espontánea para cada uno de ellos.

En primera instancia se parte del conjunto ampliado de unidades subléxicas incontextuales descrito en el capítulo anterior, pero también se explora la combinación de unidades contextuales con modelos específicos de grupos de hablantes. Se propone un algoritmo de *clustering* basado en el clásico algoritmo de cuantificación vectorial LBG (Linde et al., 1980). Como alternativa a esta metodología se han estimado y aplicado también modelos de género. Finalmente, la utilidad práctica de estas aproximaciones dependerá de la existencia de un método rápido y eficaz para determinar el mejor conjunto de modelos para cada señal de entrada. El propio modelo utilizado para dividir el conjunto de hablantes proporciona una solución simple y elegante a ese problema, ya se trate de grupos obtenidos automáticamente o de grupos definidos *a priori* —como en el caso de los modelos de género.

6.1. Alternativas de modelización.

La utilización de modelos específicos para cada hablante da mejores resultados que utilizar un único conjunto de modelos independientes del hablante, hasta tal punto que la tasa WER puede reducirse en un 50 % (Hazen, 1998, apartado 3.3). Sin embargo, en la mayor parte de las aplicaciones el usuario final no está incluido en el corpus de entrenamiento, por lo que es imposible entrenar modelos específicos, de modo que o bien se modifican los parámetros acústicos del nuevo hablante para normalizarlos con respecto a un conjunto de modelos genéricos, o bien son los modelos los que deben adaptarse al nuevo hablante. A continuación repasamos con cierta extensión los trabajos más destacables en ambas direcciones (una revisión más completa puede encontrarse en (Woodland, 2001)).

6.1.1. Normalización del hablante.

Las diferencias entre hablantes pueden ser de distinta índole. Fundamentalmente se trata de diferencias fisiológicas, relacionadas con la longitud y geometría del tracto vocal o con el tamaño de las cuerdas vocales, pero también encontramos diferencias que provienen del aprendizaje de la lengua, que afectan a los modos de articulación, a la entonación, etc. Generalmente se trata de diferencias dialectales, es decir, atribuibles a la forma de interpretar la lengua de un grupo humano amplio. Por otra parte, ciertos hablantes pueden mostrar defectos o características especiales que les distinguen del resto. Finalmente, pueden darse variaciones de tipo local, como cambios en la velocidad del habla, cambios en el estado emocional del hablante, etc.

Un posible planteamiento consiste en suponer la existencia de un *hablante canónico*, con respecto al cual los demás hablantes *se diferencian*. Desde este punto de vista, los parámetros acústicos de cualquier hablante estarían distorsionados con respecto a los de ese *hablante canónico*, por lo que sería necesario aplicar técnicas que compensaran esa distorsión. Esto es lo que se conoce como *normalización del hablante*.

Por ejemplo, la técnica conocida como *Cepstral Mean Normalization* (CMN) (Rosenberg et al., 1994), que consiste en restar a los cepstrales su media —calculada en un período largo de tiempo—, se utiliza para compensar en parte la variabilidad debida a características propias del hablante, si bien incrementa también la robustez frente a cambios en el ruido ambiental y de canal.

6.1.1.1. Normalización de la longitud del tracto vocal.

La más conocida y aplicada de este tipo de técnicas es la normalización de la longitud del tracto vocal (*Vocal Tract Length Normalization*, VTLN) (Lee and Rose, 1996; Eide and Gish, 1996; Wegmann et al., 1996; Zhan and Westphal, 1997; Welling et al., 1999). Típicamente las mujeres tienen trectos vocales más cortos que los hombres, lo cual afecta a la posición de los formantes, que es inversamente proporcional a la longitud del tracto vocal y que puede llegar a variar hasta un 25 % de unos hablantes a otros. Los modelos de género podrían compensar en parte estas diferencias, pero no tienen en cuenta la variabilidad intra-género. La VTLN es una transformación no lineal de los parámetros acústicos, que depende de un único parámetro α . Suele manejarse una versión simplificada de la transformación, lineal a trozos, que se aplica en el espacio de la frecuencia, aunque recientemente han aparecido versiones que la aplican directamente sobre los coeficientes cepstrales (Pitz et al., 2001; Emori and Shinoda, 2001). La VTLN considera un número finito de posibles transformaciones —es decir, un número finito de posibles valores de α —, cuyo efecto es comprimir (en el caso de las mujeres) o expandir (en el caso de los hombres) las componentes frecuenciales.

Durante el reconocimiento se utiliza un modelo simplificado —típicamente una mezcla de gaussianas— para determinar qué transformación es la más adecuada, en el sentido de maximizar la probabilidad de las muestras dado el modelo. Este modelo simplificado no requiere obtener una decodificación, y por tanto no depende de la calidad de la misma. Sin embargo, sí le pueden afectar el ruido de fondo o cambios en las condiciones del canal de transmisión, que provocarían errores en la determinación del mejor α . Por otra parte, siempre que la aplicación lo permite, se asigna un único valor óptimo de α a cada hablante, pero en muchas ocasiones no se conoce de antemano la identidad del hablante, y no queda más remedio que asignar un α distinto a cada intervención, lo cual incrementa el coste de normalización.

6.1.2. Adaptación al hablante.

En la literatura pueden encontrarse diversas metodologías de adaptación al hablante. Una solución obvia consiste en adquirir muestras del nuevo hablante, en cantidad suficiente como para reestimar los parámetros de los modelos. Tal era el caso de los primeros sistemas de dictado, que requerían del hablante que repitiera un conjunto de frases, tarea que podía llevar incluso horas. Esta solución resultaba poco ergonómica, pero merecía la pena, ya que el sistema de dictado estaba diseñado para un único usuario, y los modelos adaptados proporcionaban un rendimiento significativamente mejor que los modelos genéricos. Sin embargo, muchas aplicaciones no están ligadas indisolublemente a un único usuario, sino que trabajan sucesivamente con un grandísimo número de usuarios —tal es el caso, por ejemplo, de las máquinas expendedoras de billetes o los sistemas de información al gran público—, por lo que no es posible realizar un proceso de adaptación como el descrito. Así pues, desde principios de los 90 las investigaciones se encaminaron hacia métodos de adaptación gradual o instantánea de los modelos acústicos, que a partir de muy pocos datos fueran capaces de obtener rendimientos cercanos a los de un sistema dependiente del hablante. En los siguientes apartados damos cuenta brevemente de los más destacables.

6.1.2.1. Adaptación MAP.

El criterio de máxima similitud maximiza la probabilidad de las muestras de entrenamiento dado el modelo, $p(x|\lambda)$, pero si se tienen muy pocos datos las estimaciones resultantes son incompletas y muy poco robustas. Para obtener estimaciones más robustas, el criterio MAP (*Maximum A Posteriori*) (Gauvain and Lee, 1994) maximiza la cantidad $p(x|\lambda)p_0(\lambda)$, donde $p_0(\lambda)$ es la distribución *a priori* de los parámetros, que ha sido estimada a partir de una gran cantidad de datos. En este caso, $p_0(\lambda)$ representa a los modelos independientes del hablante. La adaptación MAP consiste en *inclinarse* ligeramente los parámetros de esa distribución general hacia el nuevo hablante, utilizando para ello los pocos datos disponibles. Con MOM continuos se suelen reestimar sólo las medias de las gaussianas y los pesos de las componentes en las mezclas. En la práctica, si se dispone de una muestra $O = O_1 O_2 \dots O_T$ perteneciente al hablante para el que se desean adaptar los modelos, las fórmulas de reestimación de las medias de las gaussianas tienen la forma siguiente:

$$\hat{\mu} = \frac{\tau\mu_0 + \sum_{t=1}^T \gamma(t)O_t}{\tau + \sum_{t=1}^T \gamma(t)} \quad (6.1)$$

donde μ_0 es la media *a priori*, $\gamma(t)$ la probabilidad de la gaussiana en el instante t , O_t la observación acústica en el instante t , y τ un parámetro fijado heurísticamente —llamado *factor de adaptación*— que determina el peso relativo de la distribución *a priori* con respecto a la estimación de los parámetros según el criterio de máxima similitud. Típicamente se utilizan valores de τ comprendidos entre 2 y 20.

Este método converge a las estimaciones por máxima similitud a medida que aumenta el número de muestras. Es decir, con un número suficientemente grande de muestras se obtendrían modelos dependientes del hablante. Sin embargo, en la práctica sólo se dispone de unas pocas muestras y un buen número de parámetros de la distribución *a priori* podrían permanecer inalterados. Esto hace que la adaptación sea lenta y parcial: los modelos resultantes podrían ser todavía demasiado generales. Para mejorarla se han propuesto varios métodos que establecen relaciones entre los parámetros de los modelos, de modo que todos ellos son reestimados,

de forma directa o indirecta, a partir de los datos de adaptación. Entre ellos destacamos los siguientes:

- ▷ **Predicción de los parámetros por regresión lineal** (Cox, 1995; Ahadi and Woodland, 1997). Se trata de obtener correlaciones entre los parámetros de los MOM, y utilizarlas para actualizar mediante regresión lineal aquellos parámetros que no han recibido suficiente número de muestras. Para ello, en primer lugar se generan modelos dependientes del hablante y a continuación se buscan grupos de parámetros que experimenten cambios similares al cambiar la identidad del hablante. Esta búsqueda de correlaciones es muy costosa, y produce como resultado una fórmula de regresión lineal para cada parámetro. Durante el reconocimiento, en primera instancia se aplica adaptación MAP; a continuación se toman únicamente los parámetros que han recibido suficientes muestras de adaptación para generar predicciones lineales del resto de parámetros; finalmente, el valor de estos parámetros se obtiene mediante una combinación ponderada de su estimación MAP y el valor obtenido por regresión lineal.
- ▷ **MAP estructurado** (Shinoda and Lee, 1997). Consiste en organizar las gaussianas de los MOM continuos en forma de árbol. Cada nodo modeliza únicamente la desviación con respecto a la distribución del nodo padre, mediante un término aditivo en las medias y un factor de escala en las varianzas. El nodo raíz es un caso especial, ya que agrupa a todas las gaussianas y contiene por tanto la auténtica distribución *a priori* $p_0(\lambda)$.

6.1.2.2. Transformación de los parámetros de los modelos.

En este caso lo que se plantea es una transformación de los parámetros de las gaussianas de MOM continuos, para compensar la diferencia entre las condiciones de entrenamiento y las condiciones de test. La ventaja de este método es que la misma transformación puede ser aplicada a un gran número de parámetros, o incluso a todos ellos. Durante el reconocimiento, unos pocos datos de adaptación son suficientes para estimar los parámetros de la transformación, de modo que la adaptación es rápida y afecta a todos los parámetros. Varias aproximaciones, muy similares entre sí, aplican este esquema. La más utilizada es la que se conoce como MLLR (*Maximum Likelihood Linear Regression*) (Leggetter and Woodland, 1995b), en la que las medias de las gaussianas se ajustan mediante la siguiente transformación lineal:

$$\hat{\mu} = \mathbf{A}\mu + b \quad (6.2)$$

La matriz \mathbf{A} y el vector b se estiman mediante el algoritmo *Expectation-Maximization* (Dempster et al., 1977) a partir de los datos de adaptación. Si se tuvieran datos suficientes, podría llegar a definirse una transformación distinta (\mathbf{A}, b) para cada gaussiana, pero lo más habitual es que un gran número de gaussianas —o incluso todas ellas— compartan una misma transformación. Esto permite adaptar *todos* los parámetros, aunque de forma menos precisa, utilizando muy pocos datos de adaptación.

Para alcanzar un mejor compromiso entre la robustez y la precisión de la adaptación, las gaussianas se agrupan por cercanía en el espacio de parámetros, lo cual produce un conjunto de grupos que forman las hojas de un árbol. A continuación las hojas se van fundiendo en nodos padre de forma jerárquica, con el mismo criterio de cercanía acústica, hasta alcanzar un único nodo raíz que agrupa a todas las gaussianas de los modelos. Para llevar a cabo la adaptación, se desciende desde el nodo raíz hasta la capa de nodos más profunda que sea posible, lo cual

dependerá de la cantidad de datos de adaptación disponibles. Finalmente, en cada nodo se estima una transformación distinta, que se aplica al conjunto de gaussianas correspondiente (Leggetter and Woodland, 1995a).

Los parámetros que parecen estar más específicamente determinados por la identidad del hablante son las medias de las gaussianas. Sin embargo, algunas aproximaciones tratan de adaptar también las varianzas, bien de forma independiente (Gales and Woodland, 1996), bien de forma conjunta con las medias (Digalakis et~al., 1995; Gales, 1998). Finalmente, existen múltiples variantes del esquema de adaptación MLLR que tratan de mejorar su robustez. En general, esto se logra mediante la combinación de estimaciones MAP y MLLR (Chesta et~al., 1999; Chou, 1999; Goronzy and Kompe, 1999; Siohan et~al., 2002).

6.1.2.3. *Clustering de hablantes.*

Consiste en identificar dentro del corpus de entrenamiento grupos o subconjuntos de hablantes con características similares, y entrenar a continuación modelos separados para cada uno de ellos. Después, durante el reconocimiento, para cada señal debe decidirse cuál es el grupo de hablantes más adecuado, y aplicar el conjunto de modelos correspondiente. La adaptación se concreta, por tanto, en tomar dicha decisión. En el peor caso, ello puede requerir que la señal de entrada sea reconocida con todos los conjuntos de modelos, dando por buena la decodificación más probable.

El ejemplo más sencillo de este tipo de aproximación son los modelos de género, que han sido ampliamente utilizados en sistemas de reconocimiento de habla leída (Gauvain et~al., 1994; Bahl et~al., 1995). Además del género, otras características —también conocidas de antemano—, como el acento o la variedad dialectal, podrían utilizarse como criterio para clasificar los hablantes en distintos grupos. Sin embargo, habitualmente son las propias muestras de entrenamiento las que se utilizan para determinar, tras un proceso de partición o agrupamiento automáticos, qué partición del conjunto de hablantes es la más adecuada (Kosaka and Sagayama, 1994; Sankar et~al., 1995; Gao et~al., 1997; Naito et~al., 2002).

Definir grupos de hablantes implica fragmentar el corpus de entrenamiento y ello puede conducir a estimaciones poco robustas. Por ello, las técnicas de partición/agrupamiento del conjunto hablantes suelen combinarse con entrenamiento MAP y/o con una transformación de los modelos —tipo MLLR— que permitan obtener modelos adaptados a partir de pocas muestras (Padmanabhan et~al., 1996; Gao et~al., 1997; Jiang and Huang, 2000; Zhang and Furui, 2000).

Por otra parte, cuando se trata de reconocer la intervención de un nuevo hablante —no presente en el corpus de entrenamiento—, tomar una decisión *fuerte* sobre cuál de los grupos es el más adecuado podría no ser la mejor opción. Recientemente se han presentado trabajos en los que cada hablante se modela mediante una cierta combinación de grupos de hablantes. Cada grupo puede verse como un eje en el espacio de hablantes, de manera que los pesos que conforman la combinación determinarían un punto en ese espacio. Una vez estimado el peso de cada grupo, se construye un modelo o una transformación específicos para cada hablante, y se utilizan para reconocer todas sus intervenciones (Gales, 2000; Hazen, 2000; Pusateri and Hazen, 2002). En estrecha relación con estos trabajos se ha desarrollado también una técnica de adaptación al hablante basada en lo que se conoce como *voces canónicas* o *eigenvoices* (Kuhn et~al., 2000).

Finalmente, algunos de los algoritmos de *clustering* propuestos en los últimos años no tratan de identificar la mejor partición de un conjunto *conocido* de hablantes, sino que ope-

ran en un conjunto de segmentos no etiquetados por hablante —por ejemplo, en conversaciones multi-persona tomadas de radio o televisión— y tratan de encontrar grupos de segmentos que muestren una cierta coherencia interna, grupos que se etiquetarían como pertenecientes al mismo hablante (Jin et al., 1997; Johnson and Woodland, 1998; McLaughlin et al., 1999; Ajmera et al., 2002). Después, en una segunda fase, podría aplicarse algún esquema de adaptación —típicamente MLLR— sobre los grupos obtenidos.

6.2. Adaptación mediante *clustering* de hablantes.

La adaptación al hablante mediante *clustering* requiere llevar a cabo tres tareas: (1) definir una partición del conjunto de hablantes según un criterio de proximidad acústica; (2) entrenar modelos de Markov específicos para cada uno de los grupos obtenidos; y (3) determinar un mecanismo de adaptación al hablante durante el reconocimiento, basado en dichos modelos.

Para llevar a cabo la partición del conjunto de hablantes se ha definido un modelo acústico simplificado, basado en las distribuciones discretas —histogramas de etiquetas VQ— obtenidas empíricamente para los hablantes del corpus de entrenamiento, en las cuatro representaciones acústicas consideradas. A partir de este modelo se define una distancia entre hablantes o grupos de hablantes que no requiere obtener la decodificación de las muestras, y a partir de dicha distancia se define el algoritmo de *clustering*, que es una variación del LBG. Finalmente, dada una señal, la elección del conjunto de modelos de Markov más adecuado puede hacerse de forma inmediata mediante el modelo simplificado, aunque, como se verá en apartados posteriores, es más fiable —y mucho más costoso— el uso de los propios modelos de Markov. A continuación definimos formalmente el modelo de hablante, las distancias entre hablantes y entre grupos de hablantes, el algoritmo de *clustering* y el resto de elementos de esta aproximación.

6.2.1. Un modelo acústico simplificado para grupos de hablantes.

Sea M el número de representaciones acústicas —que suponemos independientes entre sí— y $X(t) = (X_1(t), \dots, X_M(t))$ la observación acústica en el instante t . En primer lugar, el volumen de información se reduce mediante cuantificación vectorial: en cada representación acústica se construye un diccionario VQ utilizando las muestras del corpus de entrenamiento y aplicando el algoritmo LBG con distancia euclídea en el espacio de los parámetros acústicos. En realidad, como una de nuestras líneas de desarrollo se basa en MOM discretos, estos diccionarios ya han sido construidos. Sea N el tamaño de cada uno de los diccionarios. A continuación cada observación acústica $X(t)$ puede ser reemplazada por $Y(t) = (Y_1(t), \dots, Y_M(t))$, donde $Y_j(t) \in \{1, \dots, N\}$ es el índice del centroide más cercano al vector $X_j(t)$ en la representación acústica j .

La cuantificación vectorial de las observaciones acústicas introduce una cierta distorsión, que *suaviza* las diferencias entre las observaciones acústicas correspondientes a dos hablantes distintos. Además, si se utiliza una distancia euclídea, como en este caso, el proceso de etiquetado implícitamente ecualiza las varianzas de los centroides del diccionario VQ. Esta circunstancia no es importante desde el punto de vista de adaptación al hablante, ya que las características propias del hablante afectan principalmente a las medias de las gaussianas de las mezclas que conforman las distribuciones de emisión en MOM continuos, y en menor medida a las varianzas, que no suelen adaptarse. Por otra parte, si entre dos hablantes existen diferencias importantes, aunque el etiquetado VQ produzca un cierto suavizado —o distorsión, depende de cómo se

mire— esas diferencias deben quedar reflejadas en la distribución estadística de etiquetas. Esta es la hipótesis de partida para la definición del modelo de hablante.

Suponiendo que el corpus de entrenamiento está organizado por hablantes, considerese el hablante i , del que se tienen $s(i)$ etiquetas acústicas. Sea $s(k, j, i)$ el número de veces que aparece la etiqueta k en la representación acústica j para el hablante i :

$$s(k, j, i) = \sum_{l=1}^{L_i} \sum_{t=1}^{T_l} \delta(Y_j^{(i,l)}(t) = k) \quad (6.3)$$

$$s(i) = \sum_{l=1}^{L_i} T_l \quad (6.4)$$

donde L_i es el número de intervenciones del hablante i , T_l el número de etiquetas acústicas de la intervención l , e $Y_j^{(i,l)}(t)$ la etiqueta observada en el instante t en la representación acústica j de la intervención l del hablante i .

Se define $P_j(k|i)$ como la probabilidad de observar la etiqueta k en la representación acústica j , dado el hablante i . Si se utiliza el criterio de máxima similitud, esta distribución puede estimarse como sigue:

$$\hat{P}_j(k|i) = \frac{s(k, j, i)}{s(i)} \quad (6.5)$$

Evidentemente, cuantas más muestras dispongamos del hablante i , más robusta será la estimación $\hat{P}_j(k|i)$. Sin embargo, aún con un número grande de muestras, podría darse el caso de que algún *bin* del histograma fuera nulo. Para evitar probabilidades nulas, y también como forma primitiva de suavizado, se propone una fórmula de estimación alternativa:

$$s'(k, j, i) = s(k, j, i) + 1 \quad (6.6)$$

$$s'(i) = s(i) + N \quad (6.7)$$

$$\hat{P}_j(k|i) = \frac{s'(k, j, i)}{s'(i)} \quad (6.8)$$

Nótese que $\forall j \sum_{k=1}^N s'(k, j, i) = s'(i)$, de manera que $\sum_{k=1}^N \hat{P}_j(k|i) = 1$. Si se admite que $P_j(k|i) \approx \hat{P}_j(k|i)$, entonces la probabilidad de una muestra cuantificada $Y(t)$, dado el hablante i , será:

$$P(Y(t)|i) \approx \prod_{j=1}^M \hat{P}_j(Y_j(t)|i) \quad (6.9)$$

Finalmente, suponiendo que observaciones acústicas sucesivas son independientes entre sí —hipótesis que aplican también los MOM—, la probabilidad de una secuencia acústica cuantificada $Y = \{Y(1), \dots, Y(T)\}$, dado un hablante i , vendrá dada por:

$$P(Y|i) = \prod_{t=1}^T P(Y(t)|i) \quad (6.10)$$

Hasta aquí el modelo se ha interpretado como *modelo de hablante*. Sin embargo, si se utilizan las muestras de un grupo de hablantes g para calcular los histogramas:

$$c(k, j, g) = \sum_{i \in g} \sum_{l=1}^{L_i} \sum_{t=1}^{T_l} \delta(Y_j^{(i,l)}(t) = k) = \sum_{i \in g} s(k, j, i) \quad (6.11)$$

$$c(g) = \sum_{i \in g} \sum_{l=1}^{L_i} T_l = \sum_{i \in g} s(i) \quad (6.12)$$

entonces el *modelo de grupo de hablantes* se puede estimar mediante una expresión análoga a la 6.5:

$$\hat{P}_j(k|g) = \frac{c(k, j, g)}{c(g)} \quad (6.13)$$

En la práctica, los modelos de grupo se calculan a partir de los modelos de hablante suavizados, de modo que si cada grupo g consta de $H(g)$ hablantes, el modelo suavizado de grupo se calcula como sigue:

$$c'(k, j, g) = \sum_{i \in g} s'(k, j, i) = \sum_{i \in g} s(k, j, i) + H(g) \quad (6.14)$$

$$c'(g) = \sum_{i \in g} s'(i) = \sum_{i \in g} s(i) + NH(g) \quad (6.15)$$

$$\hat{P}_j(k|g) = \frac{c'(k, j, i)}{c'(i)} \quad (6.16)$$

En adelante, para no complicar la notación, se supondrá que todos los histogramas de etiquetas VQ, ya correspondan a hablantes o a grupos de hablantes, han sido suavizados mediante las fórmulas 6.6, 6.7, 6.14 y 6.15, y se les denominará $s(k, j, i)$, $s(i)$, $c(k, j, g)$ y $c(g)$, respectivamente.

6.2.2. Medida de disimilitud entre un hablante y un grupo de hablantes.

Supóngase que el modelo de cada hablante i ha sido ya construido a partir del conjunto de muestras $\Upsilon(i) = \{Y^{(i,l)} | l = 1, \dots, L_i\}$, y que se tiene una cierta partición del conjunto de hablantes, $\Pi = \{g_1, \dots, g_G\}$, cuyos modelos han sido construidos también. Entonces la *disimilitud* entre un cierto hablante i y un grupo de hablantes g se define como sigue:

$$d(i; g) = -\log \left(\frac{P(\Upsilon(i)|g)}{P(\Upsilon(i)|i)} \right) \quad (6.17)$$

donde $P(\Upsilon(i)|\cdot)$ es la probabilidad conjunta de las muestras del hablante i , condicionada por el modelo simplificado de un hablante o de un grupo de hablantes.

Haciendo cuentas, obtenemos una expresión equivalente en la que sólo aparecen los parámetros de los respectivos modelos de histogramas:

$$d(i; g) = \sum_{j=1}^M \sum_{k=1}^N s(k, j, i) \log \left(\frac{s(k, j, i)}{c(k, j, g)} \right) - Ms(i) \log \left(\frac{s(i)}{c(g)} \right) \quad (6.18)$$

6.2.3. Medida de disimilitud entre dos hablantes.

Si en la ecuación 6.18 el grupo g consta de un único hablante, lo que se obtiene es una medida de *disimilitud* entre los modelos de dos hablantes. Sin embargo, $d(i_1; i_2) \neq d(i_2; i_1)$. Para conseguir la conmutatividad, se *simetriza* la ecuación 6.17, como sigue:

$$\begin{aligned} D(i_1, i_2) &= d(i_1; i_2) + d(i_2; i_1) \\ &= -\log \left(\frac{P(\Upsilon(i_1)|i_2)P(\Upsilon(i_2)|i_1)}{P(\Upsilon(i_1)|i_1)P(\Upsilon(i_2)|i_2)} \right) \end{aligned} \quad (6.19)$$

Nótese que $D(i_1, i_2) \geq 0$, ya que $P(\Upsilon(a)|b) \leq P(\Upsilon(a)|a)$, $\forall b \neq a$, puesto que el modelo del hablante a se ha estimado mediante el criterio de máxima similitud sobre el conjunto $\Upsilon(a)$. Sólo puede ser $P(\Upsilon(a)|b) = P(\Upsilon(a)|a)$ si los modelos de los hablantes a y b son iguales.

Nuevamente, si se hacen cuentas, resulta una expresión equivalente en la que sólo aparecen los parámetros de los modelos de los dos hablantes i_1 y i_2 :

$$D(i_1, i_2) = \sum_{j=1}^M \sum_{k=1}^N (s(k, j, i_1) - s(k, j, i_2)) \log \left(\frac{s(k, j, i_1)}{s(k, j, i_2)} \right) - M(s(i_1) - s(i_2)) \log \left(\frac{s(i_1)}{s(i_2)} \right) \quad (6.20)$$

La nueva medida $D(a, b)$ es no negativa, siendo nula sólo si los modelos de a y b son iguales; además $D(a, b) = D(b, a)$. De hecho, $D(a, b)$ cumple todas las propiedades de una métrica *entre modelos de hablantes*, salvo la desigualdad triangular. Es fácil encontrar un contraejemplo. Dados los modelos de dos hablantes a y c , considerese el modelo de un hipotético tercer hablante b , situado justo en el punto medio entre ambos, es decir, $\forall k \in [1, M]$ y $\forall j \in [1, N]$ se tendría que:

$$s(k, j, b) = \frac{1}{2}(s(k, j, a) + s(k, j, c))$$

En tal caso, se tiene:

$$D(a, b) + D(b, c) - D(a, c) = \frac{1}{2} \sum_{j=1}^M \sum_{k=1}^N (s(k, j, c) - s(k, j, a)) \log \frac{s(k, j, a)}{s(k, j, c)} - \frac{1}{2} M(s(c) - s(a)) \log \frac{s(a)}{s(c)} \leq 0 \quad (6.21)$$

Por tanto, no podemos decir que $D(a, b)$ sea una *distancia*, sino tan sólo una medida de la proximidad acústica entre los hablantes a y b , y en adelante nos referiremos a ella como *disimilitud* entre a y b .

6.2.4. El algoritmo de *clustering*.

Se ha aplicado un algoritmo de *clustering* de tipo divisivo, basado en el clásico LBG (Linde et-al., 1980). Inicialmente todos los hablantes forman un único grupo. A continuación, dado un conjunto de grupos, uno de ellos se divide en dos subgrupos, e iterativamente se reasignan los hablantes a los grupos y se reestiman los modelos de éstos, hasta que no se producen modificaciones en la asignación de hablantes. El algoritmo finaliza o bien cuando no se dispone de suficientes muestras o hablantes para generar nuevos grupos, o bien cuando la distorsión promedio de la partición no decrece lo suficiente.

El grupo a dividir será aquél que maximice una cierta medida de distorsión interna. Para ello, en primer lugar es necesario definir dicha medida. Dado un grupo de hablantes g , se toma como representante del mismo a aquel hablante cuya distancia promedio al resto de hablantes del grupo sea mínima, es decir,

$$r(g) = \arg \min_{i \in g} \{\bar{D}(i; g)\} \quad (6.22)$$

donde

$$\bar{D}(i; g) = \frac{1}{H(g) - 1} \sum_{i' \in g, i' \neq i} D(i, i') \quad (6.23)$$

Como medida de distorsión interna de un grupo g se toma precisamente el valor de $\bar{D}(r(g); g)$.

Sea $g^* = \arg \max_{g \in \Pi} \{\bar{D}(r(g); g)\}$. Los modelos iniciales de los dos grupos, g_1^* y g_2^* , en los que se divide g^* se obtienen del modelo de $r(g^*)$ por un lado, y del modelo del hablante de g^* más cercano a $r(g^*)$, $r'(g^*) = \arg \min_{i \in g^*, i \neq r(g^*)} \{D(i, r(g^*))\}$, por otro (véase la Figura 6.1).

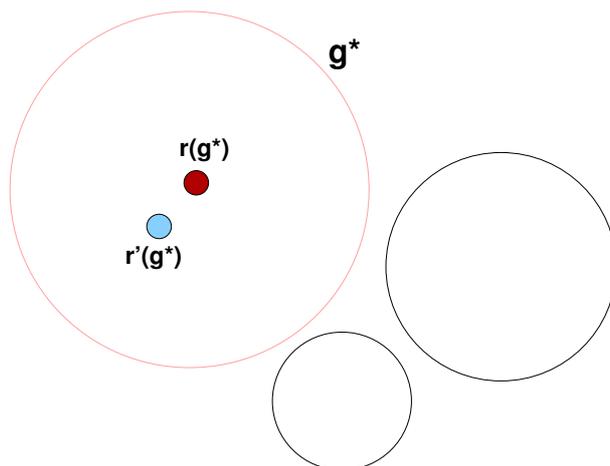


Figura 6.1: Se elige como grupo a dividir aquél, g^* , cuya distorsión interna sea máxima. Los dos grupos en los que se divide g^* se inicializan con los modelos del representante $r(g^*)$ y del hablante más próximo a éste, $r'(g^*)$.

Por último, como criterio de parada del algoritmo se utiliza la reducción relativa de la distancia promedio de un hablante al grupo al que pertenece. Sea $g(i, n)$ el grupo al que pertenece el hablante i una vez alcanzada la convergencia sobre la partición $\Pi(n)$ en una cierta iteración n del algoritmo. La distancia promedio de un hablante a su grupo en la iteración n se define:

$$R(n) = \frac{1}{C} \sum_{i=1}^H d(i; g(i, n)) = -\log \left[\prod_{i=1}^H \frac{P(\Upsilon(i)|g(i, n))}{P(\Upsilon(i)|i)} \right]^{1/C} \quad (6.24)$$

donde H es el número total de hablantes y C el número total de etiquetas acústicas en el corpus de entrenamiento. En un caso extremo, cada grupo podría estar formado por un único hablante, y la distancia promedio de un hablante a su grupo sería nula. Sin embargo, los modelos acústicos que se obtendrían de esa partición extrema no serían lo bastante robustos. De ahí que sea necesario establecer un compromiso entre el número de muestras y la coherencia interna de los grupos de hablantes.

En la Figura 6.2 se describe paso a paso el algoritmo de *clustering*. En dicho algoritmo, el parámetro $\rho > 0$, establecido heurísticamente, determina a partir de qué punto el decrecimiento relativo en la distorsión interna promedio entre dos iteraciones sucesivas $n - 1$ y n es considerado insuficiente. Además de este criterio, se aplica implícitamente otro, y es que cada vez que se ejecuta la división de un grupo, la partición resultante sólo se considera válida si los grupos resultantes tras las iteraciones de convergencia contienen todos ellos un número mínimo de hablantes y un número mínimo de muestras. Esto asegura un cierto grado de robustez de los modelos acústicos. Si una partición no se considera válida, se recupera la partición anterior y se

prueba con el siguiente grupo en la lista de candidatos l_{cand} . Como los grupos están dispuestos en orden decreciente de distorsión interna, siempre se prueba a dividir en primer lugar el grupo con mayor distorsión interna.

6.2.5. Estimación de los MOM de grupos de hablantes.

Una vez que el corpus de entrenamiento se ha dividido automáticamente en, digamos, G grupos de hablantes, es necesario estimar modelos acústicos específicos para cada uno de ellos. Una opción es aplicar directamente el mismo procedimiento empleado para los modelos generales, es decir, el algoritmo de Baum-Welch en el caso de MOM discretos o el algoritmo de entrenamiento por Viterbi en el caso de MOM continuos. Dependiendo del número de muestras disponibles, esto puede producir estimaciones muy poco robustas, especialmente con MOM continuos. Por ello, se ha aplicado el procedimiento de reestimación MAP (Gauvain and Lee, 1994), utilizando como distribución a priori la de los modelos generales y explorando distintos valores del factor de adaptación τ . En el caso de MOM discretos se han reestimado únicamente las probabilidades de emisión. En el caso de MOM continuos se han reestimado las medias de las gaussianas y los pesos de las mezclas.

La reestimación MAP puede verse como una interpolación entre los modelos generales y los modelos que se obtendrían con una estimación por máxima similitud. Véase, por ejemplo, la fórmula 6.1 de reestimación de las medias, en el apartado 6.1.2.1. Cuanto más alto es el valor de τ más peso se adjudica a los modelos generales, mientras que si $\tau = 0$, lo que se obtiene es una estimación por máxima similitud. Por otra parte, fijado un cierto $\tau > 0$, si se dispone de un gran número de muestras, la estimación por máxima similitud adquiere un peso importante, mientras que si el número de muestras es muy pequeño, el valor resultante prácticamente coincide con el de la distribución a priori.

6.2.6. Alternativas de adaptación al hablante.

En este trabajo, la adaptación al hablante se efectúa de forma independiente para cada señal de entrada, porque en las aplicaciones que nos interesan —acceso a un sistema de información para el gran público o entrevistas en medios de comunicación— es imposible conocer a priori la identidad del hablante. Este es el principal argumento para optar por la partición del conjunto de hablantes en grupos más o menos homogéneos como estrategia de adaptación. Otras estrategias requieren una mayor cantidad de información sobre el hablante, mientras que en esta aproximación bastan apenas unas décimas de segundo de señal para identificar el grupo de hablantes más próximo. Se han considerado varias formas de adaptación al hablante, que se describen en los siguientes apartados.

6.2.6.1. Selección de MOM basada en las probabilidades de las decodificaciones.

Este es el mecanismo de adaptación más costoso computacionalmente, ya que requiere efectuar G pasadas de reconocimiento distintas, una con los MOM correspondientes a cada grupo de hablantes. De las decodificaciones resultantes se escoge como óptima la que arroje un valor más alto de probabilidad. El coste de esta aproximación es G veces superior al de utilizar un único conjunto de modelos genéricos. Como contrapartida, si la estimación de los modelos acústicos es suficientemente robusta, cabe esperar mejoras notables. En adelante, nos referiremos de forma abreviada a esta alternativa como *Max-Prob*.

Algoritmo *Partición***Entradas:**

Conjunto de muestras de los hablantes: $\Omega = \{\Upsilon(i) \mid i = 1, 2, \dots, H\}$

Salidas:

Partición: $\Pi = \{g_l \mid g_l \neq \emptyset, l = 1, 2, \dots, G\}$

Conjunto de modelos: $\Lambda = \{c(k, j, l) \mid k = 1, 2, \dots, N; j = 1, 2, \dots, M; l = 1, 2, \dots, G\}$

1. A partir del conjunto de muestras cuantificadas de cada hablante i , acumular los histogramas correspondientes a cada representación acústica j , $s(k, j, i)$, y suavizarlos según la fórmula 6.6. A partir de ellos calcular el factor de normalización $s(i)$.
 2. Calcular y almacenar la disimilitud entre cada par de hablantes, $D(i_1, i_2)$. Sólo es necesario calcular $H(H - 1)/2$ valores, ya que $D(i, i) = 0$ y $D(i_1, i_2) = D(i_2, i_1)$.
 3. Inicialización:
 - $G = 1$: $c(k, j, 1) = \sum_{i=1}^H s(k, j, i)$ y $c(1) = \sum_{i=1}^H s(i)$
 - $n = 0$: calcular $R(0)$
 4. **repetir**
 - a) $n \leftarrow n + 1$
 - b) Para cada grupo $g \in \Pi(n)$, obtener el representante $r(g)$, el hablante más cercano al representante, $r'(g)$, así como la distorsión interna promedio $\bar{D}(r(g); g)$. Almacenar esta información en una lista de grupos candidatos a ser divididos, l_{cand} , en orden decreciente de $\bar{D}(r(g); g)$.
 - c) **mientras** $l_{cand} \neq \emptyset$ **hacer**
 - 1) Extraer el primer ítem de l_{cand} , $(g_d, r(g_d), r'(g_d))$, y dividir g_d en dos subgrupos, cuyos modelos se inicializan con los de los hablantes $r(g_d)$ y $r'(g_d)$.
 - 2) **repetir**
 - ▷ Asignar cada hablante i al grupo más cercano, es decir, a aquel g para el que la distancia $d(i; g)$ sea mínima.
 - ▷ Recalcular el modelo de cada grupo a partir de los histogramas de los hablantes que le han sido asignados, aplicando las fórmulas 6.11 y 6.12.
- hasta** número máximo de iteraciones **OR** grupos inalterados
- 3) **si** la partición es válida **entonces**
 - ▷ $G \leftarrow G + 1$
 - ▷ Calcular $R(n)$
 - ▷ Vaciar l_{cand}
 - si_no**
 - ▷ Recuperar la partición correspondiente a la iteración $n - 1$
 - ▷ $R(n) \leftarrow R(n - 1)$
- fsi**
- fmientras**
- hasta** $(R(n - 1) - R(n))/R(n) < \rho$

Figura 6.2: Algoritmo de *clustering*.

6.2.6.2. Selección de MOM basada en las probabilidades obtenidas de los modelos de histogramas.

En lugar de obtener G decodificaciones, a la señal de entrada se le aplican los modelos de histogramas correspondientes a los G grupos de hablantes, de donde se obtienen G valores de probabilidad. Esto implica tan sólo un cómputo sencillo, de coste lineal con la longitud de la señal. En el caso de utilizar MOM continuos, además es necesario etiquetar los vectores de parámetros, tarea relativamente poco costosa. Finalmente, se escoge como grupo óptimo de hablantes aquél que proporcione una mayor probabilidad según los modelos de histogramas, y se realiza una única pasada de reconocimiento utilizando los MOM correspondientes a dicho grupo. De forma abreviada, denominaremos a este método *Max-HProb*.

6.2.6.3. Selección de MOM basada en las probabilidades de un subconjunto de decodificaciones.

Esta aproximación es una mezcla de las dos anteriores. A la señal de entrada se le aplican los modelos de histogramas correspondientes a los G grupos de hablantes, de donde se obtienen G valores de probabilidad, pero en lugar de escoger únicamente el grupo de máxima probabilidad, se toman aquellos grupos cuya probabilidad supere un cierto umbral —típicamente un 70 % del valor máximo. A continuación se realizan tantas pasadas de reconocimiento como grupos hayan superado el umbral y se escoge como decodificación óptima aquélla que dé máxima probabilidad. En adelante, a este método se le denominará *Haz-HProb/Max-Prob*.

Se trata, en definitiva, de reducir el coste computacional de la primera aproximación, descartando aquellos grupos que, según los modelos de histogramas, sean poco probables. Esta estrategia viene motivada por el hecho de que en ocasiones, al aplicar el modelo de histogramas al grupo cuya decodificación da la máxima probabilidad, se obtiene una probabilidad cercana al máximo pero ligeramente inferior. El número de pasadas de reconocimiento dependerá del umbral establecido para descartar grupos poco probables. Típicamente, con alrededor de tres pasadas puede alcanzarse el mismo rendimiento que se obtiene con G pasadas.

6.2.6.4. Combinación de MOM basada en las probabilidades obtenidas de los modelos de histogramas.

La última estrategia —que de forma abreviada denominaremos *Comp-HProb*— consiste en componer un modelo acústico adaptado al hablante mediante una combinación lineal de los modelos de grupos de hablantes, en una línea similar a (Pusateri and Hazen, 2002). Los pesos de esa combinación se calculan de forma heurística a partir de las probabilidades de los modelos de histogramas. Dada una muestra cuantificada Y , el peso ω_g de un cierto grupo de hablantes g se define simplemente como la probabilidad normalizada de Y , dado el modelo de histogramas:

$$\omega_g = \frac{P(Y|g)}{\sum_{g'=1}^G P(Y|g')} \quad (6.25)$$

Si se utilizan MOM discretos, la probabilidad de emitir la etiqueta k en la representación acústica j de un estado q del modelo adaptado se calcula como sigue:

$$b_{qj}^{(adapt)}(k) = \sum_{g=1}^G \omega_g b_{qj}^{(g)}(k) \quad (6.26)$$

Como $\sum_{g=1}^G \omega_g = 1$ y $\forall g \sum_{k=1}^N b_{qj}^{(g)}(k) = 1$, se tiene también que $\sum_{k=1}^N b_{qj}^{(adapt)}(k) = 1$. En el caso de MOM discretos, esta estrategia de adaptación supone un incremento del coste sólo en lo que respecta a la creación de unos modelos distintos para cada señal de entrada, ya que el coste de reconocimiento no aumenta con respecto al de utilizar un conjunto de modelos genéricos.

Sin embargo, si se utilizan modelos continuos, el entrenamiento específico de modelos de grupo afecta a las medias y a los pesos de las gaussianas, por lo que si en cada mezcla se tenían M gaussianas, el modelo que resulte de la combinación de los MOM de grupo tendrá $G \cdot M$ gaussianas por mezcla. La probabilidad de emisión en la representación acústica j de un estado q del MOM continuo adaptado será:

$$b_{qj}^{(adapt)}(\mathbf{x}) = \sum_{m=1}^M \sum_{g=1}^G \omega_g c_{qjm}^{(g)} \mathcal{N}(\mathbf{x}; \mu_{qjm}^{(g)}, \sigma_{qjm}^{2(g)}) \quad (6.27)$$

Así pues, el MOM continuo adaptado tendrá G veces más parámetros que cada uno de los MOM continuos de grupo. De hecho, aunque se efectúe sólo una pasada de reconocimiento, será como si se utilizaran los G MOM de grupo simultáneamente, y el coste computacional de dicha pasada será G veces superior al de utilizar un conjunto de modelos genéricos.

6.2.7. Resultados del *clustering*.

El algoritmo de *clustering* ha sido desarrollado y probado frente a una base de datos de habla leída, concretamente SENGLAR16. Después ha sido aplicado sobre las dos bases de datos de habla espontánea que nos interesan: INFOTREN y CORLEC-EHU-1. En todos los casos, para validar un grupo se ha exigido un mínimo de 2 hablantes y 30000 tramos de señal (5 minutos). Por otra parte, se ha fijado un máximo de 20 iteraciones en la convergencia de los grupos tras cada división, aunque en la práctica raramente se pasa de las 10 iteraciones. Finalmente, el umbral de disminución relativa de la distorsión interna promedio entre dos iteraciones sucesivas se ha establecido en el 1 % ($\rho = 0,01$).

6.2.7.1. *Clustering* de hablantes en SENGLAR16.

En el corpus de entrenamiento de SENGLAR16 intervienen 57 hablantes, que provienen de tres entornos de grabación distintos, correspondientes a las tres sub-bases de datos que componen SENGLAR16: FRASES, EUROM1 y PROBA. Las sub-bases EUROM1 y PROBA fueron adquiridas en condiciones muy similares. No así la sub-base FRASES, que fue adquirida de forma independiente. Tras aplicar el algoritmo de *clustering*, los hablantes de SENGLAR16 se han repartido en 5 grupos. En la Tabla 6.1 se muestra dicho reparto y se indica el número de hombres y de mujeres, el número de intervenciones y el número de tramos —a razón de 10 milisegundos por tramo— asignados a cada grupo, así como la sub-base de datos de la que provienen.

Puede observarse que 15 de los 23 hombres de EUROM1 se asignan al grupo 2, mientras que 18 de las 20 mujeres de EUROM1 van al grupo 4. El reparto es aún más claro en el caso de PROBA: los hombres van al grupo 5 y las mujeres al grupo 4. De hecho, los grupos 2 y 5 son mayoritariamente masculinos, mientras que el grupo 4 está formado casi en su totalidad por mujeres. Finalmente, los hablantes de la sub-base FRASES han sido segregados en dos grupos específicos (1 y 3) de género mixto. Esto pone de manifiesto que el algoritmo de *clustering* distingue entre diferentes tipos de hablantes, pero también entre diferentes condiciones acústicas (ambientales y de canal).

Tabla 6.1: Reparto de hablantes al aplicar el algoritmo de *clustering* sobre SENGLAR16. Se indica el número de hombres y mujeres asignados a cada grupo, así como la sub-base de datos de la que provienen.

	FRASES		EUROM1		PROBA		Total		
	H	M	H	M	H	M	Hablantes	Intervenciones	Tramos
SENGLAR16	3	4	23	20	3	4	57	1529	484212
Grupo 1	1	2	0	0	0	0	3	251	55170
Grupo 2	0	0	15	2	0	0	17	343	118822
Grupo 3	2	2	1	0	0	0	5	287	63263
Grupo 4	0	0	3	18	0	4	25	508	195060
Grupo 5	0	0	4	0	3	0	7	140	51897

Tabla 6.2: Porcentaje de intervenciones asignadas correctamente (%IAC) mediante los modelos de histogramas, para el corpus de entrenamiento de SENGLAR16.

Grupo	1	2	3	4	5	Total
%IAC	98.01	97.67	90.24	93.31	93.57	94.51

Un aspecto fundamental de los modelos de histogramas es su coherencia interna. Una forma de verificar dicha coherencia es comprobar cuántas de las intervenciones de los hablantes del corpus de entrenamiento son asignadas al grupo *correcto*. En el caso de SENGLAR16, se ha comprobado que el 94.51 % de las intervenciones son correctamente asignadas por los modelos de histogramas, aunque los porcentajes varían de unos grupos a otros (véase la Tabla 6.2). Esto significa que los modelos tienen un alto grado de coherencia.

A título ilustrativo, en la Figura 6.3 se muestran las probabilidades normalizadas (ω_g) de las intervenciones de un hablante del grupo 1 de SENGLAR16 para los 5 modelos de histogramas. En este caso, todas las intervenciones son correctamente asignadas al grupo 1. Es destacable la cercanía del grupo 3 (la *otra mitad* de la sub-base FRASES) frente a la más o menos similar lejanía de los grupos 2, 4 y 5. Esto es debido a que los grupos 1 y 3 contienen señales adquiridas en idénticas condiciones, que sólo se diferencian en el tipo de hablantes, mientras que las señales de los grupos 2, 4 y 5 fueron adquiridas en condiciones ambientales y de canal muy distintas.

Finalmente, desde el punto de vista práctico tiene especial interés la posibilidad de determinar el grupo de hablantes más adecuado para una cierta señal a partir de tan sólo una pequeña fracción de la misma. Esto permitiría acometer una estrategia de adaptación prácticamente instantánea, ya que la asignación a uno u otro grupo apenas retrasaría el proceso de reconocimiento. Los datos de clasificación mostrados en la Tabla 6.2 se han obtenido aplicando los modelos de histogramas a las señales completas. En la Figura 6.4 se presenta el porcentaje de intervenciones correctamente asignadas en función de la longitud del segmento de señal, medida en tramos de de 10 milisegundos. Los porcentajes van aumentando desde un 77.70 % con tan sólo 0.2 segundos de señal hasta un 93.26 % con dos segundos de señal. A partir de ese punto el crecimiento se atenúa considerablemente, en parte porque muchas de las señales de SENGLAR16 apenas duran 3 segundos. Los mismos datos, desglosados por grupos, se presentan en la Tabla 6.3. En cualquier caso, es significativo que con apenas un segundo de señal se pueda clasificar correctamente más de un 90 % de las intervenciones.

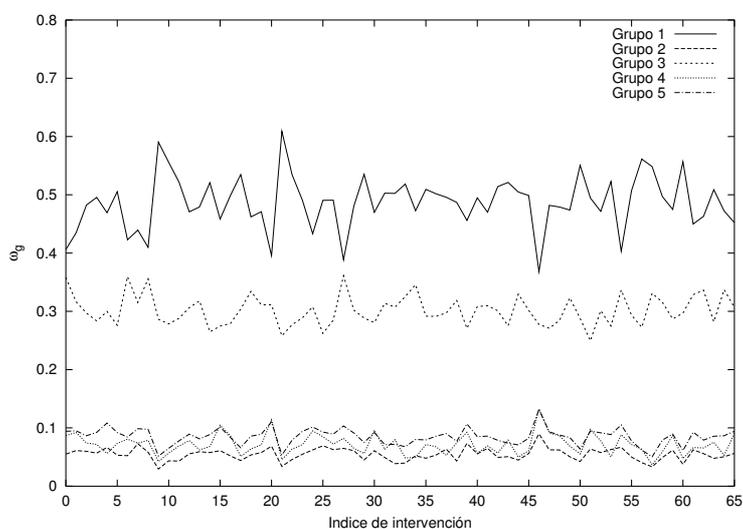


Figura 6.3: Probabilidades normalizadas (ω_g) de las intervenciones de un hablante del grupo 1 de SENGLAR16, obtenidas mediante los modelos de histogramas. El modelo del grupo 1 proporciona las probabilidades más altas en todos los casos.

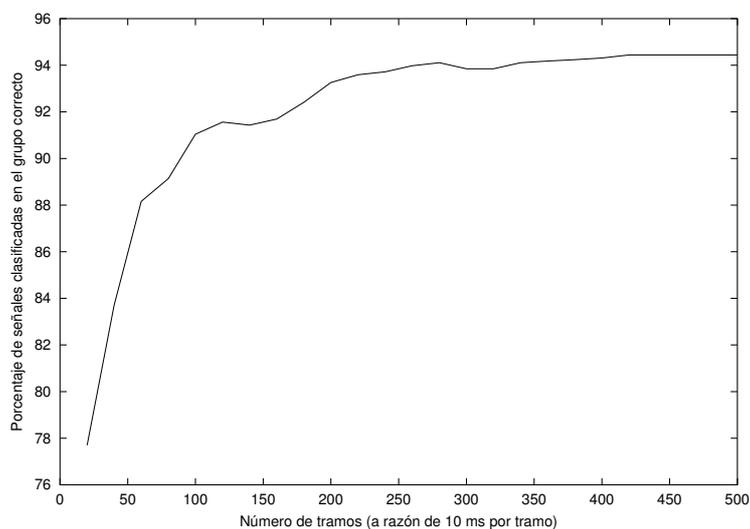


Figura 6.4: Porcentaje de intervenciones del corpus de entrenamiento de SENGLAR16 asignadas al grupo correcto, en función de la longitud del segmento de señal sobre el que se aplican los modelos de histogramas.

Tabla 6.3: Porcentaje de intervenciones asignadas al grupo correcto, desglosado por grupos, en función de la longitud del segmento de señal sobre el que se aplican los modelos de histogramas, para el corpus de entrenamiento de SENGLAR16.

Número de tramos	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Total
20	81.27	90.67	87.46	66.34	60.71	77.70
40	86.45	96.21	89.55	76.97	60.71	83.71
60	92.83	97.08	92.68	83.07	67.14	88.16
80	94.82	98.25	89.20	84.65	72.86	89.14
100	96.41	96.79	90.24	88.39	78.57	91.04
120	95.22	97.38	88.15	88.78	87.86	91.56
140	95.22	95.63	87.80	88.78	91.43	91.43
160	95.62	95.34	86.06	89.96	93.57	91.69
180	95.62	95.63	86.06	91.73	94.29	92.41
200	96.41	96.50	87.46	92.13	95.71	93.26
220	96.81	95.63	88.50	93.31	94.29	93.59
240	96.81	95.63	88.85	93.31	95.00	93.72
260	98.01	95.92	89.20	93.50	93.57	93.98
280	98.01	95.63	89.55	93.70	94.29	94.11
300	98.01	95.63	89.55	92.91	94.29	93.85
320	98.01	95.63	89.55	92.91	94.29	93.85
340	98.01	96.50	89.90	92.91	94.29	94.11
360	98.01	96.50	89.90	93.11	94.29	94.18
380	98.01	96.50	89.90	93.31	94.29	94.24
400	98.01	96.79	89.90	93.31	94.29	94.31
420	98.01	97.38	89.90	93.31	94.29	94.44
440	98.01	97.38	89.90	93.31	94.29	94.44
460	98.01	97.38	90.24	93.11	94.29	94.44
480	98.01	97.38	90.24	93.11	94.29	94.44
500	98.01	97.38	90.24	93.11	94.29	94.44

Tabla 6.4: Reparto de hablantes al aplicar el algoritmo de *clustering* sobre INFOTREN.

	Hombres	Mujeres	Turnos	Tramos
Grupo 1	6	0	96	62563
Grupo 2	0	6	187	121887
Grupo 3	1	5	142	67261
Grupo 4	6	6	226	132327
Grupo 5	14	0	317	155830
Grupo 6	6	0	150	68051
Grupo 7	4	0	75	35720
Grupo 8	0	9	156	76208

Tabla 6.5: Porcentaje de intervenciones asignadas correctamente (%IAC) mediante los modelos de histogramas, para el corpus de entrenamiento de INFOTREN.

Grupo	1	2	3	4	5	6	7	8	Total
%IAC	85.42	78.61	84.51	62.83	76.34	67.33	69.33	83.33	75.32

6.2.7.2. *Clustering* de hablantes en INFOTREN.

El corpus de entrenamiento de INFOTREN consta de 1349 intervenciones, a cargo de 63 hablantes, que tras aplicar el algoritmo de *clustering*, se reparten en 8 grupos. En la Tabla 6.4 se muestra dicho reparto y se indica el número de hombres y de mujeres, así como el número de intervenciones y el número de tramos, asignados a cada grupo. De nuevo se observa una separación entre grupos formados sólo por hombres y grupos formados sólo por mujeres. Hay dos grupos mixtos: el grupo 3, que casi no lo es, ya que consta de 5 mujeres y tan sólo un hombre, y el grupo 4, formado por 6 hombres y 6 mujeres. Este último podría interpretarse como el *grupo frontera*, integrado por los hablantes de cada género más próximos —desde el punto de vista acústico— al género opuesto.

Se observa también, con respecto a SENGLAR16, una disminución importante de la coherencia interna de estos grupos, posiblemente por el efecto conjunto de varios factores: la pérdida de información acústica intrínseca al uso de una frecuencia de muestreo menor (8 kHz), la mayor variabilidad acústica del habla espontánea, condiciones ambientales más ruidosas, presencia de pausas y sonidos extralingüísticos —que podrían *homogeneizar* los modelos de histogramas—, etc. Como consecuencia, sólo un 75.32 % de las intervenciones del corpus de entrenamiento de INFOTREN son asignadas al grupo correcto mediante los modelos de histogramas. En la Tabla 6.5 se muestra el porcentaje de asignaciones correctas para los 8 grupos definidos. El porcentaje más bajo se obtiene para el grupo 4, que es el *grupo frontera*. Por lo demás, no parece que ni el género ni el número de hablantes tengan influencia directa en dicho porcentaje. Por ejemplo, los grupos 1 y 6 son ambos grupos de hombres, formados por 6 hablantes, y sin embargo los porcentajes de acierto son muy distintos: 85.42 % y 67.33 %, respectivamente.

La indefinición acústica de los modelos de histogramas en INFOTREN hace que, aunque se produzca una distinción clara de los grupos poco probables, siempre existan dos o tres grupos asimilables a cada señal, normalmente grupos del mismo género que el hablante. Tal es el caso, por ejemplo, del hablante 33 (hombre), perteneciente al grupo 6, cuyas señales se clasifican

bien dentro del grupo 6, bien dentro del grupo 5, pero nunca dentro de otro grupo, como queda claro en la Figura 6.5. Este comportamiento de los modelos de histogramas sugiere el uso de una estrategia de adaptación basada no en una decisión *dura* sobre el grupo más probable, sino en una pre-selección de grupos más probables, como la descrita en el apartado 6.2.6.3.

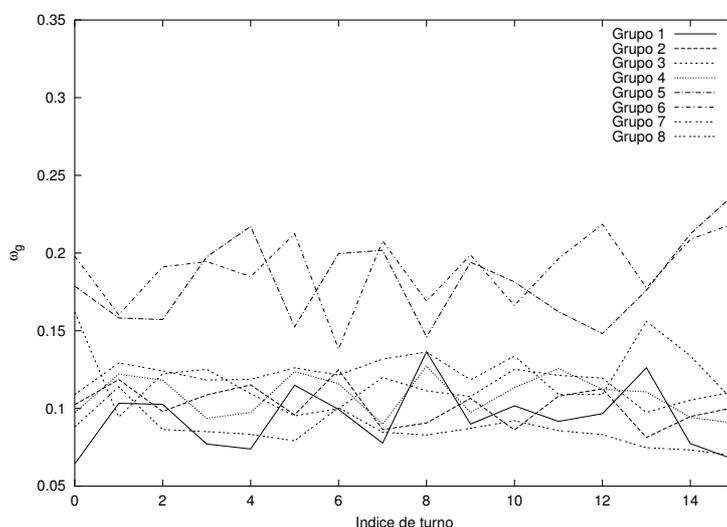


Figura 6.5: Probabilidades normalizadas (ω_g) de las intervenciones de un hablante del grupo 6 de INFOTREN, obtenidas mediante los modelos de histogramas. Los modelos de los grupos 5 y 6 proporcionan las probabilidades más altas en todos los casos.

Finalmente, tal como muestra la Figura 6.6, la utilización de una fracción pequeña de la señal para determinar el grupo de hablantes más adecuado no produce, en el caso de INFOTREN, resultados tan buenos como los obtenidos para SENGLAR16. Si se utilizan fracciones de 0.2 segundos, tan sólo el 44.63 % de las señales son asignadas al grupo correcto. Con fracciones de un segundo el porcentaje asciende al 66.94 % y con fracciones de dos segundos, al 72.05 %, valor ya bastante próximo al 75.32 % que se alcanza con las señales completas. Este bajo rendimiento puede atribuirse en parte a la pérdida de información acústica intrínseca a una frecuencia de muestreo demasiado baja, y en parte a la presencia de pausas y ruidos, particularmente al inicio de las señales, que impiden caracterizar adecuadamente la voz del hablante. Este último efecto podría evitarse identificando y descartando segmentos extralingüísticos, tanto en la construcción como en la aplicación de los modelos de histogramas, lo cual complicaría excesivamente el procedimiento de adaptación al hablante, que pretendemos sea, sobre todo, simple.

6.2.7.3. Clustering de hablantes en CORLEC-EHU-1.

En CORLEC-EHU-1 hay hablantes que aportan tan sólo unos pocos segundos de señal. Se trata de locutores de apoyo o participantes muy secundarios en las entrevistas que hacen una sola pregunta o comentario. Esto no sucede en SENGLAR16 porque cada hablante, por diseño, aporta un número mínimo de intervenciones. En el caso de INFOTREN, cada hablante mantiene tres diálogos distintos con el sistema de información ferroviaria. El número medio de intervenciones del usuario por diálogo es de 7.3 y la duración promedio de cada intervención es de 5.43 segundos, de ahí que como media se tengan casi dos minutos de señal de cada hablante.

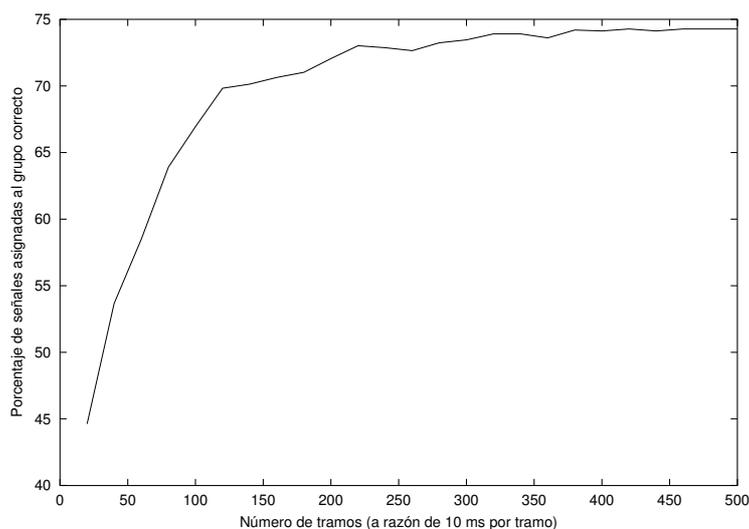


Figura 6.6: Porcentaje de intervenciones del corpus de entrenamiento de INFOTREN asignadas al grupo correcto, en función de la longitud del segmento de señal sobre el que se aplican los modelos de histogramas.

El método de suavizado propuesto mediante las fórmulas 6.6 y 6.7 es muy rudimentario, y el modelo que resulta cuando se dispone de pocas muestras es tan específico que la disimilitud con cualquier otro modelo —según las definiciones 6.18 y 6.20— se hace muy grande. Esto afecta muy negativamente al algoritmo de *clustering*, ya que continuamente aparecen hablantes muy alejados de los grupos principales, que forman un grupo por sí mismos —con muy pocas muestras—, lo cual rompe antes de tiempo el proceso de partición y reagrupamiento. Para evitar este efecto, se han descartado 8 de los 105 hablantes de CORLEC-EHU-1, en concreto aquéllos cuyo tiempo de señal no llega a 30 segundos. Así, el bloque C1 pasa de 37 a 35 hablantes, el bloque C2 pasa de 35 a 30 y el bloque C3, de 33 a 32.

Se ha aplicado el algoritmo de *clustering* a las tres particiones de CORLEC-EHU-1. Como resultado, los hablantes de C2+C3 se han organizado en 6 grupos; los de C1+C3, en 13 grupos; y los de C1+C2, en 5 grupos. En la Tabla 6.6 se muestra el número de hombres y mujeres, el número de hablantes grabados de la radio, el número de hablantes grabados de la televisión, así como el número de intervenciones y el número de tramos asignados a cada grupo en los tres casos.

El número de hombres en CORLEC-EHU-1 es más de dos veces el número de mujeres. En concreto, la relación es de 41/21 en la primera partición, de 49/18 en la segunda y de 46/19 en la tercera. Esto incide por un lado en que existan grupos *puros* de hombres pero no de mujeres, que por otro lado aparecen muy dispersas y siempre mezcladas con hombres. Esto quizá incida en que las voces femeninas no se estén modelando de forma adecuada. En la partición 1 los hablantes están casi perfectamente distribuidos por géneros: los grupos 1 y 6 reúnen a la mayor parte de las mujeres, mientras que los grupos 2, 3, 4 y 5 son mayoritariamente masculinos. En la partición 3 los grupos 3 y 4 son mixtos, mientras que los grupos 1, 2 y 5 están formados casi exclusivamente por hombres. La partición 2 aparece muy disgregada, con muchos grupos muy pequeños y las mujeres muy dispersas por todos ellos. Esto influirá negativamente en la robustez de los modelos acústicos, ya que tendrán que estimarse a partir de conjuntos de muestras muy pequeños, que en muchos casos no serán lo bastante homogéneos.

Tabla 6.6: Reparto de hablantes al aplicar el algoritmo de *clustering* sobre las tres particiones de CORLEC-EHU-1.

Partición 1: C2+C3/C1						
	Hombres	Mujeres	Radio	Televisión	Turnos	Tramos
Grupo 1	6	11	11	6	330	315632
Grupo 2	3	0	3	0	64	53622
Grupo 3	6	0	6	0	152	195285
Grupo 4	17	1	15	3	371	419805
Grupo 5	8	1	8	1	210	238007
Grupo 6	1	8	2	7	253	141831

Partición 2: C1+C3/C2						
	Hombres	Mujeres	Radio	Televisión	Turnos	Tramos
Grupo 1	3	4	1	6	181	117334
Grupo 2	12	0	12	0	282	381865
Grupo 3	3	0	3	0	69	104551
Grupo 4	2	0	2	0	30	37174
Grupo 5	5	1	0	6	63	109629
Grupo 6	2	4	0	6	167	79779
Grupo 7	6	2	8	0	172	150192
Grupo 8	3	0	3	0	80	43546
Grupo 9	3	0	3	0	99	107150
Grupo 10	3	1	0	4	46	35907
Grupo 11	2	6	7	1	129	114131
Grupo 12	2	0	2	0	21	36979
Grupo 13	3	0	3	0	82	46805

Partición 3: C1+C2/C3						
	Hombres	Mujeres	Radio	Televisión	Turnos	Tramos
Grupo 1	13	2	13	2	319	412593
Grupo 2	4	0	4	0	86	56222
Grupo 3	11	9	4	16	440	386142
Grupo 4	10	8	6	12	355	297694
Grupo 5	8	0	8	0	195	217551

Tabla 6.7: Porcentaje de intervenciones asignadas correctamente (%IAC) mediante los modelos de histogramas, para los conjuntos de entrenamiento de las tres particiones de CORLEC-EHU-1.

Grupo	%IAC		
	C2+C3	C1+C3	C1+C2
1	83.94	75.14	84.01
2	84.38	85.46	86.05
3	86.18	78.26	89.55
4	86.52	100.00	97.46
5	80.48	76.19	91.79
6	90.51	88.02	–
7	–	79.07	–
8	–	87.50	–
9	–	77.78	–
10	–	97.83	–
11	–	86.05	–
12	–	95.24	–
13	–	86.59	–
Total	85.58	83.46	90.39

El algoritmo de *clustering* no maneja las señales de forma independiente, sino que las agrupa o separa por bloques, cada uno de ellos correspondiente a un hablante. Esto favorece que sean las características propias de los hablantes las que determinen la definición de los grupos. Sin embargo, en el caso de CORLEC-EHU-1 las condiciones de grabación —ruido ambiental y de canal— son muy variables, por lo que tienen un peso importante en la definición de los grupos. Es decir, en muchas ocasiones dos hablantes forman parte del mismo grupo no porque tengan voces similares, sino porque las condiciones ambientales o de canal son muy parecidas. En la Tabla 6.6 se muestra el número de hablantes grabados directamente de la radio o con la grabadora puesta delante del televisor. Puede comprobarse que muchos de los grupos están formados únicamente por hablantes de radio o por hablantes de televisión. La homogeneidad de los grupos a este respecto es incluso mayor que la observada con respecto al género. Curiosamente, además, cuando se mezclan hablantes de distintas condiciones, en no pocas ocasiones la distribución radio/televisión coincide con la distribución hombre/mujer. Desde el punto de vista de la modelización acústica, el que los grupos se definan por el género o por las condiciones de grabación no es un problema, ya que la adaptación —independiente para cada señal— se puede plantear perfectamente como la búsqueda del conjunto de modelos acústicos que encaje mejor con las condiciones de canal, con el ruido ambiental y con las características propias del hablante, todo ello simultáneamente.

En cuanto a la coherencia interna de los grupos, el 85.58 % de las intervenciones del corpus de entrenamiento de la primera partición son asignadas al grupo correcto; en la partición 2, el porcentaje de intervenciones asignadas correctamente es del 83.46 %; y en la partición 3 se alcanza el 90.39 %. En la Tabla 6.7 se muestran los porcentajes desglosados por grupos. Curiosamente, a pesar de que las condiciones de grabación de CORLEC-EHU-1 son bastante peores que las de INFOTREN, la coherencia interna de los grupos es 10 puntos superior. En la Figura 6.7 se muestran los resultados obtenidos para un hablante del grupo 6 de la partición 1. Puede observarse que el 100 % de las señales son clasificadas correctamente, y que además

la diferencia entre las probabilidades del grupo 6 y las del resto de grupos es muy clara. La explicación hay que buscarla en la frecuencia de muestreo: 8 kHz para INFOTREN, 16 kHz para CORLEC-EHU-1 y SENGLAR16. Una frecuencia de muestreo más baja implica una pérdida de información acústica que suaviza las diferencias entre los modelos de histogramas, y de ahí que los errores de clasificación sean más numerosos en INFOTREN. Por otra parte, CORLEC-EHU-1 contiene habla espontánea adquirida en condiciones adversas y variables de ruido ambiental y de canal, por lo que la variabilidad interna de los grupos ha de ser forzosamente mayor que la observada en SENGLAR16, y menor, por tanto, su coherencia acústica.

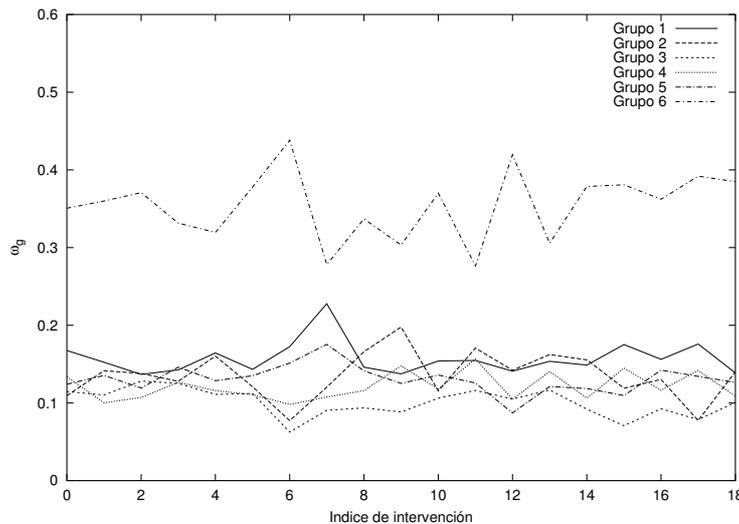


Figura 6.7: Probabilidades normalizadas (ω_g) de las intervenciones de un hablante del grupo 6 de la primera partición de CORLEC-EHU-1, obtenidas mediante los modelos de histogramas. El modelo del grupo 6 proporciona las probabilidades más altas en todos los casos.

Tal como muestra la Figura 6.8, si se utiliza una fracción pequeña de la señal para determinar el grupo de hablantes, el porcentaje de intervenciones asignadas al grupo correcto evoluciona de manera similar para las tres particiones. Con aproximadamente tres segundos de señal se alcanzan porcentajes muy cercanos a los que se obtienen con la señal completa, todos ellos por encima del 80 %: 83.26 % para la partición 1, 80.51 % para la partición 2 y 88.03 % para la partición 3. Teniendo en cuenta que muchas de las señales de esta base de datos sobrepasan los 30 segundos de duración, poder clasificarlas utilizando únicamente los 3 segundos iniciales supone una mejora importante del proceso de adaptación —aunque éste no sea perfecto.

6.2.7.4. Evolución de la distancia promedio de un hablante a su grupo.

Aunque la distancia promedio de un hablante a su grupo, $R(n)$, desciende con cada iteración del algoritmo de *clustering*, lo hace en distinta medida para cada base de datos, y además parte de valores iniciales distintos. En la Figura 6.9 se muestra la evolución de $R(n)$ para las tres bases de datos consideradas en este capítulo.

No puede decirse que $R(n)$ alcance la convergencia en ningún caso. De hecho, el algoritmo de *clustering* termina siempre debido a la imposibilidad de generar particiones en las que todos los grupos tengan suficientes muestras y hablantes. En el caso de SENGLAR16, el valor final $R(5) = 0,5445$ es un 61 % del valor inicial $R(1) = 0,8937$. En el caso de CORLEC-EHU-

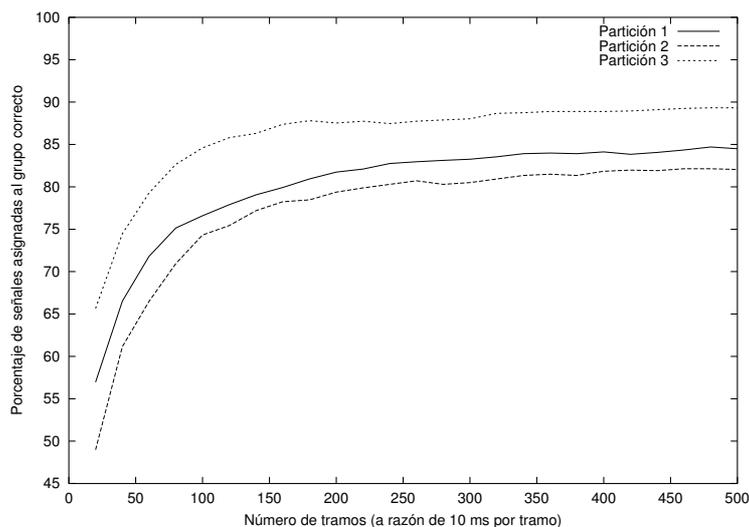


Figura 6.8: Porcentaje de intervenciones asignadas al grupo correcto en los conjuntos de entrenamiento correspondientes a las tres particiones de CORLEC-EHU-1, en función de la longitud del segmento de señal sobre el que se aplican los modelos de histogramas.

1, los valores finales de $R(n)$ son $R(6) = 0,5114$ para la partición 1, $R(13) = 0,4476$ para la partición 2, y $R(5) = 0,6187$ para la partición 3, que representan un 63 %, un 48 % y un 64 % de los valores iniciales, respectivamente. Curiosamente, el porcentaje de señales asignadas al grupo correcto parece ser proporcional al valor final de $R(n)$, cuando lo lógico sería una relación inversa. La evolución de $R(n)$ en INFOTREN se distingue porque el valor de partida ($R(1) = 0,5913$) es significativamente inferior a los de SENGLAR16 y CORLEC-EHU-1. De ahí que el valor final sea también el más bajo de todos: $R(8) = 0,3484$, un 59 % de $R(1)$. Nótese que el porcentaje de señales asignadas al grupo correcto en INFOTREN es también el más pequeño de todos. Si los valores de la Figura 6.9 se dividen por el valor inicial $R(1)$, se obtiene la evolución normalizada (Figura 6.10), que es muy parecida en todos los casos.

6.3. Modelos de género.

Se plantea en este apartado una aproximación alternativa al *clustering* de hablantes, que, de hecho, puede verse como una simplificación de la misma idea, ya que consiste en la definición *a priori* de dos grupos de hablantes, hombres y mujeres, y la utilización de dos conjuntos de modelos acústicos distintos durante el reconocimiento. Esta separación se justifica porque las características fisiológicas de hombres y mujeres determinan dos tipos de voces distintas.

Los resultados de DAF con MOM continuos de 32 gaussianas presentados en el apartado 4.3 se han desglosado por hablantes. En el caso de SENGLAR16, la tasa promedio es del 72.06 % para los hombres y del 73.25 % para las mujeres; en el caso de INFOTREN, las tasas promedio son del 59.81 % para los hombres y del 61.63 % para las mujeres; y finalmente, en el caso de CORLEC-EHU-1, al hacer un balance de las tasas de DAF sobre las tres particiones, se obtiene un promedio del 52.25 % para los hombres, frente a un 48.26 % para las mujeres. Según se deduce de estos resultados, cuando en el corpus de entrenamiento aparecen hombres y mujeres por igual, como en el caso de SENGLAR16 e INFOTREN, los modelos resultantes se comportan mejor con éstas últimas. Sin embargo, si el número de hombres es notablemente mayor

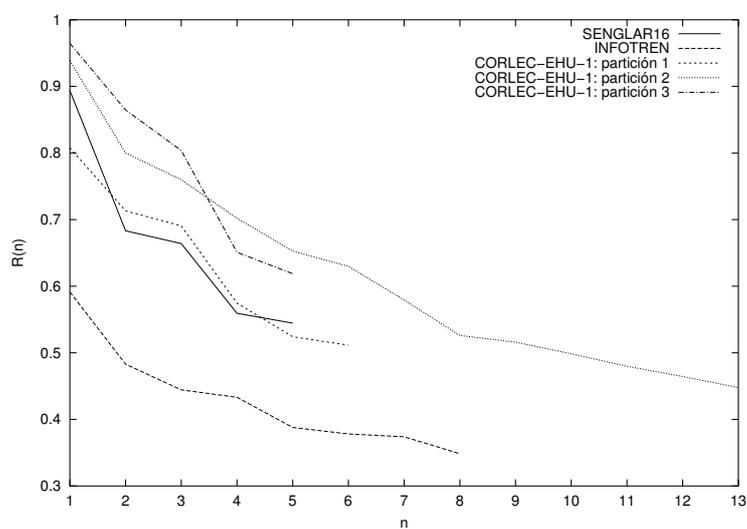


Figura 6.9: Evolución de la distancia promedio de un hablante a su grupo $R(n)$ para las tres bases de datos consideradas en este capítulo.

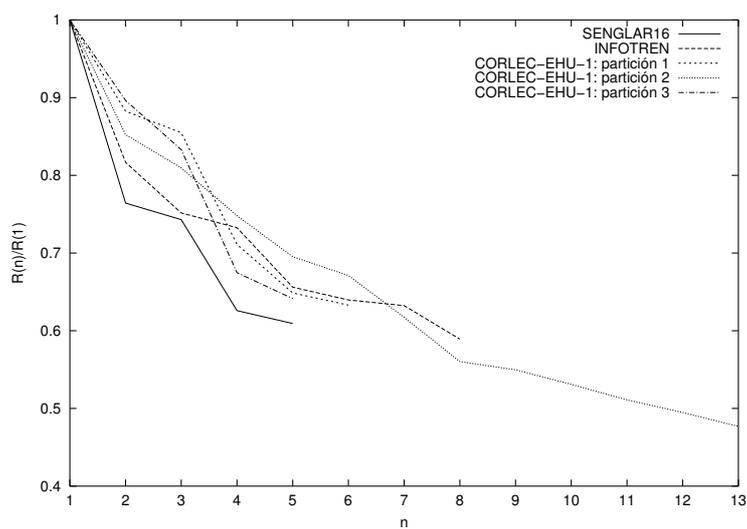


Figura 6.10: Evolución de la distancia promedio de un hablante a su grupo $R(n)$, normalizada por el valor inicial $R(1)$, para las tres bases de datos consideradas en este capítulo.

Tabla 6.8: Porcentaje de intervenciones cuyo género se ha determinado correctamente mediante los modelos de histogramas.

	Hombres	Mujeres	Total
SENGLAR16	96.76	97.75	97.25
INFOTREN	95.47	95.50	95.48
CORLEC-EHU-1: partición 1	93.15	87.44	91.51
CORLEC-EHU-1: partición 2	79.78	94.08	83.32
CORLEC-EHU-1: partición 3	86.01	89.55	87.18

—en CORLEC-EHU-1 hay más de dos hombres por cada mujer, y cada uno de ellos aporta en promedio más tramos de señal que cada una de ellas—, los modelos resultantes se ajustan, como es lógico, al grupo de hablantes dominante, y son los hombres los que presentan tasas más altas. Así pues, en los tres casos se observan diferencias significativas en el rendimiento de los modelos acústicos para hombres y mujeres, y queda margen, por tanto, para que la estimación de modelos específicos permita alcanzar un mejor ajuste de los parámetros a las muestras.

Los modelos de género se estiman utilizando exactamente el mismo procedimiento descrito en el apartado 4.2.1, dividiendo el conjunto de entrenamiento en dos subconjuntos, uno con las muestras de hombres y otro con las muestras de mujeres. Implícitamente se está suponiendo que ambos subconjuntos contienen muestras suficientes como para estimar los MOM respectivos de manera robusta. En el momento de reconocer una señal, hay que optar por hacerlo con los MOM de hombres o con los MOM de mujeres. El método habitual consiste en reconocer la señal con los dos conjuntos de modelos y dar por buena la hipótesis que tenga mayor probabilidad. Sin embargo, ello requiere reconocer dos veces la misma señal, por lo que el coste de reconocimiento se duplica.

Para evitarlo, se han construido sendos modelos de histogramas para hombres y mujeres, con objeto de aplicarlos a la señal —o a una fracción de la misma— para escoger el conjunto de modelos más adecuado. En la Tabla 6.8 se muestra el porcentaje de señales del corpus de entrenamiento correctamente clasificadas como masculinas o femeninas mediante los modelos de histogramas, valor al que en adelante nos referiremos simplemente como *coherencia de los modelos*.

Para SENGLAR16 e INFOTREN se obtienen valores superiores al 95 %, lo cual pone de manifiesto por un lado que los modelos de histogramas están muy definidos, y por otro, que la coherencia no depende significativamente de la frecuencia de muestreo. Por el contrario, los porcentajes obtenidos sobre las tres particiones de CORLEC-EHU-1 están comprendidos entre el 80 % y el 94 % y como promedio se sitúan alrededor de un 87 %. Esto significa que las condiciones de grabación de CORLEC-EHU-1, adversas y variables, afectan negativamente a la *definición* de los modelos de histogramas. Considerense los dos casos extremos: las mujeres de SENGLAR16, con una coherencia del 97.75 %, y los hombres de la partición 2 de CORLEC-EHU-1, con una coherencia del 79.78 %. En el primer caso las probabilidades normalizadas promedio son $\bar{\omega}_h = 0,3452$ y $\bar{\omega}_m = 0,6548$. En el segundo caso los dos géneros están peor definidos —menos separados— y esto se refleja en el valor promedio de las probabilidades normalizadas: $\bar{\omega}_h = 0,5697$ y $\bar{\omega}_m = 0,4303$.

6.4. Resultados de DAF.

6.4.1. Ajuste de τ en la estimación MAP.

En primer lugar se ha obtenido experimentalmente el valor óptimo de τ en la estimación MAP de los MOM de grupos de hablantes. En el caso de SENGLAR16, se han realizado barridos diferenciados para MOM discretos y continuos, partiendo de los modelos generales óptimos de unidades incontextuales y aplicando 6 iteraciones del algoritmo de Baum-Welch general para MOM discretos, y 3 iteraciones del algoritmo de entrenamiento por Viterbi para MOM continuos. En estos experimentos se ha utilizado la probabilidad de la hipótesis óptima en DAF como criterio de selección del mejor grupo de hablantes.

En la Figura 6.11 se muestran los resultados con MOM discretos y continuos de 32 gaussianas para SENGLAR16. La estimación MAP introduce una mejora apreciable en el rendimiento de los modelos con respecto a la estimación por máxima similitud ($\tau = 0$). Sin embargo, los resultados de DAF no son muy sensibles al valor de τ : los valores óptimos parecen ser $\tau = 30$ para MOM discretos y $\tau = 5$ para MOM continuos. Las tasas de DAF obtenidas, 66.51 % con MOM discretos y 74.41 % con MOM continuos, son sólo ligeramente mayores que las obtenidas sin *adaptación al hablante*, con incrementos de uno y dos puntos absolutos, respectivamente.

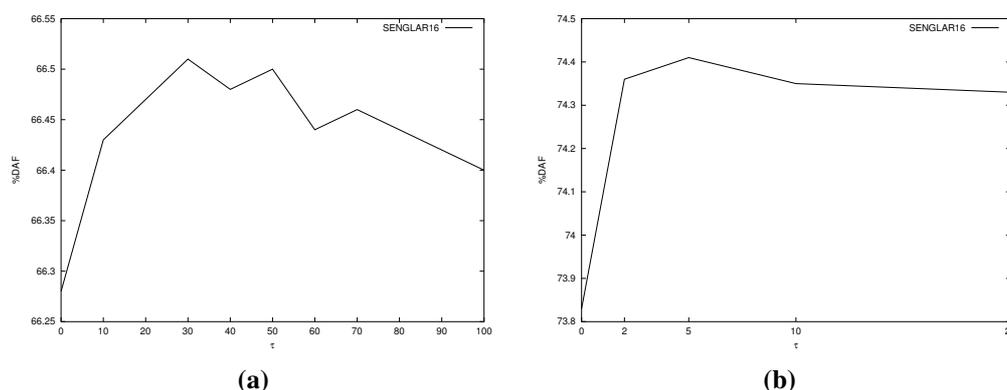


Figura 6.11: Tasas de DAF obtenidas sobre el corpus de test de SENGLAR16, utilizando modelos de grupos de hablantes con el criterio MAP, para distintos valores del factor de adaptación τ : (a) MOM discretos, (b) MOM continuos de 32 gaussianas.

En la Figura 6.12 se muestran los resultados de DAF utilizando MOM continuos de 32 gaussianas del conjunto de unidades ampliado (23 fonemas + 13 o 14 fenómenos de habla espontánea), para INFOTREN y las tres particiones de CORLEC-EHU-1. En ambos casos se observan diferencias mínimas al aplicar distintos valores de τ . En el caso de INFOTREN, el valor óptimo es $\tau = 2$, que proporciona una tasa de DAF dos puntos absolutos más alta que la obtenida sin adaptación: 63.61 % frente a 61.34 %. En CORLEC-EHU-1 los resultados varían de una partición a otra. Para la partición 1 el valor óptimo es $\tau = 2$, que proporciona una tasa un punto absoluto por encima de la tasa obtenida sin adaptación: 51.61 % frente a 50.63 %. Para las particiones 2 y 3 el valor óptimo es $\tau = 20$, y las tasas de DAF son 53.90 % para la partición 2 (frente a un 52.42 % sin adaptación), y 53.60 % para la partición 3 (frente a un 53.13 % sin adaptación).

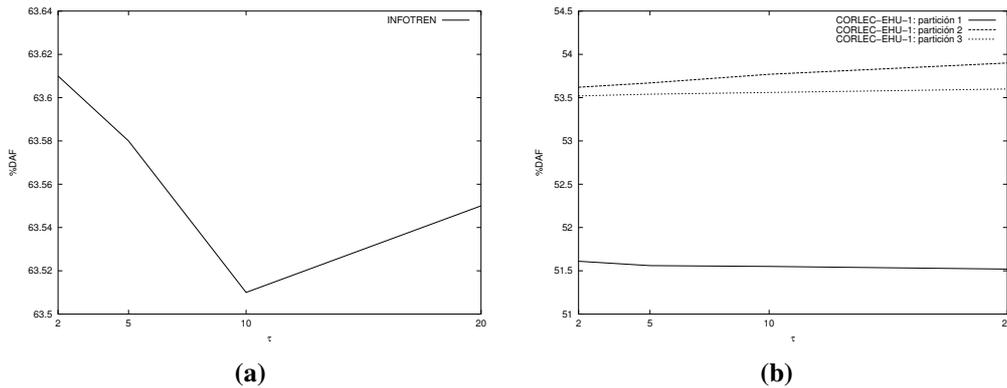


Figura 6.12: Tasas de DAF obtenidas sobre los corpus de test de INFOTREN (a) y las tres particiones de CORLEC-EHU-1 (b), utilizando modelos de grupos de hablantes (MOM continuos de 32 gaussianas) con el criterio MAP, para distintos valores del factor de adaptación τ .

Tabla 6.9: Tasas de DAF obtenidas mediante distintas alternativas de adaptación al hablante. Entre paréntesis se indica el coste computacional, expresado como el número promedio de pasadas de reconocimiento que es necesario realizar para cada señal. También se muestran las tasas de referencia, sin adaptación al hablante.

	Referencia	Max-Prob	Max-HProb	Haz-HProb (70 % Max-HProb) Max-Prob	Comp-HProb
SENGLAR16: MOMd	65.28	66.51 (5)	66.33 (1)	66.51 (2.08)	66.44 (1)
SENGLAR16: MOMc32g	72.38	74.41 (5)	73.04 (1)	74.41 (2.08)	73.68 (5)
INFOTREN: MOMc32g	61.34	63.61 (8)	63.13 (1)	63.60 (3.37)	62.14 (8)
CORLEC-EHU-1: p1, MOMc32g	50.63	51.61 (6)	51.33 (1)	51.60 (2.63)	51.20 (6)
CORLEC-EHU-1: p2, MOMc32g	52.42	53.90 (13)	53.18 (1)	53.89 (3.42)	53.35 (13)
CORLEC-EHU-1: p3, MOMc32g	53.13	53.60 (5)	53.44 (1)	53.58 (2.13)	53.64 (5)

6.4.2. Alternativas de adaptación.

Acaba de mostrarse que el uso de modelos específicos de grupos de hablantes introduce tan sólo pequeños incrementos en la tasa de DAF. Además, dichos resultados se han obtenido utilizando la probabilidad de las decodificaciones como criterio de selección del mejor grupo de hablantes —la alternativa de adaptación que hemos llamado *Max-Prob*—, lo cual multiplica por G (G : número de grupos) el coste de reconocimiento. Para mantener en lo posible esas mejoras sin incrementar el coste computacional, se han explorado las alternativas de adaptación descritas en el apartado 6.2.6, aplicando los valores óptimos de τ obtenidos previamente. Los resultados se presentan en la Tabla 6.9.

La alternativa *Max-Prob* proporciona las tasas más altas en todos los casos. Por otra parte, la alternativa *Max-HProb* proporciona tasas no tan altas pero con un coste computacional muy inferior, igual al de utilizar un único conjunto de modelos generales. La alternativa *Haz-HProb/Max-Prob* se revela como un compromiso entre las dos anteriores, ya que proporciona los mismos resultados que *Max-Prob* con un coste entre 2 y 3 veces el de *Max-HProb*. Finalmente, la composición de un modelo *ad-hoc* para cada señal, basado en los modelos de grupo y en los pesos obtenidos de las probabilidades de los modelos de histogramas, sólo parece planteable con MOM discretos, ya que proporciona tasas próximas a las de *Max-Prob* con tan sólo

una pasada de reconocimiento. Con MOM continuos *Comp-HProb* tiene el mismo coste computacional que *Max-Prob* y las tasas que proporciona son generalmente inferiores. Así pues, en la práctica, dependiendo de las restricciones de tiempo real, la adaptación al hablante deberá realizarse siguiendo una de las alternativas *Max-HProb* o *Haz-HProb/Max-Prob*.

Para completar el análisis de los resultados, en las Figuras 6.13, 6.14 y 6.15 se muestran las probabilidades normalizadas, obtenidas con los modelos de histogramas, de las intervenciones de tres hablantes de los conjuntos de test de SENGLAR16, INFOTREN y la partición 3 de CORLEC-EHU-1, respectivamente.

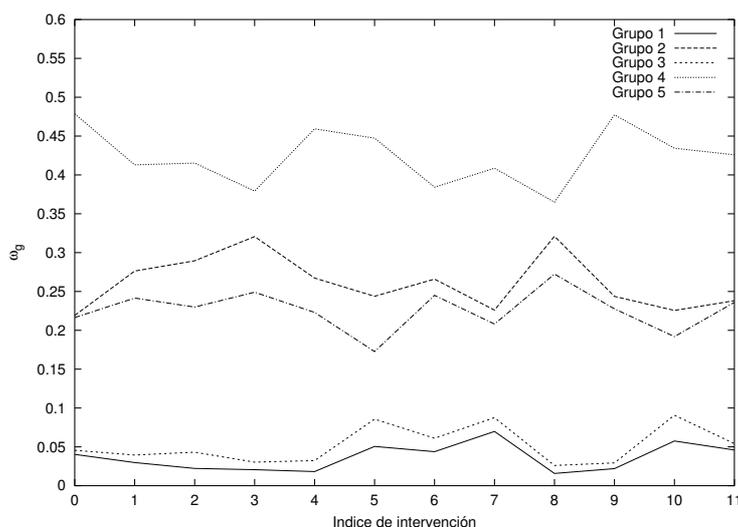


Figura 6.13: Probabilidades normalizadas (ω_g) de las intervenciones de una hablante del corpus de test de SENGLAR16, obtenidas mediante los modelos de histogramas. El modelo del grupo 4 proporciona las probabilidades más altas en todos los casos.

Todas las intervenciones del hablante de SENGLAR16 —una mujer perteneciente a la sub-base EUROM1— son clasificadas de forma similar. El modelo del grupo 4 —estimado, recuérdese, a partir de las intervenciones de 22 mujeres y 3 hombres, todos ellos pertenecientes a las sub-bases EUROM1 y PROBA— proporciona las probabilidades más altas. A continuación aparecen los modelos de los grupos 2 y 5, formados mayoritariamente por hombres, todos ellos pertenecientes también a las sub-bases EUROM1 y PROBA. En último lugar, con unas probabilidades muy bajas, se sitúan los grupos 1 y 3, es decir, los dos grupos que reúnen a los hablantes de la sub-base FRASES.

El hablante de INFOTREN es también una mujer, y son precisamente dos grupos formados exclusivamente por mujeres, el grupo 2 y el grupo 8, los que proporcionan las probabilidades más altas para casi todas las intervenciones. El resto de grupos aparece por debajo, en algunos casos muy claramente: grupos 1, 5, 6 y 7, todos ellos mayoritariamente masculinos, y en otros casos no tanto: grupo 2, que es casi puramente femenino, y grupo 4, que es mixto. El hablante de CORLEC-EHU-1 muestra también un comportamiento coherente. Se trata de un hombre y es el grupo 1, formado casi exclusivamente por hombres, el que proporciona las probabilidades más altas para casi todas las intervenciones. Estos resultados se reproducen para el resto de hablantes que forman parte de los conjuntos de test, lo cual confirma la validez de los modelos de histogramas en la tarea de identificar el grupo más próximo acústicamente a la intervención de un hablante.

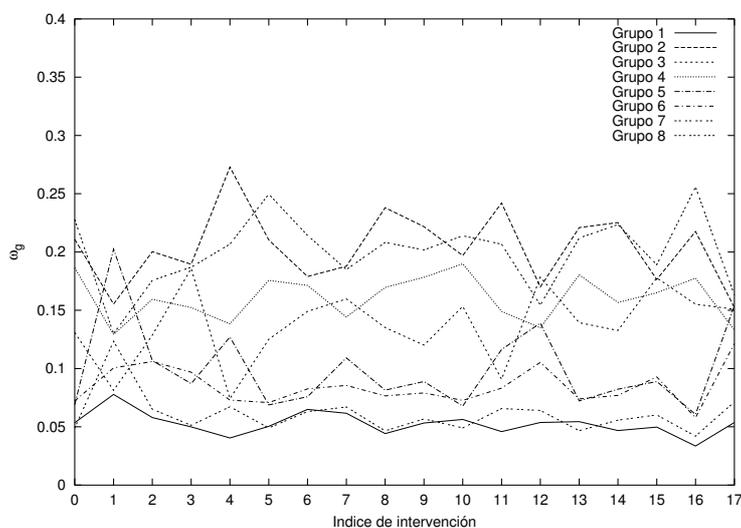


Figura 6.14: Probabilidades normalizadas (ω_g) de las intervenciones de una hablante del corpus de test de INFOTREN, obtenidas mediante los modelos de histogramas. Los modelos de los grupos 2 y 8 proporcionan las probabilidades más altas en la mayor parte de los casos.

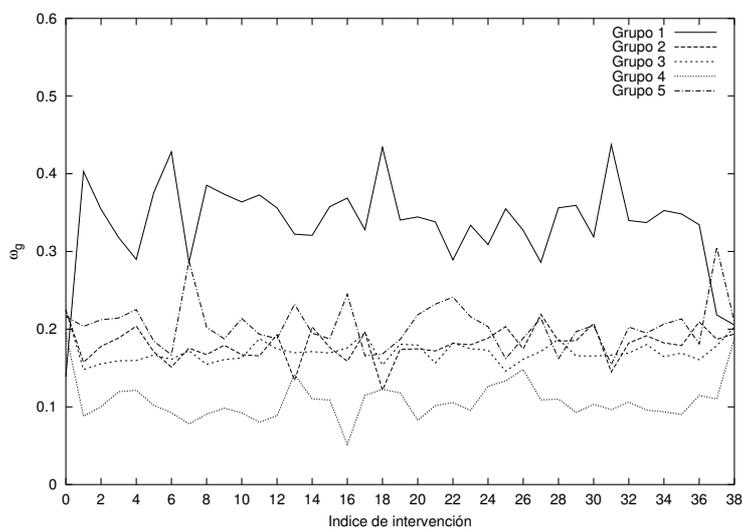


Figura 6.15: Probabilidades normalizadas (ω_g) de las intervenciones de un hablante del corpus de test de la partición 3 de CORLEC-EHU-1, obtenidas mediante los modelos de histogramas. El modelo del grupo 1 proporciona las probabilidades más altas en casi todos los casos.

Tabla 6.10: Porcentaje de intervenciones del corpus de test para las que el grupo que da *Max-Prob*: (a) coincide con el que da *Max-HProb*, y (b) está incluido en el subconjunto *Haz-HProb*.

	(a)	(b)
SENGLAR16: MOMd	87.43	100.00
SENGLAR16: MOMc32g	86.00	100.00
INFOTREN: MOMc32g	62.99	98.38
CORLEC-EHU-1: p1, MOMc32g	73.31	98.06
CORLEC-EHU-1: p2, MOMc32g	64.63	93.47
CORLEC-EHU-1: p3, MOMc32g	87.18	99.58

Tabla 6.11: Porcentaje de intervenciones del corpus de test para las que el grupo que da la tasa de DAF más alta: (a) coincide con el grupo que da *MaxProb*, y (b) coincide con el grupo que da *MaxHProb*.

	(a)	(b)
SENGLAR16: MOMd	46.86	45.00
SENGLAR16: MOMc32g	52.14	48.71
INFOTREN: MOMc32g	28.90	26.30
CORLEC-EHU-1: p1, MOMc32g	34.02	34.16
CORLEC-EHU-1: p2, MOMc32g	30.11	25.14
CORLEC-EHU-1: p3, MOMc32g	36.20	34.37

El éxito de la adaptación basada en grupos de hablantes depende fuertemente de la selección/composición de los modelos acústicos. Aún suponiendo que los modelos de histogramas consiguieran asignar sin ambigüedad un único grupo a cada hablante del conjunto de test, cabe hacerse varias preguntas.

Por ejemplo, ¿en qué medida coinciden las decisiones tomadas a partir de los modelos de histogramas con las que se tomarían a partir de las probabilidades de las decodificaciones? En la Tabla 6.10 se muestra el porcentaje de intervenciones del corpus de test para las que el grupo que proporciona la decodificación de máxima probabilidad coincide con el que da la máxima probabilidad según los modelos de histogramas. También se indica el porcentaje de intervenciones para las que el grupo que proporciona la decodificación de máxima probabilidad está incluido en el subconjunto de grupos cuyos modelos de histogramas dan una probabilidad por encima de $0,7 \cdot \text{MaxHProb}$. Es destacable que aunque las coincidencias entre *MaxProb* y *MaxHProb* apenas superan el 85 % con habla leída y pueden descender hasta el 63 % con habla espontánea, la decodificación de máxima probabilidad está incluida en el subconjunto *Haz-HProb* en un porcentaje igual o muy cercano al 100 % en casi todos los casos. De ahí que las tasas de DAF obtenidas con la alternativa *Haz-HProb/Max-Prob* sean prácticamente iguales a las obtenidas con la alternativa *Max-Prob*.

Otra pregunta: ¿es la decodificación de máxima probabilidad la que proporciona la tasa de DAF más alta? En el caso de SENGLAR16, esto es cierto sólo para el 50 % de las intervenciones —véase la Tabla 6.11, columna (a). Para INFOTREN y CORLEC-EHU-1 el porcentaje es aún más bajo, de alrededor del 30 %. Si se utilizan los modelos de histogramas para calcular la probabilidad, los resultados son muy similares —Tabla 6.11, columna (b).

Y una última cuestión: si diéramos con la función de selección perfecta, ¿qué tasa de DAF

Tabla 6.12: Tasa máxima de DAF alcanzable con los grupos de hablantes definidos sobre SENGLAR16, INFOTREN y las tres particiones de CORLEC-EHU-1.

	%DAF máximo
SENGLAR16: MOMd	68.78
SENGLAR16: MOMc32g	76.17
INFOTREN: MOMc32g	66.87
CORLEC-EHU-1: p1, MOMc32g	53.23
CORLEC-EHU-1: p2, MOMc32g	55.37
CORLEC-EHU-1: p3, MOMc32g	54.84

podría alcanzarse? Para cada intervención del corpus de test se ha elegido, de entre las decodificaciones correspondientes a los distintos grupos, aquella que produce la tasa de DAF más alta, de donde se obtiene un conjunto de decodificaciones. A partir de este conjunto de decodificaciones se calcula una tasa de DAF, que aquí consideramos como una estimación de la *tasa máxima alcanzable* con una cierta partición del corpus de entrenamiento en grupos de hablantes. Las *tasas máximas* obtenidas para SENGLAR16, INFOTREN y las tres particiones de CORLEC-EHU-1 revelan que la función de selección podría aportar aún entre uno y tres puntos más a la tasa de DAF, dependiendo del caso —véase la Tabla 6.12. Sin embargo, definir una función de selección capaz de determinar *a priori* cuál de los grupos va a proporcionar la tasa de DAF más alta es sencillamente imposible. Como demuestra implícitamente la Tabla 6.12, la decodificación de máxima probabilidad no es, en general, la que proporciona la tasa de DAF más alta.

6.4.3. Resultados de DAF con modelos de género.

En la Tabla 6.13 se muestran los resultados de DAF obtenidos con modelos de género. Se incluyen tres tasas: la que se obtiene sobre el subcorpus de test de mujeres con los modelos de mujeres, la que se obtiene sobre el subcorpus de test de hombres con los modelos de hombres, y la que se obtiene sobre el corpus de test completo seleccionando aquel conjunto de modelos que produzca la decodificación más probable. No se han utilizado en este caso los modelos de histogramas. Además de las tasas, se muestra también el porcentaje de intervenciones cuyo género ha sido identificado correctamente.

Sobre SENGLAR16 se han continuado las dos líneas de experimentación, estimando MOM discretos y continuos de hombres y mujeres y utilizando el conjunto de unidades subléxicas básico (23 fonemas + silencio). En ambos casos la definición de modelos de género produce mejoras similares a las que se obtienen con grupos de hablantes —más marcadas en el caso de las mujeres—, y además los costes de estimación y reconocimiento son más bajos. El porcentaje de acierto en la determinación del género, utilizando la probabilidad de las decodificaciones, está comprendido entre el 98 % y el 99 %. Si se utilizan modelos de histogramas de hombres y mujeres, dicho porcentaje desciende tan sólo hasta el 96 %.

Para INFOTREN y CORLEC-EHU-1 se han estimado únicamente MOM continuos de hombres y mujeres, utilizando el conjunto de unidades subléxicas ampliado (23 fonemas + 13 o 14 fenómenos de habla espontánea). En ambos casos los modelos de género igualan o incluso superan claramente el rendimiento de los modelos de grupos de hablantes. En el caso de INFOTREN, con éstos últimos se obtenía una tasa del 63.61 %, mientras que con modelos de

Tabla 6.13: Tasas de DAF obtenidas mediante modelos de género: (a) sobre el subcorpus de mujeres con los modelos de mujeres, (b) sobre el subcorpus de hombres con los modelos de hombres, y (c) sobre el corpus de test completo con los modelos (hombres/mujeres) que proporcionen la decodificación más probable. En la última columna se indica el porcentaje de intervenciones cuyo género se ha identificado correctamente en el caso (c).

	Mujeres	Hombres	Test completo (Max-Prob)	% Acierto en la determinación del género
SENGLAR16: MOMd	67.61	66.13	66.75	98.29
SENGLAR16: MOMc32g	74.93	73.83	74.31	98.86
INFOTREN: MOMc32g	65.57	63.99	64.73	95.45
CORLEC-EHU-1: p1, MOMc32g	48.45	52.49	51.97	82.71
CORLEC-EHU-1: p2, MOMc32g	54.77	53.77	54.04	90.91
CORLEC-EHU-1: p3, MOMc32g	52.86	54.88	54.57	90.00

género se llega hasta el 64.73 %, y, como sucedía con SENGLAR16, la mejora es más marcada para las mujeres. En el caso de CORLEC-EHU-1, la mejora más significativa se obtiene sobre la partición 3 (54.57 %, un punto por encima de la tasa obtenida con grupos de hablantes); sobre las particiones 1 y 2 los resultados son similares a los que se obtienen con grupos de hablantes. Por otra parte, el acierto en la determinación del género es del 95.45 % en el caso de INFOTREN, y de alrededor del 90 % para las particiones 2 y 3 de CORLEC-EHU-1. La partición 1 se revela como particularmente difícil: el acierto en la determinación del género es tan sólo del 82.71 %, lo cual ha influido, sin duda, en que la tasa de DAF sea sensiblemente inferior a la que se obtiene sobre las particiones 2 y 3. Por último, si se utilizan modelos de histogramas de hombres y mujeres para determinar el género, el acierto es del 90.26 % para INFOTREN, y del 70.68 %, 79.69 % y 86.34 % para las particiones 1, 2 y 3 de CORLEC-EHU-1, respectivamente.

En definitiva, los modelos de género proporcionan tasas de DAF iguales o superiores a las que se obtienen con modelos de grupos de hablantes (véase la Tabla 6.14). La variabilidad interna de los conjuntos de hombres y mujeres utilizados para estimar los MOM de género ha de ser forzosamente mayor que la de los grupos de hablantes. Esto significa que la *precisión* de los modelos de género debería ser menor que la de los modelos de grupos. Su mejor rendimiento sólo puede ser debido, por tanto, a una mayor robustez de las estimaciones, y estaría poniendo de manifiesto que ninguna de las bases de datos consideradas dispone de muestras/hablantes suficientes como para entrenar de forma robusta MOM de grupos de hablantes. Y ello a pesar de que se ha realizado una estimación MAP de los MOM de grupos de hablantes —que garantiza un grado mínimo de robustez—, mientras que los MOM de género se han estimado aplicando el criterio de máxima similitud.

Finalmente, los porcentajes de acierto en la determinación del género obtenidos mediante modelos de histogramas de hombres y mujeres son lo bastante elevados, especialmente en los casos de SENGLAR16 e INFOTREN, como para permitir la utilización de modelos de género sin incrementar apenas el coste de reconocimiento.

6.4.4. Resultados de DAF con unidades subléxicas contextuales.

En este apartado se presenta el resultado de combinar modelos de grupos de hablantes con unidades subléxicas contextuales. Los modelos de unidades contextuales correspondientes a

Tabla 6.14: Resumen de las tasas de DAF obtenidas mediante modelos de grupos de hablantes (Haz-HProb) y de género (Max-Prob). Entre paréntesis se indica el número promedio de pasadas de reconocimiento que es necesario realizar en cada caso. También se muestran las tasas de referencia, sin adaptación al hablante, obtenidas con MOM de unidades incontextuales.

	Referencia	Grupos	Género
SENGLAR16: MOMd	65.28 (1)	66.51 (2.08)	66.75 (2)
SENGLAR16: MOMc32g	72.38 (1)	74.41 (2.08)	74.31 (2)
INFOTREN: MOMc32g	61.34 (1)	63.60 (3.37)	64.73 (2)
CORLEC-EHU-1: p1, MOMc32g	50.63 (1)	51.60 (2.63)	51.97 (2)
CORLEC-EHU-1: p2, MOMc32g	52.42 (1)	53.89 (3.42)	54.04 (2)
CORLEC-EHU-1: p3, MOMc32g	53.13 (1)	53.58 (2.13)	54.57 (2)

Tabla 6.15: Tasas de DAF obtenidas con modelos de grupos de hablantes y unidades subléxicas contextuales. Se han aplicado dos criterios de selección de grupo: *MaxProb* y *MaxHProb*. En este último caso, entre paréntesis se indica el porcentaje de coincidencias con el criterio *MaxProb*.

	Bifonemas Izquierda		Bifonemas Derecha		Trifonemas	
	MaxProb	MaxHProb	MaxProb	MaxHProb	MaxProb	MaxHProb
SENGLAR16: MOMd	69.00	68.82 (81.00)	69.17	69.05 (79.14)	65.36	65.34 (56.29)
SENGLAR16: MOMc32g	75.79	75.63 (80.29)	75.93	75.71 (81.86)	72.42	72.42 (54.14)
INFOTREN: MOMc32g	65.57	65.15 (54.22)	67.40	66.99 (51.62)	62.76	62.69 (36.03)
CORLEC-EHU-1: p2, MOMc32g	58.05	57.52 (56.96)	58.10	57.53 (58.81)	56.20	56.06 (51.42)

cada grupo de hablantes se inicializan con los modelos generales, obtenidos a partir del corpus de entrenamiento completo según se ha explicado en el apartado 5.2.3. A continuación se aplica el método de reestimación MAP, utilizando los factores de adaptación τ óptimos obtenidos previamente. Los resultados presentados en la Tabla 6.15 corresponden a la utilización por separado de bifonemas a izquierda, bifonemas a derecha y trifonemas —completados con las unidades incontextuales con objeto de garantizar la cobertura— para hacer un test de DAF sobre las distintas bases de datos. Se han aplicado los criterios de selección de grupo *MaxProb* y *MaxHProb*. En este último caso, se ha obtenido el porcentaje de coincidencias con el criterio *MaxProb*, que en la tabla aparece indicado entre paréntesis. Al aumentar el número de unidades subléxicas, la estimación de los modelos y los experimentos de DAF se hacen mucho más costosos, especialmente en el caso de CORLEC-EHU-1, razón por la cual la experimentación sobre dicha base de datos se ha limitado a la partición 2.

La escasez de muestras observada en los experimentos con modelos de grupos de hablantes y unidades incontextuales se hace aún más notoria en estos experimentos, ya que se tiene un mayor número de unidades subléxicas. Así pues, las mejoras que se producen —si se producen— son muy pequeñas con respecto a los resultados obtenidos sin adaptación al hablante (véanse las Tablas 5.12 y 5.13).

En el caso de SENGLAR16, las tasas con MOM discretos son, de hecho, inferiores a las que se obtienen sin adaptación al hablante: con MOM continuos de bifonemas las mejoras no llegan a medio punto, y con MOM continuos de trifonemas se obtiene una tasa ligeramente inferior a la que se obtiene sin adaptación al hablante. En todos los casos es más importante la aportación de los bifonemas que la de los trifonemas, ya que éstos últimos, además de ser escasos, reciben

Tabla 6.16: Tasas de DAF obtenidas mediante modelos de género (MOM continuos de 32 gaussianas) y unidades subléxicas contextuales: (a) sobre el subcorpus de test de mujeres con los modelos de mujeres, (b) sobre el subcorpus de test de hombres con los modelos de hombres, y (c) sobre el corpus de test completo con los modelos (hombres/mujeres) que proporcionen la decodificación más probable. Junto a esta última tasa, entre paréntesis, se indica el porcentaje de intervenciones cuyo género se ha identificado correctamente.

SENGLAR16, MOMc32g, MAP($\tau = 5$)			
	(a)	(b)	(c)
Bifonemas izquierda	77.29	76.00	76.58 (99.71)
Bifonemas derecha	77.19	76.38	76.75 (99.00)
Trifonemas	74.81	73.89	74.30 (99.14)

INFOTREN, MOMc32g, MAP($\tau = 2$)			
	(a)	(b)	(c)
Bifonemas izquierda	68.81	67.26	67.98 (95.45)
Bifonemas derecha	69.45	68.05	68.69 (95.78)
Trifonemas	66.52	64.81	65.63 (95.78)

CORLEC-EHU-1, p2, MOMc32g, MAP($\tau = 20$)			
	(a)	(b)	(c)
Bifonemas izquierda	58.16	58.33	58.47 (91.76)
Bifonemas derecha	58.23	58.30	58.44 (91.48)
Trifonemas	56.73	57.40	56.31 (89.77)

muy pocas muestras. En las dos bases de datos de habla espontánea, la utilización de bifonemas puede incrementar las tasas hasta en un punto. En lo que se refiere al criterio de selección de grupo, los porcentajes de coincidencia de *MaxHProb* y *MaxProb* son bastante bajos, especialmente en el caso de los trifonemas. Con bifonemas, los porcentajes están en torno al 80 % en el caso de SENGLAR16, y entre el 50 % y el 60 % para las bases de datos de habla espontánea. Aún así, las tasas que se obtienen con *MaxHProb* son sólo ligeramente inferiores a las que se obtienen con *MaxProb*, lo cual confirma nuevamente la utilidad de los modelos de histogramas.

También se ha explorado, como alternativa, la combinación de modelos de género con unidades subléxicas contextuales. Se han llevado a cabo tres series de experimentos: sobre el subcorpus de test de mujeres con los modelos de mujeres, sobre el subcorpus de test de hombres con los modelos de hombres, y sobre el corpus de test completo con los modelos (hombres/mujeres) que proporcionen en cada caso la decodificación más probable. Los modelos de género de unidades contextuales se inicializan con los modelos de unidades contextuales generales. A continuación se aplica el método de reestimación MAP, utilizando los subconjuntos de entrenamiento de hombres y mujeres y los factores de adaptación τ óptimos obtenidos previamente.

Como puede observarse en la Tabla 6.16, la adición de modelos de género supone una mejora significativa del rendimiento en DAF, mayor que la observada utilizando modelos de grupos de hablantes. En el caso de SENGLAR16 las tasas se incrementan entre 1 y 1.5 puntos con respecto a los modelos generales de unidades contextuales (véase la Tabla 5.13). En INFOTREN las mejoras son todavía mayores: en torno a 3 puntos para los bifonemas a izquierda y los trifonemas, y cerca de 2 puntos para los bifonemas a derecha. Para ambas bases de datos, las mejoras son más acusadas en el caso de las mujeres.

Tabla 6.17: Tasas de DAF obtenidas sobre los corpus de test de SENGLAR16, INFOTREN y la partición 2 de CORLEC-EHU-1, utilizando modelos de género y modelos de grupos de hablantes (MOM continuos de 32 gaussianas), con unidades subléxicas incontextuales (ICXT) y contextuales (CXT, bifonemas a derecha). También se muestran las tasas de DAF obtenidas sin adaptación al hablante.

		SENGLAR16	INFOTREN	CORLEC-EHU-1: p2
Sin adaptación al hablante	ICXT	72.38	61.34	52.42
	CXT	75.55	66.98	56.98
Modelos de género	ICXT	74.31	64.73	54.04
	CXT	76.75	68.69	58.44
Modelos de grupos	ICXT	74.41	63.60	53.89
	CXT	75.93	67.40	58.10

En la partición 2 de CORLEC-EHU-1 los modelos de género de unidades contextuales introducen, con respecto a los modelos generales (véase, de nuevo, la Tabla 5.13), incrementos de 1.5 puntos para los bifonemas a izquierda y derecha, y de tan sólo 0.4 puntos para los trifonemas. No hay diferencias significativas entre las tasas para hombres y las tasas para mujeres. En el caso de los bifonemas, la tasa conjunta (c) es mayor que las tasas que se obtienen separadamente para mujeres (a) y hombres (b). Esto significa que en ocasiones los modelos del género opuesto —elegidos *erróneamente* por la función de selección— resultan más adecuados que los del propio género. Curiosamente, en el caso de los trifonemas sucede justo lo contrario: la tasa conjunta (c) es inferior a las dos tasas separadas (a) y (b), aún cuando el porcentaje de acierto del género está en el 90 %. Precisamente, el porcentaje de acierto del género es muy alto en todos los casos, muy cercano al 100 % para SENGLAR16, superior al 95 % para INFOTREN y en torno al 90 % para CORLEC-EHU-1, lo cual indica que los modelos acústicos realmente consiguen diferenciar entre ambos tipos de voces. En la Tabla 6.17 se resumen las tasas de DAF obtenidas sobre BDGEO-16, INFOTREN y la partición 2 de CORLEC-EHU-1, utilizando modelos de grupos de hablantes y modelos de género (MOM continuos de 32 gaussianas) con unidades subléxicas incontextuales y contextuales.

6.5. Resultados de reconocimiento.

Para completar el análisis de la modelización del hablante propuesta en este capítulo, se han llevado a cabo experimentos de reconocimiento, tanto sobre el corpus de aplicación de habla leída BDGEO-16, utilizada para medir el rendimiento de los modelos acústicos obtenidos sobre SENGLAR16, como sobre las dos bases de datos de habla espontánea, INFOTREN y CORLEC-EHU-1, objeto de este trabajo.

En primer lugar se han llevado a cabo experimentos con los modelos de género de unidades incontextuales obtenidos sobre SENGLAR16, utilizando el vocabulario, el modelo de lenguaje ($k = 3$) y el corpus de test definidos para BDGEO-16, tal como se describe en el apartado 4.4. En la Tabla 6.18 se muestran los resultados obtenidos en los cuatro experimentos realizados: (a) utilizando los modelos de mujeres sobre el subcorpus de mujeres del corpus de test de BDGEO-16, (b) utilizando los modelos de hombres sobre el subcorpus de hombres del corpus de test de BDGEO-16, y utilizando ambos modelos sobre el corpus de test de BDGEO-16, (c) tomando como función de selección *MaxProb*, o (d) tomando como función de selección *MaxHProb*.

Tabla 6.18: Error de reconocimiento (%WER) sobre el corpus de test de BDGEO-16, utilizando los modelos de género de unidades incontextuales obtenidos sobre SENGLAR16 y el modelo de lenguaje de BDGEO-16 ($k = 3$, $\alpha = 5$ para MOM discretos y $\alpha = 9$ para MOM continuos).

	%WER			
	Mujeres	Hombres	Max-Prob	Max-HProb
MOMd	11.37	10.48	12.78	11.65
MOMc32g	6.48	6.57	9.78	7.28

Tabla 6.19: Resultados de reconocimiento (%WER) sobre los corpus de test de BDGEO-16, INFOTREN y la partición 2 de CORLEC-EHU-1, utilizando modelos de género y modelos de grupos de hablantes (MOM continuos de 32 gaussianas), con unidades subléxicas incontextuales (ICXT) y contextuales (CXT). Entre paréntesis se indica el valor óptimo de α en cada caso. También se muestra el error de reconocimiento obtenido sin adaptación al hablante.

		%WER		
		BDGEO-16	INFOTREN	CORLEC-EHU-1: C2 (ML123)
Sin adaptación al hablante	ICXT	6.41 ($\alpha = 9$)	29.32 ($\alpha = 8$)	36.36 ($\alpha = 10$)
	CXT	6.21 ($\alpha = 9$)	27.22 ($\alpha = 8$)	28.64 ($\alpha = 10$)
Modelos de género	ICXT	7.28 ($\alpha = 9$)	29.00 ($\alpha = 7$)	36.92 ($\alpha = 10$)
	CXT	10.15 ($\alpha = 9$)	26.85 ($\alpha = 8$)	28.49 ($\alpha = 10$)
Modelos de grupos	ICXT	6.83 ($\alpha = 9$)	29.12 ($\alpha = 8$)	35.73 ($\alpha = 10$)
	CXT	10.85 ($\alpha = 9$)	30.74 ($\alpha = 8$)	29.14 ($\alpha = 10$)

Nótese que la probabilidad de la mejor decodificación no es ya la probabilidad acústica sino la probabilidad conjunta de la secuencia acústica y de la secuencia de palabras, mientras que la probabilidad que proporcionan los modelos de histogramas sigue siendo únicamente una probabilidad acústica. Esto hace que la selección del género —que debe estar basada exclusivamente en información acústica— se vea distorsionada por el modelo de lenguaje en el caso (c) y se obtengan peores resultados que si se utilizan los modelos de histogramas de hombres y mujeres. Así pues, en adelante se aplicarán los modelos de histogramas tanto en la selección del género como en la selección del grupo —cuando se utilicen modelos de grupos de hablantes. Obsérvese que ni siquiera en el caso (d) consiguen superarse las tasas obtenidas con modelos generales. Ello indica que la ganancia en *precisión* de los modelos de género no llega a compensar la pérdida de robustez que se produce al fragmentar el corpus de entrenamiento.

En la Tabla 6.19 se muestra el error de reconocimiento para BDGEO-16, INFOTREN y la partición 2 de CORLEC-EHU-1, utilizando modelos acústicos de género y de grupos de hablantes, en concreto los modelos óptimos obtenidos en DAF (MOM continuos de 32 gaussianas), unidades subléxicas incontextuales y unidades subléxicas contextuales, y modelos de lenguaje k -explorables en sentido estricto con $k = 3$. Como quedó establecido en el capítulo 5, los modelos de lenguaje de INFOTREN y CORLEC-EHU-1 se han estimado a partir de textos que incluyen fenómenos de habla espontánea, los cuales se han integrado en el modelo de manera natural, como pseudo-palabras. Por otra parte, como ya se indicó en el capítulo 4, el modelo de lenguaje de CORLEC-EHU-1 no es independiente del corpus de test, condición que se ha mantenido inalterada en toda la experimentación sobre dicha base de datos, con objeto de evaluar las distintas mejoras incorporadas a los modelos acústicos.

El error de reconocimiento sobre BDGEO-16 es del 6.83 %, casi medio punto por debajo del obtenido con modelos de género, pero no llega al 6.41 % que se obtenía con los modelos generales. Esto pone de manifiesto nuevamente que los modelos de grupos de hablantes pueden ser más precisos pero no tan robustos como los modelos generales, y en el balance resultan mejor éstos últimos. La estimación de modelos contextuales diferenciados para cada género o, en general, para cada grupo de hablantes, no hace sino empeorar la situación, como ha quedado demostrado en los experimentos de DAF, ya que las muestras se dispersan aún más. Si se aplican modelos contextuales de género sobre BDGEO-16, se obtiene un error del 10.15 %. Si se aplican modelos contextuales de grupos de hablantes el error es aún mayor, del 10.85 %. Así pues, al menos en lo que se refiere a SENGLAR16, queda claro que el número de muestras no permite definir grupos de hablantes. En realidad, ni siquiera los dos grupos *naturales* hombres/mujeres parecen alcanzar una robustez suficiente.

En el caso de INFOTREN, los modelos de género suponen mejoras del 1 % o ligeramente superiores con respecto a los modelos generales; por su parte, los modelos de grupos de hablantes sólo producen una pequeña mejora (del 0.7 %) cuando se combinan con unidades subléxicas incontextuales, lo cual podría indicar que los modelos acústicos no ofrecen la suficiente robustez, o que la forma de construir modelos de palabra a partir de modelos contextuales no es la más adecuada. En lo que respecta a CORLEC-EHU-1, los modelos de género sólo aportan una pequeña mejora (del 0.52 %) cuando se combinan con unidades contextuales, e inversamente, los modelos de grupos de hablantes introducen una mejora del 1.73 % sólo cuando se combinan con unidades incontextuales. El hecho de que esta fragmentación del corpus de entrenamiento en grupos de hablantes sólo produzca pequeños descensos —o incluso, leves incrementos— del error de reconocimiento debe interpretarse como una señal de falta de robustez, es decir, de falta de muestras. En reconocimiento de habla espontánea, los errores más pequeños se obtienen combinando los modelos de género con unidades subléxicas contextuales: 26.85 % para INFOTREN y 28.49 % para CORLEC-EHU-1.

Resumen del capítulo.

En este capítulo se ha presentado un algoritmo de *clustering* de hablantes que trata de reducir la variabilidad acústica interna en cada uno de los grupos definidos, y obtener, por tanto, modelos más precisos. El algoritmo es una variación del conocido LBG utilizado en cuantificación vectorial, y utiliza un modelo acústico simplificado para cada hablante o grupo de hablantes, basado en histogramas de etiquetas microfonéticas. A partir de este modelo se definen sendas medidas de disimilitud, entre un hablante y un grupo de hablantes, y entre dos hablantes, que determinan la configuración final de la partición. Se han explorado distintas estrategias para la selección o combinación de los modelos más adecuados a cada señal de test. En el caso de utilizar como criterio la probabilidad de las muestras condicionada por el propio modelo simplificado, se ha comprobado que basta con 2 o 3 segundos de señal para alcanzar el rendimiento óptimo de esta metodología.

Los resultados del *clustering* sugieren que *realmente* se están identificando grupos de hablantes diferenciados, ya que se separan por género y también, en determinados casos, por condiciones de grabación. Por otro lado, la función de selección basada en el modelo simplificado tiene unos índices de acierto muy notables entre los hablantes del conjunto de entrenamiento. Esa misma función de selección muestra asimismo una gran coherencia entre los hablantes del conjunto de test.

Una vez definida la partición, se estiman MOM diferentes para cada grupo de hablantes. Los resultados de DAF obtenidos con estos modelos mejoran los obtenidos con modelos generales, aunque en general son inferiores a los obtenidos con modelos de género. La combinación de modelos de género y unidades subléxicas contextuales es la que proporciona las mejores tasas de DAF tanto sobre habla espontánea como sobre habla leída. Los modelos de grupos de hablantes no son lo bastante robustos, debido a una excesiva fragmentación del corpus de entrenamiento. Sin embargo, si se dispone de bases de datos con un número suficiente de hablantes, y las condiciones ambientales y de canal son más o menos estacionarias, la partición del corpus de entrenamiento en grupos de hablantes con características similares y la estimación de modelos acústicos diferenciados debería superar a los modelos de género, que resultarían poco precisos. Por otra parte, las tasas máximas de DAF alcanzables revelan que los modelos acústicos de grupos de hablantes tienen mayor capacidad de la que muestran en la práctica, y que un margen de mejora importante depende de la función de selección.

En definitiva, el algoritmo de *clustering* propuesto ha mostrado un comportamiento muy coherente tanto sobre habla leída como sobre habla espontánea, incluso, en este último caso, con condiciones ambientales y de canal muy variables. Se ha planteado una estrategia de adaptación al hablante basada en elegir el grupo de hablantes más afín a la señal de entrada. Esta estrategia no aporta mejoras significativas en reconocimiento, debido fundamentalmente a la falta de robustez de los modelos acústicos. Es de esperar que al aplicarla sobre bases de datos suficientemente grandes, que permitan obtener estimaciones robustas de los modelos acústicos, las mejoras sean mucho más notables.

Capítulo 7

Conclusiones y trabajo futuro.

7.1. Conclusiones.

Esta tesis parte del concepto de *fenómeno de habla espontánea*, definido como cualquier característica acústica, léxica, sintáctica o incluso pragmática que distinga el habla espontánea del habla leída, frente al concepto más restringido de *disfluencia*, que se refiere a cualquier ruptura de la continuidad acústica o sintáctica del habla. Las disfluencias son consideradas todas ellas como fenómenos de habla espontánea, ya que no aparecen en habla leída, pero también lo son otros fenómenos, como por ejemplo los alargamientos de sonidos, que, aunque no rompen la continuidad del habla, tampoco aparecen en habla leída. Teniendo en cuenta esta definición, se ha elaborado un inventario de los fenómenos más comunes en habla espontánea, y se han discutido sus características, sus realizaciones más comunes y el tratamiento que han recibido en la literatura. A partir de este inventario de fenómenos se han realizado los trabajos que conforman la tesis, cuyas principales aportaciones se enumeran a continuación:

1. Tras una revisión bibliográfica y una discusión de las ventajas y desventajas de los distintos esquemas de anotación y de los formatos y herramientas asociados, todos ellos desarrollados para la lengua inglesa, se ha propuesto un esquema de anotación de fenómenos de habla espontánea en diálogos en castellano, que consta de un formato basado en XML, un formato simplificado que facilita y acelera el proceso de anotación y mejora la legibilidad de las anotaciones, herramientas de libre distribución para la edición de marcas y la visualización de señales, y una guía de anotación que trata de resolver las posibles ambigüedades o problemas de interpretación que puedan surgir, mediante una descripción clara de los fenómenos —ejemplos incluidos—, todo ello encaminado a incrementar la coherencia de las anotaciones.
2. Se han anotado los fenómenos de habla espontánea en dos bases de datos, una de ellas (INFOTREN) formada por diálogos hombre-máquina en una tarea de acceso a información sobre trayectos, horarios y precios de viajes en tren, y la otra (CORLEC-EHU-1) formada por entrevistas tomadas de radio y televisión, completamente libres en cuanto al léxico y la gramática, y por tanto mucho menos restringidas que los diálogos hombre-máquina. La presión cognitiva e interactiva es más alta en el caso de las entrevistas, lo cual, *a priori*, debería propiciar la aparición de un mayor número de fenómenos de habla espontánea. Sin embargo, no se observan diferencias significativas en el número relativo de fenómenos que aparecen en diálogos hombre-máquina y en conversaciones naturales

entre personas. Las diferencias se refieren más bien a la distribución de fenómenos: la proporción de distorsiones léxicas y reformulaciones es notablemente mayor en el habla espontánea natural (más errores y más correcciones). Muchos de los fenómenos aparecen no como respuesta a perturbaciones externas —la presión cognitiva o interactiva a que aludíamos—, sino como recursos del habla espontánea, que permiten corregir o planificar el discurso. En este sentido, es destacable el gran número de pausas habladas, alargamientos y repeticiones, que no deben verse como errores o problemas del habla sino que cumplen una función concreta: marcar la apropiación del turno mientras se planifica el discurso. Depende de los usos del hablante el que se utilicen unas u otras formas. En particular, los alargamientos y las pausas habladas muestran porcentajes de aparición comparables, si bien en la literatura las pausas habladas han recibido una atención mucho mayor que los alargamientos.

Por otra parte, el número de fenómenos que aparecen en una intervención muestra una correlación notable con la longitud de la misma —expresada ésta en segundos o en número de palabras efectivas—, particularmente en el caso de los ruidos producidos por el propio hablante y las disfluencias acústicas. Se han confirmado los resultados de Shriberg en cuanto a que la probabilidad de que una intervención sea fluida decrece exponencialmente con la longitud de la misma, y depende de un parámetro b , conocido como *tasa de fluencia*. Se ha comprobado asimismo que la tasa de fluencia de INFOTREN ($b = 0,8838$) es prácticamente igual a la de CORLEC-EHU-1 ($b = 0,8996$). La probabilidad de que tras una palabra se produzca una disfluencia es aproximadamente de 0,1 en ambos casos. Por otra parte, el número de fenómenos varía notablemente de unos hablantes a otros. Ello es debido principalmente a que unos hablantes hablan durante más tiempo que otros. No obstante, se han presentado ejemplos que demuestran que tanto en INFOTREN como en CORLEC-EHU-1 existen hablantes intrínsecamente más disfluentes que otros —con tiempos de elocución similares. Por último, a juzgar por los datos obtenidos, no parece posible determinar la presencia de disfluencias sintácticas utilizando como indicador únicamente la presencia de disfluencias acústicas, distorsiones léxicas o marcadores de discurso. Así, por ejemplo, entre el 50 % y el 60 % de las disfluencias sintácticas contienen alargamientos (típicamente en el último fonema del *reparandum*), pero alrededor del 70 % de los alargamientos aparece fuera de disfluencias sintácticas. En el mismo sentido, el 80 % de las palabras cortadas aparece como elemento desencadenante de una reformulación, pero sólo una pequeña fracción de las reformulaciones están motivadas por la presencia de una palabra cortada.

3. Se ha comparado el rendimiento de los modelos acústicos sobre habla espontánea con respecto al rendimiento obtenido sobre habla leída, utilizando la configuración básica del sistema para habla leída: unidades subléxicas incontextuales (23 fonemas + silencio), modelos léxicos lineales y modelos de lenguaje k -explorables en sentido estricto (típicamente, con $k = 3$). Las diferencias son muy notables, tanto en DAF (entre 10 y 20 puntos absolutos) como en reconocimiento (entre 30 y 50 puntos absolutos). Las causas hay que buscarlas en las condiciones ambientales y de canal, en la presencia de fenómenos acústicos y léxicos no modelados y en la mayor variabilidad acústica, de pronunciaciones y de hablantes.
4. Se han modelado explícitamente los fenómenos de habla espontánea de tipo acústico y se han incluido en el léxico y en el modelo de lenguaje como pseudo-palabras —13 en el

caso de INFOTREN, 14 en el caso de CORLEC-EHU-1. Como resultado, el rendimiento del reconocedor mejora significativamente. La inclusión de las pseudo-palabras en el corpus de entrenamiento y su integración plena en el modelo de lenguaje desde el principio ha mostrado un mejor rendimiento que si simplemente se añaden mediante suavizado como palabras no vistas. Las mejoras son muy claras en INFOTREN, pero no tanto en CORLEC-EHU-1, debido fundamentalmente al nivel de ruido de las grabaciones, que distorsiona los parámetros acústicos e impide que los modelos se definan con la suficiente precisión durante el entrenamiento. También hay que tener en cuenta que los fenómenos añadidos representan alrededor del 6 % del total de palabras en INFOTREN, mientras que en CORLEC-EHU-1 suponen menos del 3 %. A la modelización explícita de los fenómenos de habla espontánea se ha añadido la modelización de los efectos de coarticulación en las unidades subléxicas, mediante la definición de unidades contextuales, que en DAF aportan reducciones del error de entre el 10 % y el 13 %, mientras que en reconocimiento la reducción del error es del 7 % sobre INFOTREN, y superior al 20 % sobre la partición 2 de CORLEC-EHU-1.

5. Finalmente, se ha presentado un nuevo algoritmo de *clustering* del conjunto de hablantes que trata de reducir la variabilidad debida al hablante dentro de los modelos acústicos, de modo que éstos den cuenta solamente de las diferencias entre los distintos sonidos de la lengua. El algoritmo propuesto es una variación del LBG utilizado en cuantificación vectorial. Parte de un único conjunto de hablantes, el cual se va dividiendo sucesivamente hasta que o bien alguno de los conjuntos resultantes no dispone de muestras suficientes, o bien se cumple un cierto criterio de convergencia basado en una medida de disimilitud entre grupos de hablantes. A su vez, esta medida de disimilitud se basa en un modelo acústico simplificado (histogramas de etiquetas microfonéticas). Una vez obtenida la partición, se estima un modelo acústico diferente para cada grupo de hablantes. En la fase de test, se aplica una técnica de selección del grupo más adecuado a cada señal de entrada. Se ha comprobado que basta con utilizar un fragmento de la señal (de entre 2 y 3 segundos) para alcanzar el rendimiento óptimo de esta metodología. Los resultados del *clustering* sugieren que *realmente* se están identificando grupos de hablantes diferenciados, ya que se separan por género y también, en determinados casos, por condiciones de grabación. Por otro lado, los índices de acierto de la función de selección entre los hablantes del conjunto de entrenamiento son muy notables. Esa misma función de selección muestra asimismo una gran coherencia entre los hablantes del conjunto de test.

Los resultados de DAF obtenidos con modelos de grupos de hablantes mejoran los obtenidos con modelos generales, aunque no superan los obtenidos con modelos de género, debido a la mayor robustez de éstos últimos. Precisamente, la combinación de modelos de género y unidades subléxicas contextuales es la que proporciona las mejores tasas de DAF de esta tesis, tanto sobre habla espontánea como sobre habla leída. Con respecto al rendimiento en reconocimiento, la modelización del hablante —en particular, el uso de modelos de género— sólo introduce pequeñas mejoras. En cualquier caso, si nos atenemos a los resultados del *clustering* y al acierto y coherencia de la función de selección, la metodología propuesta tiene muchas posibilidades de éxito si se aplica sobre bases de datos suficientemente grandes, que permitan obtener estimaciones robustas de los modelos acústicos.

7.1.1. Publicaciones relacionadas con esta tesis.

- ▷ Publicaciones relacionadas con el esquema de anotación de fenómenos de habla espontánea y el análisis de los fenómenos (capítulos 2 y 3):
 1. L.J. Rodríguez, I. Torres, A. Varona. *Annotation of disfluencies in Spanish dialogues*. Proceedings of the IX Symposium Nacional de Reconocimiento de Formas y Análisis de Imágenes (SNRFAI), Volumen 2, páginas 409-416. Benicassim (Castellón), 14-18 de mayo, 2001.
 2. L.J. Rodríguez, I. Torres, A. Varona. *Annotation and analysis of disfluencies in a spontaneous speech corpus in Spanish*. Proceedings of the ISCA Workshop on Disfluency in Spontaneous Speech (DiSS), pages 1-4. University of Edinburgh, Scotland, August 29-31, 2001.
 3. L.J. Rodríguez, I. Torres, A. Varona. *Anotación de disfluencias en un corpus de habla espontánea no específico*. Segundas Jornadas en Tecnología del Habla. Granada, 16-18 de diciembre, 2002.
 4. L.J. Rodríguez, I. Torres. *Annotation and analysis of acoustic and lexical events in a generic corpus of spontaneous speech in Spanish*. Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR), pages 187-190. Tokyo Institute of Technology, April 14-16, 2003.
 5. L.J. Rodríguez, I. Torres, A. Varona. *Annotation and analysis of spontaneous speech events in Spanish*. Enviado a la revista *Language & Speech*.
- ▷ Publicaciones relacionadas con la metodología de estimación de los modelos acústicos (capítulo 4):
 1. L.J. Rodríguez, I. Torres. *Comparative study of the Baum-Welch and Viterbi training algorithms applied to read and spontaneous speech recognition*. Pattern Recognition and Image Analysis (IbPRIA 2003), Lecture Notes in Computer Science, LNCS 2652, pages 847-857. F.J. Perales, A.J.C. Campilho, N. Pérez de la Blanca and A. Sanfeliú (Eds.). Springer Verlag, 2003.
- ▷ Publicaciones relacionadas con la modelización explícita de los fenómenos de habla espontánea de tipo acústico (capítulo 5):
 1. L.J. Rodríguez, I. Torres, A. Varona. *Evaluation of sublexical and lexical models of acoustic disfluencies for spontaneous speech recognition in Spanish*. Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH), pages 1665-1668. Aalborg, Denmark, September 2-7, 2001.
- ▷ Publicaciones relacionadas con el *clustering* de hablantes y la adaptación al hablante mediante selección/combinación del grupo o los grupos más adecuados (capítulo 6):
 1. L.J. Rodríguez, I. Torres. *A speaker clustering algorithm for fast speaker adaptation in continuous speech recognition*. 7th International Conference on TEXT, SPEECH and DIALOGUE, Lecture Notes in Artificial Intelligence (LNAI) series, Springer-Verlag (aceptado, en prensa). Brno (Czech Republic), September 8-11, 2004.

7.2. Trabajo futuro.

Tras la finalización de esta tesis, quedan pendientes varias tareas. A continuación se enumeran las más importantes:

1. Aplicar técnicas de extracción de características robustas frente al ruido o técnicas de compensación del ruido. Los resultados de DAF sobre INFOTREN con la parametrización del ETSI para entornos ruidosos (ETSI, 2003b) (véase el apéndice D) ponen de manifiesto que aún es posible obtener mejoras muy significativas en reconocimiento de habla espontánea modelando o compensando las fuentes de ruido —muy habituales en estas bases de datos— o eliminando variabilidades indeseadas de los parámetros acústicos.
2. Integrar los modelos de grupos de hablantes —o los modelos de género— en los prototipos del GRFyTH, utilizando una fracción de la señal de entrada (alrededor de 2 segundos) para determinar el grupo más adecuado. Asimismo, comprobar el rendimiento de esta metodología, y la de posibles mejoras (nuevos criterios de partición y reagrupamiento y nuevos criterios de selección/combinación de modelos), sobre bases de datos más grandes, con cientos o incluso miles de hablantes.
3. Generar automáticamente modelos de palabra en forma de grafo. Es previsible que estos modelos recojan mejor que los modelos lineales la variedad de pronunciaciones que aparece en habla espontánea. Hay una gran cantidad de trabajos en la literatura que tratan este asunto, desde aproximaciones basadas en reglas fonológicas, en reglas introducidas por expertos o en diccionarios, hasta otras que se basan en datos puramente estadísticos (matrices de confusión), pasando por metodologías mixtas. Los modelos de palabra se elaborarán muy probablemente a partir de unidades contextuales (trifonemas generalizados), y afectarán incluso a los bordes. Precisamente, los modelos de palabra deben ser capaces de recoger los efectos de coarticulación entre palabras —especialmente frecuentes en habla espontánea—, que hacen que ciertos sonidos sean borrados, asimilados o sustituidos por otros. La transición de una palabra a otra se convierte, por tanto, en un elemento crítico del proceso de búsqueda de hipótesis, más aún si, como se ha planteado más arriba, los modelos de palabra se construyen con unidades contextuales, porque en determinados casos el contexto acústico no vendrá dado por el fonema que ocupa el borde de la palabra, sino por el fonema anterior o posterior. Por último, los modelos de palabra deberán contar con la posibilidad de que existan alargamientos iniciales o finales, dependiendo de los fonemas implicados. Esto es particularmente importante, por su frecuencia, en determinado tipo de palabras: conjunciones, preposiciones, artículos y pronombres.
4. Estimar modelos de lenguaje específicos para las reformulaciones. Los modelos de reformulaciones estarán basados en categorías, ya que, en general, no se dispone de muestras suficientes como para estimar modelos basados en palabras. A su vez, estos modelos constituirán supercategorías dentro de un modelo de lenguaje jerárquico, al igual que otras entidades lingüísticas de rango superior a la palabra (sujeto, complementos, frases, etc.). Aunque la definición de estos modelos no debería afectar al rendimiento del reconocedor como tal, sí que tendrá una enorme utilidad desde el punto de vista de la *comprensión del habla*. De no aplicarlos, la *absorción* de las reformulaciones en el modelo de lenguaje

general podría provocar errores graves de comprensión, ya que la secuencia de palabras reconocida —aún con un 100 % de acierto— estaría sujeta a varias interpretaciones.

Apéndice A

Definición del formato de anotación de fenómenos de habla espontánea.

```
<?xml encoding="ISO-8859-1"?>

<!-- elemento raíz -->

<!ELEMENT anotacion (descripcion?,hablante+,turno+)>
<!ATTLIST anotacion
  anotador      CDATA      #IMPLIED
  fecha        CDATA      #IMPLIED
>

<!-- descripción del contenido -->

<!ELEMENT descripcion EMPTY>
<!ATTLIST descripcion
  base_de_datos  CDATA      #IMPLIED
  lengua         (español|eusquera|inglés) #IMPLIED
  tipo_habla     (leída|planeada|espontánea) #IMPLIED
  registro       (familiar|coloquial|formal) #IMPLIED
  tipo_discurso  (tarea|monólogo|entrevista|conversación) #IMPLIED
  dominio        CDATA      #IMPLIED
  fichero_audio  CDATA      #IMPLIED
  frecuencia_kHz (8|16)      #IMPLIED
  codificacion   (mulaw|lineal) #IMPLIED
  bits_muestra   (8|16|32)      #IMPLIED
  lugar          (laboratorio|oficina|exterior) #IMPLIED
  canal          (micrófono|teléfono) #IMPLIED
  nivel_ruido    (nulo|bajo|alto|variable) #IMPLIED
>

<!-- hablante -->

<!ELEMENT hablante EMPTY>
<!ATTLIST hablante
  id            ID          #REQUIRED
  nombre       CDATA      #IMPLIED
  sexo         (hombre|mujer) #IMPLIED
  edad         (niño|adulto|anciano) #IMPLIED
  acento       CDATA      #IMPLIED
```

```

>
<!-- turno -->
<!ELEMENT turno (o?,t?,o?,(no_transcrito|corte)?,continua?)>
<!ATTLIST turno
  id          ID          #REQUIRED
  hablante    IDREF       #REQUIRED
  t_ini       CDATA       #IMPLIED
  t_fin       CDATA       #IMPLIED
  fichero_audio CDATA     #IMPLIED
  t_ss_ini    CDATA       #IMPLIED
  t_ss_fin    CDATA       #IMPLIED
  fichero_audio_ss CDATA   #IMPLIED
>

<!-- segmento de texto sin solapamientos -->
<!ELEMENT t (#PCDATA|a|n|p|f|l|r|b|d|N|sync)*>
<!ATTLIST t
  t_ini CDATA          #IMPLIED
  t_fin CDATA          #IMPLIED
>

<!-- solapamiento -->
<!ELEMENT o (#PCDATA|a|n|p|f|l|r|b|d|N|sync)*>
<!ATTLIST o
  t_ini CDATA          #IMPLIED
  t_fin CDATA          #IMPLIED
>

<!-- ruido aislado -->
<!ELEMENT n EMPTY>
<!ATTLIST n
  fuente (hablante|exterior)          #REQUIRED
  tipo   (aspiración|labios|tos|genérico) #REQUIRED
  t_ini  CDATA                          #IMPLIED
  t_fin  CDATA                          #IMPLIED
>

<!-- ruido externo que abarca señal de voz -->
<!ELEMENT N (#PCDATA|a|n|p|f|l|r|b|d|sync)*>
<!ATTLIST N
  fuente CDATA          #FIXED          "exterior"
  tipo   CDATA          #FIXED          "genérico"
  t_ini  CDATA          #IMPLIED
  t_fin  CDATA          #IMPLIED
>

<!-- alargamiento de un sonido -->
<!ELEMENT a (#PCDATA)>
<!ATTLIST a

```

```

    t_ini CDATA    #IMPLIED
    t_fin CDATA    #IMPLIED
  >

<!-- pausa de silencio -->

<!ELEMENT p EMPTY>
<!ATTLIST p
  t_ini CDATA    #IMPLIED
  t_fin CDATA    #IMPLIED
>

<!-- pausa hablada -->

<!ELEMENT f EMPTY>
<!ATTLIST f
  tipo (a|e|m|sin_identificar) #REQUIRED
  t_ini CDATA                    #IMPLIED
  t_fin CDATA                    #IMPLIED
>

<!-- distorsión léxica -->

<!ELEMENT l (#PCDATA|a)*>
<!ATTLIST l
  tipo (cortada|mal_pronunciada|gutural|acrónimo|extranjera) #REQUIRED
  palabra CDATA #IMPLIED
  pronunciación CDATA #IMPLIED
  t_ini CDATA #IMPLIED
  t_fin CDATA #IMPLIED
>

<!-- reformulación -->

<!ELEMENT r (m,s?,c)>
<!ATTLIST r
  tipo (repetición|sustitución|inserción|borrado) #REQUIRED
  t_ini CDATA #IMPLIED
  t_fin CDATA #IMPLIED
>

<!-- reparandum -->

<!ELEMENT m (#PCDATA|a|n|p|f|l|r|d|N|sync)*>
<!ATTLIST m
  t_ini CDATA    #IMPLIED
  t_fin CDATA    #IMPLIED
>

<!-- señal de edición -->

<!ELEMENT s (#PCDATA|a|n|p|f|l|r|d|N|sync)*>
<!ATTLIST s
  t_ini CDATA    #IMPLIED
  t_fin CDATA    #IMPLIED
>

```

```

<!-- corrección -->

<!ELEMENT c (#PCDATA|a|n|p|f|l|r|d|N|sync)*>
<!ATTLIST c
  t_ini CDATA    #IMPLIED
  t_fin CDATA    #IMPLIED
>

<!-- frase abandonada -->

<!ELEMENT b (#PCDATA|a|n|p|f|l|r|d|N|sync)*>
<!ATTLIST b
  t_ini CDATA    #IMPLIED
  t_fin CDATA    #IMPLIED
>

<!-- marcador de discurso -->

<!ELEMENT d (#PCDATA|a|n|p|f|l|r|N|sync)*>
<!ATTLIST d
  tipo (abrir|cerrar|aceptar|rechazar|editar|pedir|rellenar|exclamar) #REQUIRED
  t_ini CDATA    #IMPLIED
  t_fin CDATA    #IMPLIED
>

<!-- corte en la grabación -->

<!ELEMENT corte EMPTY>
<!ATTLIST corte
  t_ini CDATA    #IMPLIED
  t_fin CDATA    #IMPLIED
>

<!-- fragmento no transcrito o ininteligible -->

<!ELEMENT no_transcrito EMPTY>
<!ATTLIST no_transcrito
  t_ini CDATA    #IMPLIED
  t_fin CDATA    #IMPLIED
>

<!-- marca de continuación -->

<!ELEMENT continua EMPTY>
<!ATTLIST continua>

<!-- marca de sincronización -->

<!ELEMENT sync EMPTY>
<!ATTLIST sync
  t      CDATA    #REQUIRED
>

```

Apéndice B

Ejemplos de anotación.

B.1. Diálogo hombre-máquina.

B.1.1. Versión en formato simplificado.

M0: bienvenido al sistema automático de información de trenes regionales y de largo recorrido . por favor , utilice frases cortas . ¿ qué desea ?

U0: (na) (df mire) . quisiera sab(a e)r (nw)(fa)(fm) el horario de trenes que hay (na)(nw) para ir de Valencia a (lm Madrid Madrit) . (na)(fa) me gustaría que (rr (m l(a a)) (s (fm)) (c la)) llegada a Madrid (ri (m (lu sobre so)) (c (rs (m fuese) (c fuera)) sobre)) las diez de la mañana .

M1: desde Valencia con destino a Madrid . ¿ desea viajar hoy mismo ?

U1: (na)(fa)(fm)(na) no . hoy no . (na) todavía no sé exactamente (a e)l día (rr (m qu(a e)) (c que)) vamos a ir (p) ¿ tengo que decirle un día exacto ?

M2: lo siento , creo que no le he entendido bien . ¿ qué día desea viajar ?

U2: (nl)(na) (df mire) . (df pues) mañana .

M3: le consulto horarios de Valencia a Madrid , el jueves treinta de octubre . un momento , por favor .

U3:

M4: por la mañana hay tres trenes . el primero llega a las diez horas y cuarenta y ocho minutos de la mañana , el segundo a las doce horas y cuarenta y ocho minutos de la mañana y el tercero a las trece horas y cincuenta y ocho minutos . ¿ algo más ?

U4: (nl)(na)(fa)(fm)(p)(na) (df mire) . quisiera volver el mismo día , (de o sea) , mañana por la tarde , (na) a última hora . (na) (df entonces) ¿ puede decirme los trenes que hay (na) (rs (m par(a a) ir) (c para volver)) de Madrid a Valencia ?

M5: le consulto horarios de Madrid a Valencia , el jueves treinta de octubre . un momento , por favor .

U5:

M6: ese día hay diez trenes . el primero sale a siete de la mañana y el último a las veintitrés horas . ¿ algo más ?

U6: (nl)(na) sí . (df mire) . me interesaría sab(a e)r (na)(fe)(p) cuál de los trenes (fe)(fm) es (fe) intercity o talgo .

M7: hay tres intercity . el primero sale a las siete horas y quince minutos de la mañana y el último a las veintitrés horas . ¿ desea más información ?

U7: (na) sí . (dq por favor) . (df mire) . quisiera saber e(a l) primer tren que me ha dado de Valencia a Madr(a i)d . (na) me ha dicho que

- llega a las diez cuarenta y ocho . (na) me interesaría saber si hay otro tren que llegue un poco antes .
- M8: antes de las diez horas y cuarenta y ocho minutos no llega ningún tren a Madrid .
- U8: (nl)(na) (da muy bien) . (da gracias) (na) .
- M9: gracias por utilizar este servicio . por favor , diga el código del escenario que ha utilizado .
- U9: (nl)(na) no . (de es que) aún no he acabado . (p) (b (nw tenía más)) (fe)(fm) (nw quería hacerle) otra pregunta (nw) ¿ es posible ?
- M10: ¿ qué desea ?
- U10: (na) (df mire) . quisiera sab(a e)r (rs (m e(a l)) (c las)) tarifas (na)(fa)(fm) e(a n) (fe) los trenes de tipo intercity y talgo (na) los días laborables y luego los días festivos . (na) quiero saber el que sale más barato .
- M11: el precio del viaje de ida en segunda clase en un intercity es de cuatro mil trescientas pesetas , y de seis mil novecientas pesetas el de ida y vuelta . ¿ desea realizar otra consulta ?
- U11: (na) sí . quisiera que me dijese el precio en el talgo .
- M12: no hay ningún tren talgo a Madrid . ¿ desea otro tipo de tren ?
- U12: (nl)(na) no . (fb) (b lo único qu(a e)) (df entonces) ¿ me puede decir s(a i) l(a a) tarifa es la misma (na)(fe) durante todos los días de la semana ? ¿ (na) o esta tarifa cambia los fines de semana , por ejemplo ?
- M13: lo siento . no he entendido su pregunta .
- U13: (nl)(na) (df mire) . quisiera saber el importe que me ha dado de cuatro mil trescientas pesetas (na) el trayecto de ida , (na) y luego seis mil novecientas para ida y vuelta (na)(fa)(fm) ¿ este precio es válido para todos los días de la semana de lunes a domingo ? (p) esto es lo que quiero saber .
- M14: ¿ desea viajar el domingo tres de octubre ?
- U14: (na)(fa) s(a í) . (p) por ejemplo , podría ser . sí .
- M15: le consulto precios de Valencia a Madrid en intercity y segunda clase . un momento , por favor .
- U15: (nw) (da gracias) .
- M16: el precio del viaje de ida en segunda clase en un intercity es de cuatro mil cuatrocientas pesetas , y de siete mil pesetas el de ida y vuelta . ¿ desea realizar otra consulta ?
- U16: (na) no . (dc muchas gracias) . (df entonces) ya veo qu(a e) viajar en domingo es más caro que entre semana . hay poca diferencia pero (nw bueno) . (na) (da muy bien) . (dc muchas (nw gracias)) .
- M17: gracias por utilizar este servicio . feliz viaje .

B.1.2. Versión en formato XML.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<!DOCTYPE
  anotacion
  SYSTEM
  "http://sinatra.we.lc.ehu.es:8080/~luisja/resources/anotacion.dtd"
>

<anotacion
  anotador="Amparo_Varona"
  fecha="13 June 2003 15:58:19"
```

```

>

<descripcion
  base_de_datos="INFOTREN1"
  lengua="español"
  tipo_habla="espontánea"
  registro="coloquial"
  tipo_discurso="tarea"
  frecuencia_kHz="8"
  codificacion="lineal"
  bits_muestra="16"
  lugar="oficina"
  canal="teléfono"
  nivel_ruido="bajo"
  fichero_audio=""
/>

<hablante id="sistema"/>

<hablante
  id="usuario"
  nombre="_9"
  sexo="mujer"
  edad="adulto"
/>

<!-- Turno 0 sistema -->
<turno id="M_000" hablante="sistema" fichero_audio_ss="">
<t>
  bienvenido al sistema automático de información de trenes
  regionales y de largo recorrido . por favor , utilice frases
  cortas . ¿ qué desea ?
</t>
</turno>

<!-- Turno 0 usuario -->
<turno id="U_000" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U0">
<t>
  <n fuente="hablante" tipo="aspiración"/>
  <d tipo="rellenar"> mire </d> . quisiera sab<a>e</a>r
  <n fuente="exterior" tipo="genérico"/>
  <f tipo="a"/> <f tipo="m"/> el horario de trenes que hay
  <n fuente="hablante" tipo="aspiración"/>
  <n fuente="exterior" tipo="genérico"/> para ir de Valencia a
  <l tipo="mal_pronunciada" palabra="Madrid"> Madrit </l> .
  <n fuente="hablante" tipo="aspiración"/> <f tipo="a"/> me gustaría que
  <r tipo="repetición"> <m> l<a>a</a> </m> <s> <f tipo="m"/> </s>
  <c> la </c> </r> llegada a Madrid <r tipo="inserción">
  <m> <l tipo="cortada" palabra="sobre"> so </l> </m>
  <c> <r tipo="sustitución"> <m> fuese </m> <c> fuera </c> </r>
  sobre </c> </r> las diez de la mañana .
</t>
</turno>

<!-- Turno 1 sistema -->

```

```

<turno id="M_001" hablante="sistema" fichero_audio_ss="">
<t>
    desde Valencia con destino a Madrid . ¿ desea viajar hoy mismo ?
</t>
</turno>

<!-- Turno 1 usuario -->
<turno id="U_001" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U1">
<t>
    <n fuente="hablante" tipo="aspiración"/> <f tipo="a"/> <f tipo="m"/>
    <n fuente="hablante" tipo="aspiración"/> no . hoy no .
    <n fuente="hablante" tipo="aspiración"/> todavía no sé exactamente
    <a>e</a>l día <r tipo="repetición"> <m> qu<a>e</a> </m> <c> que </c>
    </r> vamos a ir <p/> ¿ tengo que decirle un día exacto ?
</t>
</turno>

<!-- Turno 2 sistema -->
<turno id="M_002" hablante="sistema" fichero_audio_ss="">
<t>
    lo siento , creo que no le he entendido bien . ¿ qué día desea viajar ?
</t>
</turno>

<!-- Turno 2 usuario -->
<turno id="U_002" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U2">
<t>
    <n fuente="hablante" tipo="labios"/>
    <n fuente="hablante" tipo="aspiración"/> <d tipo="rellenar"> mire </d> .
    <d tipo="rellenar"> pues </d> mañana .
</t>
</turno>

<!-- Turno 3 sistema -->
<turno id="M_003" hablante="sistema" fichero_audio_ss="">
<t>
    le consulto horarios de Valencia a Madrid , el jueves treinta de
    octubre . un momento , por favor .
</t>
</turno>

<!-- Turno 3 usuario -->
<turno id="U_003" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U3">

</turno>

<!-- Turno 4 sistema -->
<turno id="M_004" hablante="sistema" fichero_audio_ss="">
<t>
    por la mañana hay tres trenes . el primero llega a las diez horas y
    cuarenta y ocho minutos de la mañana , el segundo a las doce horas y
    cuarenta y ocho minutos de la mañana y el tercero a las trece horas y
    cincuenta y ocho minutos . ¿ algo más ?

```

```

</t>
</turno>

<!-- Turno 4 usuario -->
<turno id="U_004" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U4">
<t>
  <n fuente="hablante" tipo="labios"/>
  <n fuente="hablante" tipo="aspiración"/> <f tipo="a"/> <f tipo="m"/>
  <p/> <n fuente="hablante" tipo="aspiración"/>
  <d tipo="rellenar"> mire </d> . quisiera volver el mismo día ,
  <d tipo="editar"> o sea </d> , mañana por la tarde ,
  <n fuente="hablante" tipo="aspiración"/> a última hora .
  <n fuente="hablante" tipo="aspiración"/>
  <d tipo="rellenar"> entonces </d> ¿ puede decirme los trenes que hay
  <n fuente="hablante" tipo="aspiración"/> <r tipo="sustitución">
  <m> par<a>a</a> ir </m> <c> para volver </c> </r> de Madrid a Valencia ?
</t>
</turno>

<!-- Turno 5 sistema -->
<turno id="M_005" hablante="sistema" fichero_audio_ss="">
<t>
  le consulto horarios de Madrid a Valencia , el jueves treinta de
  octubre . un momento , por favor .
</t>
</turno>

<!-- Turno 5 usuario -->
<turno id="U_005" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U5">

</turno>

<!-- Turno 6 sistema -->
<turno id="M_006" hablante="sistema" fichero_audio_ss="">
<t>
  ese día hay diez trenes . el primero sale a siete de la mañana y el
  último a las veintitrés horas . ¿ algo más ?
</t>
</turno>

<!-- Turno 6 usuario -->
<turno id="U_006" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U6">
<t>
  <n fuente="hablante" tipo="labios"/>
  <n fuente="hablante" tipo="aspiración"/> sí .
  <d tipo="rellenar"> mire </d> . me interesaría sab<a>e</a>r
  <n fuente="hablante" tipo="aspiración"/> <f tipo="e"/> <p/> cuál de los
  trenes <f tipo="e"/> <f tipo="m"/> es <f tipo="e"/> intercity o talgo .
</t>
</turno>

<!-- Turno 7 sistema -->
<turno id="M_007" hablante="sistema" fichero_audio_ss="">

```

```
<t>
    hay tres intercity . el primero sale a las siete horas y quince minutos
    de la mañana y el último a las veintitrés horas . ¿ desea más
    información ?
```

```
</t>
</turno>
```

```
<!-- Turno 7 usuario -->
```

```
<turno id="U_007" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U7">
<t>
```

```
    <n fuente="hablante" tipo="aspiración"/> sí . <d tipo="pedir">
    por favor </d> . <d tipo="rellenar"> mire </d> . quisiera saber
    e<a>l</a> primer tren que me ha dado de Valencia a Madr<a>i</a>d .
    <n fuente="hablante" tipo="aspiración"/> me ha dicho que llega a las
    diez cuarenta y ocho . <n fuente="hablante" tipo="aspiración"/> me
    interesaría saber si hay otro tren que llegue un poco antes .
```

```
</t>
</turno>
```

```
<!-- Turno 8 sistema -->
```

```
<turno id="M_008" hablante="sistema" fichero_audio_ss="">
<t>
```

```
    antes de las diez horas y cuarenta y ocho minutos no llega ningún tren
    a Madrid .
```

```
</t>
</turno>
```

```
<!-- Turno 8 usuario -->
```

```
<turno id="U_008" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U8">
<t>
```

```
    <n fuente="hablante" tipo="labios"/>
    <n fuente="hablante" tipo="aspiración"/>
    <d tipo="aceptar"> muy bien </d> . <d tipo="aceptar"> gracias </d>
    <n fuente="hablante" tipo="aspiración"/> .
```

```
</t>
</turno>
```

```
<!-- Turno 9 sistema -->
```

```
<turno id="M_009" hablante="sistema" fichero_audio_ss="">
<t>
```

```
    gracias por utilizar este servicio . por favor , diga el código del
    escenario que ha utilizado .
```

```
</t>
</turno>
```

```
<!-- Turno 9 usuario -->
```

```
<turno id="U_009" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U9">
<t>
```

```
    <n fuente="hablante" tipo="labios"/>
    <n fuente="hablante" tipo="aspiración"/> no .
    <d tipo="editar"> es que </d> aún no he acabado . <p/> <b> <N>
    tenía más </N> </b> <f tipo="e"/> <f tipo="m"/> <N> quería hacerle </N>
    otra pregunta <n fuente="exterior" tipo="genérico"/> ¿ es posible ?
```

```

</t>
</turno>

<!-- Turno 10 sistema -->
<turno id="M_010" hablante="sistema" fichero_audio_ss="">
<t>
    ¿ qué desea ?
</t>
</turno>

<!-- Turno 10 usuario -->
<turno id="U_010" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U10">
<t>
    <n fuente="hablante" tipo="aspiración"/> <d tipo="rellenar"> mire </d> .
    quisiera sab<a>e</a>r <r tipo="sustitución"> <m> e<a>l</a> </m> <c> las
    </c> </r> tarifas <n fuente="hablante" tipo="aspiración"/> <f tipo="a"/>
    <f tipo="m"/> e<a>n</a> <f tipo="e"/> los trenes de tipo intercity y
    talgo <n fuente="hablante" tipo="aspiración"/> los días laborables y
    luego los días festivos . <n fuente="hablante" tipo="aspiración"/>
    quiero saber el que sale más barato .
</t>
</turno>

<!-- Turno 11 sistema -->
<turno id="M_011" hablante="sistema" fichero_audio_ss="">
<t>
    el precio del viaje de ida en segunda clase en un intercity es de
    cuatro mil trescientas pesetas , y de seis mil novecientas pesetas
    el de ida y vuelta . ¿ desea realizar otra consulta ?
</t>
</turno>

<!-- Turno 11 usuario -->
<turno id="U_011" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U11">
<t>
    <n fuente="hablante" tipo="aspiración"/> sí . quisiera que me dijese el
    precio en el talgo .
</t>
</turno>

<!-- Turno 12 sistema -->
<turno id="M_012" hablante="sistema" fichero_audio_ss="">
<t>
    no hay ningún tren talgo a Madrid . ¿ desea otro tipo de tren ?
</t>
</turno>

<!-- Turno 12 usuario -->
<turno id="U_012" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U12">
<t>
    <n fuente="hablante" tipo="labios"/>
    <n fuente="hablante" tipo="aspiración"/> no .
    <f tipo="sin_identificar"/> <b> lo único qu<a>e</a> </b>

```

```

<d tipo="rellenar"> entonces </d> ¿ me puede decir s<a>i</a> l<a>a</a>
tarifa es la misma <n fuente="hablante" tipo="aspiración"/>
<f tipo="e"/> durante todos los días de la semana ?
¿ <n fuente="hablante" tipo="aspiración"/> o esta tarifa cambia los
fines de semana , por ejemplo ?
</t>
</turno>

<!-- Turno 13 sistema -->
<turno id="M_013" hablante="sistema" fichero_audio_ss="">
<t>
    lo siento . no he entendido su pregunta .
</t>
</turno>

<!-- Turno 13 usuario -->
<turno id="U_013" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U13">
<t>
    <n fuente="hablante" tipo="labios"/>
    <n fuente="hablante" tipo="aspiración"/> <d tipo="rellenar"> mire </d> .
    quisiera saber el importe que me ha dado de cuatro mil trescientas
    pesetas <n fuente="hablante" tipo="aspiración"/> el trayecto de ida ,
    <n fuente="hablante" tipo="aspiración"/> y luego seis mil novecientas
    para ida y vuelta <n fuente="hablante" tipo="aspiración"/> <f tipo="a"/>
    <f tipo="m"/> ¿ este precio es válido para todos los días de la semana
    de lunes a domingo ? <p/> esto es lo que quiero saber .
</t>
</turno>

<!-- Turno 14 sistema -->
<turno id="M_014" hablante="sistema" fichero_audio_ss="">
<t>
    ¿ desea viajar el domingo tres de octubre ?
</t>
</turno>

<!-- Turno 14 usuario -->
<turno id="U_014" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U14">
<t>
    <n fuente="hablante" tipo="aspiración"/> <f tipo="a"/> s<a>í</a> .
    <p/> por ejemplo , podría ser . sí .
</t>
</turno>

<!-- Turno 15 sistema -->
<turno id="M_015" hablante="sistema" fichero_audio_ss="">
<t>
    le consulto precios de Valencia a Madrid en intercity y segunda clase .
    un momento , por favor .
</t>
</turno>

<!-- Turno 15 usuario -->
<turno id="U_015" hablante="usuario"

```

```

fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U15">
<t>
  <n fuente="exterior" tipo="genérico"/> <d tipo="aceptar"> gracias </d> .
</t>
</turno>

<!-- Turno 16 sistema -->
<turno id="M_016" hablante="sistema" fichero_audio_ss="">
<t>
  el precio del viaje de ida en segunda clase en un intercity es de
  cuatro mil cuatrocientas pesetas , y de siete mil pesetas el de ida y
  vuelta . ¿ desea realizar otra consulta ?
</t>
</turno>

<!-- Turno 16 usuario -->
<turno id="U_016" hablante="usuario"
fichero_audio_ss="/home/grah/bd/castellano/oz1/audio/c730/c730.U16">
<t>
  <n fuente="hablante" tipo="aspiración"/> no .
  <d tipo="cerrar"> muchas gracias </d> .
  <d tipo="rellenar"> entonces </d> ya veo qu<a>e</a> viajar en domingo
  es más caro que entre semana . hay poca diferencia pero <N> bueno </N> .
  <n fuente="hablante" tipo="aspiración"/>
  <d tipo="aceptar"> muy bien </d> .
  <d tipo="cerrar"> muchas <N> gracias </N> </d> .
</t>
</turno>

<!-- Turno 17 sistema -->
<turno id="M_017" hablante="sistema" fichero_audio_ss="">
<t>
  gracias por utilizar este servicio . feliz viaje .
</t>
</turno>

</anotacion>

```

B.2. Conversación entre personas.

B.2.1. Versión en formato simplificado.

H1[1]: Muchas graci(a a)s Benign(a o) Santiño . Vamos a entrar ya de lleno en los minutos destinados a la información deportiva cuand(a o) restan (na) (p) veintisiete minutos para las dos de la tarde . Vamos a entrar (p) de lleno en el capítulo d(a e) Protagonistas y para ello tenemos (na) como invitado especial en el estudio central de Radio Segovia (na) al que es ahora capitán de la Sociedad Deportiva Gimnástica Segoviana . (na) Carmelo del Pozo , buenas tardes .

H2[2]: (fe) Buenas tardes .

H1[3]: (na) También estuv(a o) el señ(a o)r Blanco Gutiérrez , el colegiado burgalés qu(a e) (na) , que actuó el pasad(a o) (p) domingo e(a n) tierras palentinas , en Venta de Baños .

H2[4]: Pues no . A mi modo de entender no estuv(a o) ni siquiera bien ,

- estuvo bastante mal aunqu(a e) según parec(a e) me habéis (lm comentado comentao) , las crónicas le ponen bien .
- H1[5]: (na) Sí . Le ponen buena actuación . Buena actuación a lo mejor fue para los de Venta de Baños ¿ no ?
- H2[6]: Sí , claro , s(a i) piensas qu(a e) (p) , coge , expulsa a un jugador del equipo co(a n) , visitante a los treinta minutos por dos entrada(a s) (na) que son las dos primeras que hace el jugador (na) (fm) , sus (lm jueces jue(a e)) d(a e) (fb) , sus jueces de línea no le ayudan bastante y comet(a e)n bastantes fueras de juego que no se señalan y a parte pue(a s) (na) , en el último minuto tiene que expulsar al portero por la nueva norma (a y) , no le expulsa , le saca simplemente tarjeta amarilla porque se queda con un jugador menos (p) y no pueden hacer cambios (na) , pues hombre , la actuación se puede calificar par(a a) el equipo local de buena .
- H1[7]: L(a a) , dos doble tarjeta para Hoyos (fe) ¿ bien enseñadas ?
- H2[8]: (na) Hombr(a e) , (fm) es que estamos en lo de siempre so(a n) do(a s) entradas fuertes (p) que sí que puede señalarse tarjeta pero si son las dos primeras entradas de un jugador , (nw) pues seguramente si fuera del equipo local a lo mejor no le hubiera (lu enseñado enseña) (p) (lm enseñado enseño) las dos , (p) en las dos entradas quiero decir , (fe) tarjeta amarilla .
- H1[9]: ¿ Es cierto qu(a e) el colegiado señor Blanco Gutiérrez os había amenazado ya en pretemporada de que os iba a expulsar a algún jugador ?
- H2[10]: (nl) (na) Hombre , es que fue curioso cuando e(a l) domingo anterior el partido amistoso en Aranda contra el Arandina (na) nos dice , hombre , probabement(a e) n(a o) (fb) sea yo el que os arbitra (p) cuand(a o) siendo el partido el domingo , hasta el lunes o el martes ellos no saben la designación d(a e) (nw) , qué (lu equipo equi) de qué (lu equipo equip) (a o) de qué partido le toc(a a) arbitrar . (na) Entonces a nosotros nos chocó un poco , sí que en ese partido estuvo bastant(a e) amenazante (na) y hombre , pue(a s) luego en el partid(a o) ya de liga , en el primero , el del pasado domingo pues sí que estuvo bastant(a e) (p) (fm) en mala situación con nosotros .
- H1[11]: Hay una jugada un tanto conflictiv(a a) a tenor de la nueva reglamentación para los porteros y es que s(a e) sitúa un jugador , creo que era Montes ¿ no ? (na) (nw) (fe) ante Zapatero (p) , no sé quién era (o exactamente) .
- H2[12]: (o Alvaro) , (o era Alvaro) .
- H1[13]: (o Alvaro) (na) (fe) , se sitúa ant(a e) la posición del portero , (fb) se va en inmejorable posición par(a a) , para lanzar (na) y le entra en clara falta (p) dentro del área (na) (a e)l guardameta Zapatero . (p) Eso según la nueva reglamentación (na) es (fe) expulsión d(a e) (na) , del propio Zapatero y penalty (fa) favorable a la Sociedad Deportiva Gimnástica Segoviana .
- H2[14]: Sí , además es qu(a e) (na) no ofrece dudas de(a l) (fb) , desde el momento en que se encuentra (lu casi cas) en el área pequeña , o se(a a) (na) , s(a e) le encara directamente en un pas(a e) , en un cambio de juego que hacemo(a s) y que le llega a Alvaro perfectamente (na) , le encara el portero , el porter(a o) le cierra un ángulo , él busca otro regateándole (na) y al regatearle y el portero ver que no llega se tira sobre Alvaro y l(a e) , y le hace un placaje claro . (na) E(a l) , lo que pasa que esta norma es sumamente subjetiva , (p) (fb) luego posteriormente el árbitro com(a o) , me llamó para hablar con él , bueno , nos llamó a varios jugadores de nuestro equipo (na) para intentar explicárnoslo pero es una norma tan sumamente subjetiva que va a crear muchos problemas .

H1[15]: Esto me imagino yo qu(a e) puede ser motivo , la expulsión de Adolfo Redondo que se enfadó ya (fm) por aquello de la expulsión de Hoyos (na) en los minutos inicial del partido cuando corría el minuto treinta y cinco aproximadamente , (na) luego es(a a) falta clarísima d(a e) (p) , de Zapatero a Alvaro (na) y naturalmente pues (fe) (fe) (lu influiría in) (na) influiría en el ánimo que tenía Adolfo Redondo en esos momentos (na) (fe) con un resultado todavía adverso (fb) .

H2[16]: Sí , bueno , la expulsión de Adolfo vino un poco anterior que también fue por una jugada de esta nueva norma . (nw) Fue una jugada en la qu(a e) hicimos u(a n) pase en profundidad (p) o bueno , la cogió el balón Tomás , dio un pase largo , se iba a plantar solo delante del portero (na) y el defens(a a) no llegando al balón se lanzó a por Tomás y le hizo una falta clara . (na) El árbitro para evitarse complicaciones (p) no pitó la falta y así no tenía que expulsar al (lu jugador jua) , al defens(a a) local . (na) Entonces Adolfo Redondo salió de la caseta y al salir de la caset(a a) (na) (a y) al acabar la jugada qu(a e) el árbitro no había parado pues se acercó al línea , l(a e) , le coment(a ó) el línea que le habí(a a) , que había salido de la caseta , le preguntó si era tarjeta amarilla o tarjeta roja y (fb) sacó tarjeta roja (a y) (fb) (p) , y vamos , le expulsó .

H1[17]: Entonces (lm juzgando jugando) la actuación del señor (fe) Blanco , el señ(a o)r , (p) (fb) el colegiado burgalés , señor Blanco (na) l(a a) (p) , no victoria de la Sociedad Deportiva Gimnástica Segoviana ¿ se la puede achacar (a a) (p) , al colegiado o no ?

H2[18]: No . Desde lueg(a o) , que tuvo una mala actuación per(a o) (na) , aunque no(a s) complic(a ó) y nos estrope(a ó) (fb) el buen sistema y la buena situación que teníamos en los primeros minutos (na) (lm pero puer(a o)) desde luego que ya después con un jugador menos sí qu(a e) (na) , fuimos nosotros los que tuvimos que echarnos para atrás y aguantar el vendaval qu(a e) se nos venía encima con el Venta de Baños .

H1[19]: En el plano particular , Carmelo del Pozo está satisfecho de cóm(a o) (p) se discurrían los primeros minutos de partido .

H2[20]: (na) Hombre es que s(a i) (a a)l principio de partid(a o) , en los primeros diez minutos estamos muy bien (lm situados situaos) , les cortamos , les presionamos muy bien en el centro del campo , cortamos balones y disponemos de dos o tres claras ocasiones (na) que no materializamos que ahí fue la pena porque si a los diez o quince minutos primeros ya vamos con un gol o dos a favor (na) pues el partido hubiera tenido un giro totalment(a e) , vamos , una situación totalmente distinta . (na) Entonces n(a o) (lm logramos labramos) materializar esas situaciones (na) luego sí alcanzamos el primer gol per(a o) , nos expulsaron a un jugador y y(a a) trastocó todo . Si además en el último minuto de la primera parte nos meten (a e)l gol del empate (na) pues hombr(a e) , ya vas al vestuario de una forma distinta que si hubieras ido con un cero dos .

H1[21]: (na) ¿ Va a ser difícil puntuar para los equipos que visiten el campo de Venta (nw de Baños) ?

H2[22]: (na) Yo creo que va a ser bastante difícil porqu(a e) es un equipo muy peleó(a n) , sab(a e) , vamos , tiene un entrenador qu(a e) ya conoce como es la tercera división y lo que verdaderamente influy(a e) (na) o lo que vale en esta tercera división que son los puntos de casa , (na) entonces ellos cuentan co(a n) un equip(a o) ya te he dicho , muy peleón y qu(a e) buscan (fb) al máximo lo(a s) (na) , los puntos de casa . S(a i) (na) , si consigues mantener esos primeros (p) partidos o esos primer(a o)s minutos d(a e) los partidos , (na) pues entonces sí

que mantendrá(a n) en su casa una buena disposición para ganar (o los partidos) .

H1[23]: (o Puede) engañar el hecho de que haya sido un equipo recién ascendido a la tercera división ¿ no ?

H2[24]: Sí , engaña porqu(a e) si te das cuenta tiene , vamo(a s) (p) (fm) tien(a e) el mismo esquema que el año pasado cuando a falta de ocho partidos ya estaba en tercera división (na) . Y ha unido a ello el fichar a un portero porque el suyo se había (lm marchado marchao) a Cantabria (p) como es Zapatero que era un portero que estaba en la Herrera (na) qu(a e) vamos , algunos que ya estamos algunos en tercera división le conocíamos y sabíamos que era un buen portero .

H1[25]: Y además (fe) importantes refuerzos en la delantera como es el caso del ex gimnástico Catón (na) y hombres qu(a e) ya llevan veteran(a o) (p) , o son veteranos en (na) , en el grupo octavo .

H2[26]: Sí . Tienen (p) dos o tres hombres muy peleones arriba (a y) es muy importante tener ahí en tercera divisió(a n) en estos partidos (na) gente que pelea y no da un balón por perdido pue(a s) como es el caso de Catón , qu(a e) (na) , que si además e(a n) , s(a e) vamos (na) jugara contra su (lm ex es) equipo pues mucho más . S(a e) (na) , se multiplicó las acciones (a y) (fb) , y enseguida quería meter uno o dos goles para (na) demostrar a su (lm ex es) equipo que estaba confundido con él .

H1[27]: (na) Yo prefiero que Carmelo no tenga que actuar mucho en los partidos (p) per(a o) te han vuelto a destacar , es decir (na) ¿ Carmelo vuelve a sus fueros (fm) como destacado portero en el grupo octavo ?

H2[28]: (na) No , no se trata de que vuelvas a tus fueros , simplemente se trata de qu(a e) (na) en estos partidos de toma y daca tienes muchas (lu opciones of) , bueno , tienes que trabajar más de lo normal (na) y si además pue(a s) (fb) , ocurre que nos meten el dos uno (a y) simplemente no hay otra solución que , y que da igual perder por dos uno que por tres uno (na) pues lanzas el equipo al centro del camp(a o) (a y) (na) , y hombre , las situaciones de peligro (na) (a n)o es que sean más sino que son mucho más claras porque te enfrentas directamente contra los delanteros y hay que tener además (p) con estas nuevas normas hay que tener mucho (lm cuidado cuidao) porque a la mínim(a a) vez que hagas algo pues te (na) , y si es más fuera de casa (na) van a mirar mucho más al portero y enseguida vas a esta(a r) (p) bueno , le van a expulsar , vamos .

B.2.2. Versión en formato XML.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<!DOCTYPE
  anotacion
  SYSTEM
  "http://sinatra.we.lc.ehu.es:8080/~luisja/resources/anotacion.dtd"
>

<anotacion
  anotador="Luis_Javier_Rodríguez"
  fecha="13 June 2003 16:03:07"
>

<descripcion
  base_de_datos="CORLEC_EHU_1"
  lengua="español"
```

```

        tipo_habla="espontánea"
        registro="coloquial"
        tipo_discurso="entrevista"
        frecuencia_kHz="16"
        codificacion="lineal"
        bits_muestra="16"
        lugar="exterior"
        canal="micrófono"
        nivel_ruido="variable"
        fichero_audio=
            "/home/grah/bd/castellano/corpus_oral_UAM/audio/ent/cent001b.esps"
/>

<hablante
    id="H1"
    sexo="hombre"
    edad="adulto"
/>

<hablante
    id="H2"
    sexo="hombre"
    edad="adulto"
/>

<!-- Turno 1 -->
<turno id="_001" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.001.H1">
<t>
    Muchas graci<a>a</a>s Benign<a>o</a> Santiago . Vamos a entrar ya de
    lleno en los minutos destinados a la información deportiva cuand<a>o</a>
    restan <n fuente="hablante" tipo="aspiración"/> <p/> veintisiete minutos
    para las dos de la tarde . Vamos a entrar <p/> de lleno en el capítulo
    d<a>e</a> Protagonistas y para ello tenemos
    <n fuente="hablante" tipo="aspiración"/> como invitado especial en el
    estudio central de Radio Segovia <n fuente="hablante" tipo="aspiración"/>
    al que es ahora capitán de la Sociedad Deportiva Gimnástica Segoviana .
    <n fuente="hablante" tipo="aspiración"/> Carmelo del Pozo , buenas tardes .
</t>
</turno>

<!-- Turno 2 -->
<turno id="_002" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.002.H2">
<t>
    <f tipo="e"/> Buenas tardes .
</t>
</turno>

<!-- Turno 3 -->
<turno id="_003" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.003.H1">
<t>

```

```

<n fuente="hablante" tipo="aspiración"/> También estuv<a>o</a> el
señ<a>o</a>r Blanco Gutiérrez , el colegiado burgalés qu<a>e</a>
<n fuente="hablante" tipo="aspiración"/> , que actuó el pasad<a>o</a>
<p/> domingo e<a>n</a> tierras palentinas , en Venta de Baños .
</t>
</turno>

<!-- Turno 4 -->
<turno id="_004" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.004.H2">
<t>
    Pues no . A mi modo de entender no estuv<a>o</a> ni siquiera bien ,
    estuvo bastante mal aunqu<a>e</a> según parec<a>e</a> me habéis
    <l tipo="mal_pronunciada" palabra="comentado"> comentao </l> , las
    crónicas le ponen bien .
</t>
</turno>

<!-- Turno 5 -->
<turno id="_005" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.005.H1">
<t>
    <n fuente="hablante" tipo="aspiración"/> Sí . Le ponen buena actuación .
    Buena actuación a lo mejor fue para los de Venta de Baños ¿ no ?
</t>
</turno>

<!-- Turno 6 -->
<turno id="_006" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.006.H2">
<t>
    Sí , claro , s<a>i</a> piensas qu<a>e</a> <p/> , coge , expulsa a un
    jugador del equipo co<a>n</a> , visitante a los treinta minutos por dos
    entrada<a>s</a> <n fuente="hablante" tipo="aspiración"/> que son las dos
    primeras que hace el jugador <n fuente="hablante" tipo="aspiración"/>
    <f tipo="m"/> , sus <l tipo="mal_pronunciada" palabra="jueces">
    jue<a>e</a> </l> d<a>e</a> <f tipo="sin_identificar"/> , sus jueces de
    línea no le ayudan bastante y comet<a>e</a> n bastantes fueras de juego
    que no se señalan y a parte pue<a>s</a>
    <n fuente="hablante" tipo="aspiración"/> , en el último minuto tiene que
    expulsar al portero por la nueva norma <a>y</a> , no le expulsa , le
    saca simplemente tarjeta amarilla porque se queda con un jugador menos
    <p/> y no pueden hacer cambios <n fuente="hablante" tipo="aspiración"/> ,
    pues hombre , la actuación se puede calificar par<a>a</a> el equipo
    local de buena .
</t>
</turno>

<!-- Turno 7 -->
<turno id="_007" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.007.H1">
<t>

```

```

L<a>a</a> , dos doble tarjeta para Hoyos <f tipo="e"/>
¿ bien enseñadas ?
</t>
</turno>

<!-- Turno 8 -->
<turno id="_008" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.008.H2">
<t>
  <n fuente="hablante" tipo="aspiración"/> Hombr<a>e</a> , <f tipo="m"/>
  es que estamos en lo de siempre so<a>n</a> do<a>s</a> entradas fuertes
  <p/> que sí que puede señalarse tarjeta pero si son las dos primeras
  entradas de un jugador , <n fuente="exterior" tipo="genérico"/> pues
  seguramente si fuera del equipo local a lo mejor no le hubiera
  <l tipo="cortada" palabra="enseñado"> enseña </l> <p/>
  <l tipo="mal_pronunciada" palabra="enseñado"> enseñao </l> las dos ,
  <p/> en las dos entradas quiero decir , <f tipo="e"/> tarjeta amarilla .
</t>
</turno>

<!-- Turno 9 -->
<turno id="_009" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.009.H1">
<t>
  ¿ Es cierto qu<a>e</a> el colegiado señor Blanco Gutiérrez os había
  amenazado ya en pretemporada de que os iba a expulsar a algún jugador ?
</t>
</turno>

<!-- Turno 10 -->
<turno id="_010" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.010.H2">
<t>
  <n fuente="hablante" tipo="labios"/>
  <n fuente="hablante" tipo="aspiración"/> Hombre , es que fue curioso
  cuando e<a>l</a> domingo anterior el partido amistoso en Aranda contra
  el Arandina <n fuente="hablante" tipo="aspiración"/> nos dice , hombre ,
  probablemente<a>e</a> n<a>o</a> <f tipo="sin_identificar"/> sea yo el que
  os arbitra <p/> cuand<a>o</a> siendo el partido el domingo , hasta el
  lunes o el martes ellos no saben la designación d<a>e</a>
  <n fuente="exterior" tipo="genérico"/> , qué
  <l tipo="cortada" palabra="equipo"> equi </l> de qué
  <l tipo="cortada" palabra="equipo"> equip </l> <a>o</a> de qué partido
  le toc<a>a</a> arbitrar . <n fuente="hablante" tipo="aspiración"/>
  Entonces a nosotros nos chocó un poco , sí que en ese partido estuvo
  bastant<a>e</a> amenazante <n fuente="hablante" tipo="aspiración"/> y
  hombre , pue<a>s</a> luego en el partid<a>o</a> ya de liga , en el
  primero , el del pasado domingo pues sí que estuvo bastant<a>e</a>
  <p/> <f tipo="m"/> en mala situación con nosotros .
</t>
</turno>

<!-- Turno 11 -->

```

```

<turno id="_011" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.011.H1">
<t>
    Hay una jugada un tanto conflictiv<a>a</a> a tenor de la nueva
    reglamentación para los porteros y es que s<a>e</a> sitúa un jugador ,
    creo que era Montes ¿ no ? <n fuente="hablante" tipo="aspiración"/>
    <n fuente="exterior" tipo="genérico"/> <f tipo="e"/> ante Zapatero
    <p/> , no sé quién era
</t>
<o>
    exactamente .
</o>
</turno>

<!-- Turno 12 -->
<turno id="_012" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.012.H2">
<o>
    Alvaro ,
</o>
<o>
    era Alvaro .
</o>
</turno>

<!-- Turno 13 -->
<turno id="_013" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.013.H1">
<o>
    Alvaro
</o>
<t>
    <n fuente="hablante" tipo="aspiración"/> <f tipo="e"/> , se sitúa
    ant<a>e</a> la posición del portero , <f tipo="sin_identificar"/> se va
    en inmejorable posición par<a>a</a> , para lanzar
    <n fuente="hablante" tipo="aspiración"/> y le entra en clara falta <p/>
    dentro del área <n fuente="hablante" tipo="aspiración"/> <a>e</a>l
    guardameta Zapatero . <p/> Eso según la nueva reglamentación
    <n fuente="hablante" tipo="aspiración"/> es <f tipo="e"/> expulsión
    d<a>e</a> <n fuente="hablante" tipo="aspiración"/> , del propio Zapatero
    y penalty <f tipo="a"/> favorable a la Sociedad Deportiva Gimnástica
    Segoviana .
</t>
</turno>

<!-- Turno 14 -->
<turno id="_014" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.014.H2">
<t>
    Sí , además es qu<a>e</a> <n fuente="hablante" tipo="aspiración"/> no
    ofrece dudas de<a>l</a> <f tipo="sin_identificar"/> , desde el momento
    en que se encuentra <l tipo="cortada" palabra="casi"> cas </l> en el

```

área pequeña , o se<a>a <n fuente="hablante" tipo="aspiración"/> ,
 s<a>e le encara directamente en un pas<a>e , en un cambio de
 juego que hacemos<a>s y que le llega a Alvaro perfectamente
 <n fuente="hablante" tipo="aspiración"/> , le encara el portero , el
 porter<a>o le cierra un ángulo , él busca otro regateándole
 <n fuente="hablante" tipo="aspiración"/> y al regatearle y el portero
 ver que no llega se tira sobre Alvaro y l<a>e , y le hace un placaje
 claro . <n fuente="hablante" tipo="aspiración"/> E<a>l , lo que pasa
 que esta norma es sumamente subjetiva , <p/> <f tipo="sin_identificar"/>
 luego posteriormente el árbitro com<a>o , me llamó para hablar con
 él , bueno , nos llamó a varios jugadores de nuestro equipo
 <n fuente="hablante" tipo="aspiración"/> para intentar explicárnoslo
 pero es una norma tan sumamente subjetiva que va a crear muchos
 problemas .

</t>

</turno>

<!-- Turno 15 -->

<turno id="_015" hablante="H1"
 fichero_audio_ss=
 "/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.015.H1">

<t>

Esto me imagino yo qu<a>e puede ser motivo , la expulsión de Adolfo
 Redondo que se enfadó ya <f tipo="m"/> por aquello de la expulsión de
 Hoyos <n fuente="hablante" tipo="aspiración"/> en los minutos inicial
 del partido cuando corría el minuto treinta y cinco aproximadamente ,
 <n fuente="hablante" tipo="aspiración"/> luego es<a>a falta
 clarísima d<a>e <p/> , de Zapatero a Alvaro
 <n fuente="hablante" tipo="aspiración"/> y naturalmente pues
 <f tipo="e"/> <f tipo="e"/>
 <l tipo="cortada" palabra="influiría"> in </l>
 <n fuente="hablante" tipo="aspiración"/> influiría en el ánimo que tenía
 Adolfo Redondo en esos momentos <n fuente="hablante" tipo="aspiración"/>
 <f tipo="e"/> con un resultado todavía adverso
 <f tipo="sin_identificar"/> .

</t>

</turno>

<!-- Turno 16 -->

<turno id="_016" hablante="H2"
 fichero_audio_ss=
 "/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.016.H2">

<t>

Sí , bueno , la expulsión de Adolfo vino un poco anterior que también
 fue por una jugada de esta nueva norma .
 <n fuente="exterior" tipo="genérico"/> Fue una jugada en la qu<a>e
 hicimos u<a>n pase en profundidad <p/> o bueno , la cogió el balón
 Tomás , dio un pase largo , se iba a plantar solo delante del portero
 <n fuente="hablante" tipo="aspiración"/> y el defens<a>a no llegando
 al balón se lanzó a por Tomás y le hizo una falta clara .
 <n fuente="hablante" tipo="aspiración"/> El árbitro para evitarse
 complicaciones <p/> no pitó la falta y así no tenía que expulsar al
 <l tipo="cortada" palabra="jugador"> jua </l> , al defens<a>a local .
 <n fuente="hablante" tipo="aspiración"/> Entonces Adolfo Redondo salió
 de la caseta y al salir de la caset<a>a
 <n fuente="hablante" tipo="aspiración"/> <a>y al acabar la jugada

```

    qu<a>e</a> el árbitro no había parado pues se acercó al línea ,
    l<a>e</a> , le coment<a>ó</a> el línea que le habí<a>a</a> , que había
    salido de la caseta , le preguntó si era tarjeta amarilla o tarjeta roja
    y <f tipo="sin_identificar"/> sacó tarjeta roja <a>y</a>
    <f tipo="sin_identificar"/> <p/> , y vamos , le expulsó .
</t>
</turno>

<!-- Turno 17 -->
<turno id="_017" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.017.H1">
<t>
    Entonces <l tipo="mal_pronunciada" palabra="jugando"> jugando </l>
    la actuación del señor <f tipo="e"/> Blanco , el señ<a>o</a>r , <p/>
    <f tipo="sin_identificar"/> el colegiado burgalés , señor Blanco
    <n fuente="hablante" tipo="aspiración"/> l<a>a</a> <p/> , no victoria de
    la Sociedad Deportiva Gimnástica Segoviana ¿ se la puede achacar
    <a>a</a> <p/> , al colegiado o no ?
</t>
</turno>

<!-- Turno 18 -->
<turno id="_018" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.018.H2">
<t>
    No . Desde lueg<a>o</a> , que tuvo una mala actuación per<a>o</a>
    <n fuente="hablante" tipo="aspiración"/> , aunque no<a>s</a>
    complic<a>ó</a> y nos estrope<a>ó</a> <f tipo="sin_identificar"/>
    el buen sistema y la buena situación que teníamos en los primeros
    minutos <n fuente="hablante" tipo="aspiración"/>
    <l tipo="mal_pronunciada" palabra="pero"> puer<a>o</a> </l> desde luego
    que ya después con un jugador menos sí qu<a>e</a>
    <n fuente="hablante" tipo="aspiración"/> , fuimos nosotros los que
    tuvimos que echarnos para atrás y aguantar el vendaval qu<a>e</a> se
    nos venía encima con el Venta de Baños .
</t>
</turno>

<!-- Turno 19 -->
<turno id="_019" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.019.H1">
<t>
    En el plano particular , Carmelo del Pozo está satisfecho de cóm<a>o</a>
    <p/> se discurrían los primeros minutos de partido .
</t>
</turno>

<!-- Turno 20 -->
<turno id="_020" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.020.H2">
<t>
    <n fuente="hablante" tipo="aspiración"/> Hombre es que s<a>i</a>

```

```

<a>a</a>l principio de partid<a>o</a> , en los primeros diez minutos
estamos muy bien
<l tipo="mal_pronunciada" palabra="situados"> situaos </l> , les
cortamos , les presionamos muy bien en el centro del campo , cortamos
balones y disponemos de dos o tres claras ocasiones
<n fuente="hablante" tipo="aspiración"/> que no materializamos que ahí
fue la pena porque si a los diez o quince minutos primeros ya vamos con
un gol o dos a favor <n fuente="hablante" tipo="aspiración"/> pues el
partido hubiera tenido un giro totalment<a>e</a> , vamos , una situación
totalmente distinta . <n fuente="hablante" tipo="aspiración"/> Entonces
n<a>o</a> <l tipo="mal_pronunciada" palabra="logramos"> labramos </l>
materializar esas situaciones <n fuente="hablante" tipo="aspiración"/>
luego sí alcanzamos el primer gol per<a>o</a> , nos expulsaron a un
jugador y y<a>a</a> trastocó todo . Si además en el último minuto de
la primera parte nos meten <a>e</a>l gol del empate
<n fuente="hablante" tipo="aspiración"/> pues hombr<a>e</a> , ya vas al
vestuario de una forma distinta que si hubieras ido con un cero dos .
</t>
</turno>

<!-- Turno 21 -->
<turno id="_021" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.021.H1">
<t>
    <n fuente="hablante" tipo="aspiración"/> ¿ Va a ser difícil puntuar
    para los equipos que visiten el campo de Venta <N> de Baños </N> ?
</t>
</turno>

<!-- Turno 22 -->
<turno id="_022" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.022.H2">
<t>
    <n fuente="hablante" tipo="aspiración"/> Yo creo que va a ser bastante
    difícil porqu<a>e</a> es un equipo muy peleó<a>n</a> , sab<a>e</a> ,
    vamos , tiene un entrenador qu<a>e</a> ya conoce como es la tercera
    división y lo que verdaderamente influy<a>e</a>
    <n fuente="hablante" tipo="aspiración"/> o lo que vale en esta tercera
    división que son los puntos de casa ,
    <n fuente="hablante" tipo="aspiración"/> entonces ellos cuentan
    co<a>n</a> un equip<a>o</a> ya te he dicho , muy peleón y qu<a>e</a>
    buscan <f tipo="sin_identificar"/> al máximo lo<a>s</a>
    <n fuente="hablante" tipo="aspiración"/> , los puntos de casa .
    S<a>i</a> <n fuente="hablante" tipo="aspiración"/> , si consigues
    mantener esos primeros <p/> partidos o esos primer<a>o</a>s minutos
    d<a>e</a> los partidos , <n fuente="hablante" tipo="aspiración"/> pues
    entonces sí que mantendrá<a>n</a> en su casa una buena disposición para
    ganar
</t>
<o>
    los partidos .
</o>
</turno>

```

```
<!-- Turno 23 -->
<turno id="_023" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.023.H1">
<o>
    Puede
</o>
<t>
    engañar el hecho de que haya sido un equipo recién ascendido a la
    tercera división ¿ no ?
</t>
</turno>
```

```
<!-- Turno 24 -->
<turno id="_024" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.024.H2">
<t>
    Sí , engaña porqu<a>e</a> si te das cuenta tiene , vamo<a>s</a> <p/>
    <f tipo="m"/> tien<a>e</a> el mismo esquema que el año pasado cuando a
    falta de ocho partidos ya estaba en tercera división
    <n fuente="hablante" tipo="aspiración"/> . Y ha unido a ello el fichar
    a un portero porque el suyo se había
    <l tipo="mal_pronunciada" palabra="marchado"> marchao </l> a Cantabria
    <p/> como es Zapatero que era un portero que estaba en la Herrera
    <n fuente="hablante" tipo="aspiración"/> qu<a>e</a> vamos , algunos que
    ya estamos algunos en tercera división le conocíamos y sabíamos que era
    un buen portero .
</t>
</turno>
```

```
<!-- Turno 25 -->
<turno id="_025" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.025.H1">
<t>
    Y además <f tipo="e"/> importantes refuerzos en la delantera como es el
    caso del ex gimnástico Catón <n fuente="hablante" tipo="aspiración"/>
    y hombres qu<a>e</a> ya llevan veteran<a>o</a> <p/> , o son veteranos
    en <n fuente="hablante" tipo="aspiración"/> , en el grupo octavo .
</t>
</turno>
```

```
<!-- Turno 26 -->
<turno id="_026" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.026.H2">
<t>
    Sí . Tienen <p/> dos o tres hombres muy peleones arriba <a>y</a> es muy
    importante tener ahí en tercera divisió<a>n</a> en estos partidos
    <n fuente="hablante" tipo="aspiración"/> gente que pelea y no da un
    balón por perdido pue<a>s</a> como es el caso de Catón , qu<a>e</a>
    <n fuente="hablante" tipo="aspiración"/> , que si además e<a>n</a> ,
    s<a>e</a> vamos <n fuente="hablante" tipo="aspiración"/> jugara contra
    su <l tipo="mal_pronunciada" palabra="ex"> es </l> equipo pues mucho más
    . S<a>e</a> <n fuente="hablante" tipo="aspiración"/> , se multiplicó las
```

acciones <a>y <f tipo="sin_identificar"/> , y enseguida quería meter uno o dos goles para <n fuente="hablante" tipo="aspiración"/> demostrar a su <l tipo="mal_pronunciada" palabra="ex"> es </l> equipo que estaba confundido con él .

</t>
</turno>

<!-- Turno 27 -->
<turno id="_027" hablante="H1"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.027.H1">
<t>
<n fuente="hablante" tipo="aspiración"/> Yo prefiero que Carmelo no tenga que actuar mucho en los partidos <p/> per<a>o te han vuelto a destacar , es decir <n fuente="hablante" tipo="aspiración"/>
¿ Carmelo vuelve a sus fueros <f tipo="m"/> como destacado portero en el grupo octavo ?

</t>
</turno>

<!-- Turno 28 -->
<turno id="_028" hablante="H2"
fichero_audio_ss=
"/home/grah/bd/castellano/corpus_oral_UAM/audio_EHU/ent/cent001b/cent001b.028.H2">
<t>
<n fuente="hablante" tipo="aspiración"/> No , no se trata de que vuelvas a tus fueros , simplemente se trata de qu<a>e
<n fuente="hablante" tipo="aspiración"/> en estos partidos de toma y daca tienes muchas <l tipo="cortada" palabra="opciones"> of </l> , bueno , tienes que trabajar más de lo normal
<n fuente="hablante" tipo="aspiración"/> y si además pue<a>s
<f tipo="sin_identificar"/> , ocurre que nos meten el dos uno <a>y simplemente no hay otra solución que , y que da igual perder por dos uno que por tres uno <n fuente="hablante" tipo="aspiración"/> pues lanzas el equipo al centro del camp<a>o <a>y
<n fuente="hablante" tipo="aspiración"/> , y hombre , las situaciones de peligro <n fuente="hablante" tipo="aspiración"/> <a>n es que sean más sino que son mucho más claras porque te enfrentas directamente contra los delanteros y hay que tener además <p/> con estas nuevas normas hay que tener mucho
<l tipo="mal_pronunciada" palabra="cuidado"> cuidao </l> porque a la mínim<a>a vez que hagas algo pues te
<n fuente="hablante" tipo="aspiración"/> , y si es más fuera de casa
<n fuente="hablante" tipo="aspiración"/> van a mirar mucho más al portero y enseguida vas a esta<a>r <p/> bueno , le van a expulsar , vamos .

</t>
</turno>

</anotacion>

Apéndice C

Fórmulas de reestimación de los parámetros de MOM discretos, continuos y semicontinuos.

C.1. Notación.

Con $O = O_1 O_2 \dots O_T$ nos referiremos a la secuencia de observaciones acústicas correspondientes a una muestra/pronunciación, donde T es la logitud de la secuencia, y O_t la observación acústica en el instante t . Si se tienen múltiples observaciones acústicas, O_t consta en realidad de varias componentes: $O_t = (O_t(1), \dots, O_t(R))$. Cada una de dichas componentes será una etiqueta microfonética en el caso de MOM discretos, o un vector de parámetros acústicos en el caso de MOM continuos y semicontinuos.

Sea $\lambda = (Q, \pi, A, B, q_F)$ la representación compacta de un MOM, donde Q es el conjunto de estados, con $|Q| = E$; $\pi = (\pi_i)$ es la distribución de estados iniciales, con π_i : probabilidad de que un estado $i \in Q$ sea inicial; $A = (a_{ij})$ es la matriz de transiciones, con a_{ij} : probabilidad de transitar de un estado $i \in Q$ a otro estado $j \in Q \cup \{q_F\}$, de modo que $\sum_{j \in Q \cup \{q_F\}} a_{ij} = 1$; q_F es un pseudo-estado auxiliar en el que no se emiten observaciones acústicas y cuya única utilidad es definir la probabilidad a_{iq_F} de transitar desde el estado i a otro modelo, que es donde se emitiría la siguiente observación acústica; finalmente, $B = \{b_i(O)\}$ es la distribución de probabilidades de emisión, con $b_i(O)$: probabilidad de emitir la observación acústica O en el estado $i \in Q$.

Con observaciones acústicas múltiples, $b_i(O) = \prod_{r=1}^R b_{ir}(O(r))$. En el caso de MOM discretos, $b_{ir}(s)$ es la probabilidad de emitir el símbolo $s \in [1, C]$ en la representación acústica $r \in [1, R]$ del estado $i \in Q$. En el caso de MOM continuos y semicontinuos, $b_{ir}(x)$ es la probabilidad de emitir el vector acústico $x \in \mathcal{R}^{D_r}$ en la representación acústica $r \in [1, R]$ del estado $i \in Q$, que se modela mediante una *función densidad de probabilidad (fdp)*, en concreto con una mezcla de gaussianas:

$$b_{ir}(x) = \sum_{g=1}^G c_{irg} \mathcal{N}(x; \mu_{r,h(i,r,g)}, \Sigma_{r,h(i,r,g)}) \quad (\text{C.1})$$

donde G es el tamaño de la mezcla, c_{irg} es el peso de la componente de índice g en la mezcla correspondiente a la representación r del estado i , y $h(i, r, g)$ indica a qué gaussiana, dentro del conjunto definido para la representación acústica r , corresponde la componente g del estado i .

En el caso de MOM continuos, $h(j, r, f) = h(i, r, g)$ si y sólo si $j = i$ y $f = g$, es decir, en cada representación acústica r todas las gaussianas de todos los estados son distintas, lo cual significa que hay EG gaussianas distintas en cada modelo. En el caso de MOM semicontinuos, $h(j, r, f) = h(i, r, g)$ si y sólo si $f = g$, es decir, en cada representación acústica r hay únicamente G gaussianas distintas compartidas por todos los estados de todos los modelos. En adelante, cuando hablemos de MOM continuos nos estaremos refiriendo al caso más general, en el que un cierto número de gaussianas son compartidas por varios estados de distintos modelos.

C.2. Algoritmo de Baum-Welch restringido.

En este caso, cada MOM λ dispone de un conjunto específico de muestras $\Omega = \{O^{(l)} \mid l = 1, \dots, L\}$, y se estima, en principio, de forma independiente. Las fórmulas de reestimación se basan en dos funciones, α y β , que se conocen como *probabilidad hacia adelante* y *probabilidad hacia atrás*. Estas funciones permiten simplificar extraordinariamente la estimación de parámetros, que de otra forma tendría complejidad exponencial. El cálculo de las funciones α y β de una muestra $O^{(l)}$ tiene una complejidad $O(E^2 T_l)$.

La función $\alpha^{(l)}(i, t)$ representa la probabilidad de la secuencia parcial $O_1^{(l)} O_2^{(l)} \dots O_t^{(l)}$ en el estado $i \in Q$ y en el instante $t \in [1, T_l]$, y se calcula de forma recursiva como sigue:

$$\begin{aligned} \alpha^{(l)}(i, 1) &= \pi_i b_i(O_1^{(l)}) \\ \alpha^{(l)}(i, t) &= \left[\sum_{j \in Q} \alpha^{(l)}(j, t-1) a_{ji} \right] b_i(O_t^{(l)}) \quad t \in [2, T_l] \end{aligned} \quad (\text{C.2})$$

La función $\beta^{(l)}(i, t)$ representa la probabilidad de la secuencia parcial $O_{t+1}^{(l)} O_{t+2}^{(l)} \dots O_{T_l}^{(l)}$ en el estado $i \in Q$ y en el instante $t \in [1, T_l]$, y se calcula también de forma recursiva:

$$\begin{aligned} \beta^{(l)}(i, T_l) &= a_{i q_F} \\ \beta^{(l)}(i, t) &= \sum_{j \in Q} a_{ij} b_j(O_{t+1}^{(l)}) \beta^{(l)}(j, t+1) \quad t \in [1, T_l - 1] \end{aligned} \quad (\text{C.3})$$

Una vez calculadas las funciones α y β , la probabilidad *a posteriori* de transitar de un estado i a otro estado j en el instante t , dados el modelo λ y la muestra $O^{(l)}$, se define:

$$\begin{aligned} \gamma^{(l)}(i, j, t) &= Pr(s_t = i, s_{t+1} = j | O^{(l)}, \lambda) \\ &= \frac{\alpha^{(l)}(i, t) a_{ij} b_j(O_{t+1}^{(l)}) \beta^{(l)}(j, t+1)}{Pr(O^{(l)} | \lambda)} \end{aligned} \quad (\text{C.4})$$

La probabilidad *a posteriori* de encontrarse en un estado i en el instante $t \in [1, T_l - 1]$, dados el modelo λ y la muestra $O^{(l)}$, será:

$$\begin{aligned} \gamma^{(l)}(i, t) &= Pr(s_t = i | O^{(l)}, \lambda) \\ &= \sum_{j \in Q} \gamma^{(l)}(i, j, t) \\ &= \frac{\alpha^{(l)}(i, t) \beta^{(l)}(i, t)}{Pr(O^{(l)} | \lambda)} \end{aligned} \quad (\text{C.5})$$

La transición al pseudo-estado final q_F ha de tener lugar obligatoriamente en el instante T_l . Así pues, se tiene:

$$\gamma^{(l)}(i, T_l) = \frac{\alpha^{(l)}(i, T_l)\beta^{(l)}(i, T_l)}{Pr(O^{(l)}|\lambda)} = \frac{\alpha^{(l)}(i, T_l)a_{iq_F}}{\sum_{j \in Q} \alpha^{(l)}(j, T_l)a_{jq_F}} \quad (C.6)$$

En cualquier instante t se verifica: $\sum_{i \in Q} \alpha^{(l)}(i, t)\beta^{(l)}(i, t) = Pr(O^{(l)}|\lambda)$, de modo que $\forall t \in [1, T_l]$ $\sum_{i \in Q} \gamma^{(l)}(i, t) = 1$. Las fórmulas de reestimación de las probabilidades de transición se expresan como un promedio normalizado para todas las muestras del corpus de entrenamiento:

$$\bar{\pi}_i = \frac{1}{L} \sum_{l=1}^L \gamma^{(l)}(i, 1) \quad (C.7)$$

$$\bar{a}_{ij} = \frac{\sum_{l=1}^L \sum_{t=1}^{T_l-1} \gamma^{(l)}(i, j, t)}{\sum_{l=1}^L \sum_{t=1}^{T_l} \gamma^{(l)}(i, t)} \quad (C.8)$$

$$\bar{a}_{iq_F} = \frac{\sum_{l=1}^L \gamma^{(l)}(i, T_l)}{\sum_{l=1}^L \sum_{t=1}^{T_l} \gamma^{(l)}(i, t)} \quad (C.9)$$

Hasta aquí no hay diferencia entre MOM discretos y MOM continuos, con observaciones acústicas simples o múltiples. En el caso de MOM discretos con observaciones acústicas múltiples, cada observación $O_t^{(l)}(r)$ contribuye a la distribución de emisión en la representación acústica r del estado i con el peso $\gamma^{(l)}(i, t)$:

$$\bar{b}_{ir}(k) = \frac{\sum_{l=1}^L \sum_{t=1}^{T_l} \gamma^{(l)}(i, t) \delta(O_t^{(l)}(r) = k)}{\sum_{l=1}^L \sum_{t=1}^{T_l} \gamma^{(l)}(i, t)} \quad (C.10)$$

donde:

$$\delta(X) = \begin{cases} 1 & \text{si } X = \text{VERDADERO} \\ 0 & \text{si } X = \text{FALSO} \end{cases} \quad (C.11)$$

En el caso de MOM continuos, para estimar la *fdp* definida por la ecuación C.1, en primer lugar se supone que las matrices de covarianza $\Sigma_{r,p}$ son diagonales, de manera que tan sólo es necesario estimar el peso de cada componente c_{irg} y los vectores de medias y de varianzas que definen cada gaussiana, $\mu_{r,p}$ y $\sigma_{r,p}^2$. A continuación se define la función auxiliar $\zeta^{(l)}(i, t, r, g)$:

$$\zeta^{(l)}(i, t, r, g) = \gamma^{(l)}(i, t) \frac{c_{irg} \mathcal{N}(O_t^{(l)}(r); \mu_{r,h(i,r,g)}, \sigma_{r,h(i,r,g)}^2)}{\sum_{f=1}^G c_{irf} \mathcal{N}(O_t^{(l)}(r); \mu_{r,h(i,r,f)}, \sigma_{r,h(i,r,f)}^2)} \quad (C.12)$$

Esta función representa la contribución de la observación $O_t^{(l)}(r)$ a la componente g de la mezcla correspondiente a la representación acústica r en el estado i . Una vez calculada esta

función, la fórmula de reestimación para el peso c_{irg} es:

$$\bar{c}_{irg} = \frac{\sum_{l=1}^L \sum_{t=1}^{T_l} \zeta^{(l)}(i, t, r, g)}{\sum_{l=1}^L \sum_{t=1}^{T_l} \gamma^{(l)}(i, t)} \quad (\text{C.13})$$

En general, los parámetros de las gaussianas no pueden reestimarse de forma independiente para cada modelo, sino que deben tenerse en cuenta, en cada representación acústica, todos los estados de todos los modelos, ya que una gaussiana puede formar parte de varias mezclas en distintos estados de distintos modelos. En la literatura esto se conoce como *ataadura* o *ligadura* de gaussianas. Supongamos que hay M modelos acústicos $\lambda_1, \dots, \lambda_M$, y que para cada modelo λ_n se tienen L_n muestras. Sea $O^{(n,l)} = (O^{(n,l)}(1), \dots, O^{(n,l)}(R))$ la muestra de índice l del modelo n . La componente c del vector de medias de la gaussiana de índice p en la representación acústica r se reestima según la siguiente expresión:

$$\bar{\mu}_{r,p}(c) = \frac{\sum_{n=1}^M \sum_{i \in Q^{(n)}} \sum_{g=1}^G \delta(h(n, i, r, g) = p) \sum_{l=1}^{L_n} \sum_{t=1}^{T_l} \zeta^{(n,l)}(i, t, r, g) O_t^{(n,l)}(r, c)}{\sum_{n=1}^M \sum_{i \in Q^{(n)}} \sum_{g=1}^G \delta(h(n, i, r, g) = p) \sum_{l=1}^{L_n} \sum_{t=1}^{T_l} \zeta^{(n,l)}(i, t, r, g)} \quad (\text{C.14})$$

donde $\zeta^{(n,l)}(i, t, r, g)$ es la misma función definida en C.12, y $h(n, i, r, g)$ tiene el mismo significado que $h(i, r, g)$, sólo que haciendo explícito en ambos casos el índice del modelo n ; por otro lado, $O_t^{(n,l)}(r, c)$ es la componente c del vector acústico $O_t^{(n,l)}(r)$. Análogamente, la componente c del vector de varianzas de la gaussiana de índice p en la representación acústica r se reestima como sigue:

$$\overline{\sigma^2}_{r,p}(c) = \frac{\sum_{n=1}^M \sum_{i \in Q^{(n)}} \sum_{g=1}^G \delta(h(n, i, r, g) = p) \sum_{l=1}^{L_n} \sum_{t=1}^{T_l} \zeta^{(n,l)}(i, t, r, g) [O_t^{(n,l)}(r, c)]^2}{\sum_{n=1}^M \sum_{i \in Q^{(n)}} \sum_{g=1}^G \delta(h(n, i, r, g) = p) \sum_{l=1}^{L_n} \sum_{t=1}^{T_l} \zeta^{(n,l)}(i, t, r, g)} - [\bar{\mu}_{r,p}(c)]^2 \quad (\text{C.15})$$

C.3. Algoritmo de Baum-Welch general.

En este caso se dispone de L pronunciaciones y para cada una de ellas se crea un MOM específico concatenando los MOM correspondientes a su transcripción fonética. Con cada pronunciación se entrena únicamente el MOM de la misma, sólo que ligando los parámetros de éste con los de los MOM simples que lo forman. De esta forma es posible entrenar MOM de unidades subléxicas sin disponer de una segmentación explícita.

Con respecto al algoritmo restringido, los cambios fundamentales tienen que ver con la definición de las funciones α y β . También van a ser distintas las definiciones de π_i y a_{iq_F} , ya que deben tenerse en cuenta las transiciones entre MOM simples que tienen lugar dentro del MOM específico de la pronunciación. El resto de funciones y las estimaciones van a cambiar sólo en la medida en que lo hacen α y β , que se definen como sigue:

$$\alpha^{(l)}(m, i, 1) = A^{(l)}(m, 1) \pi_i^{(H(m))} b_i^{(H(m))} (O_1^{(l)})$$

$$\alpha^{(l)}(m, i, t) = \left[A^{(l)}(m, t) \pi_i^{(H(m))} + \sum_{j \in Q^{(H(m))}} \alpha^{(l)}(m, j, t-1) a_{ji}^{(H(m))} \right] b_i^{(H(m))}(O_t^{(l)})$$

$$t \in [2, T_l] \quad (\text{C.16})$$

$$\beta^{(l)}(m, i, T_l) = a_{iq_F}^{(H(m))} B^{(l)}(m, T_l)$$

$$\beta^{(l)}(m, i, t) = \sum_{j \in Q^{(H(m))}} a_{ij}^{(H(m))} b_j^{(H(m))}(O_{t+1}^{(l)}) \beta^{(l)}(m, j, t+1) + a_{iq_F}^{(H(m))} B^{(l)}(m, t)$$

$$t \in [1, T_l - 1] \quad (\text{C.17})$$

donde el índice $m \in [1, F_l]$ recorre la secuencia de MOM simples que componen el MOM de la muestra l , con F_l : longitud de la transcripción fonética de la muestra l ; $H(m)$ indica qué MOM simple ocupa la posición m en el modelo de la muestra l , y se aplica como superíndice sobre los parámetros; $A^{(l)}(m, t)$ representa la probabilidad de entrar en el MOM simple de índice m justo en el instante t (sin haber emitido aún el símbolo correspondiente); y $B^{(l)}(m, t)$ representa la probabilidad de abandonar el MOM simple de índice m justo en el instante t (habiendo emitido ya el símbolo correspondiente). Las funciones $A^{(l)}(m, t)$ y $B^{(l)}(m, t)$ se definen como sigue:

$$A^{(l)}(m, t) = \sum_{j \in Q^{(H(m-1))}} \alpha^{(l)}(m-1, j, t-1) a_{jq_F}^{(H(m-1))}$$

$$m \in [2, F_l]$$

$$t \in [2, T_l] \quad (\text{C.18})$$

$$B^{(l)}(m, t) = \sum_{j \in Q^{(H(m+1))}} \pi_j^{(H(m+1))} b_j^{(H(m+1))}(O_{t+1}^{(l)}) \beta^{(l)}(m+1, j, t+1)$$

$$m \in [1, F_l - 1]$$

$$t \in [1, T_l - 1] \quad (\text{C.19})$$

En este trabajo, los MOM simples utilizan topologías izquierda-derecha como la de la Figura 4.2, en las que es necesario emitir una o más observaciones acústicas, y por tanto se tiene que $A^{(l)}(m, 1) = \delta(m=1)$, $A^{(l)}(1, t) = \delta(t=1)$, $B^{(l)}(m, T_l) = \delta(m=F_l)$ y $B^{(l)}(F_l, t) = \delta(t=T_l)$.

La probabilidad *a posteriori* de transitar en el instante t de un estado i a otro estado j del MOM simple que ocupa la posición m , dados la muestra $O^{(l)}$ y el modelo de la muestra $\Lambda^{(l)}$, se define:

$$\gamma^{(l)}(m, i, j, t) = Pr(s_t = (m, i), s_{t+1} = (m, j) | O^{(l)}, \Lambda^{(l)})$$

$$= \frac{\alpha^{(l)}(m, i, t) a_{ij}^{(H(m))} b_j^{(H(m))}(O_{t+1}^{(l)}) \beta^{(l)}(m, j, t+1)}{Pr(O^{(l)} | \Lambda^{(l)})} \quad (\text{C.20})$$

Análogamente, la probabilidad *a posteriori* de encontrarse en el instante t en el estado i del MOM simple que ocupa la posición m , dados la muestra $O^{(l)}$ y el modelo de la muestra $\Lambda^{(l)}$, se define:

$$\gamma^{(l)}(m, i, t) = Pr(s_t = (m, i) | O^{(l)}, \Lambda^{(l)})$$

$$= \frac{\alpha^{(l)}(m, i, t) \beta^{(l)}(m, i, t)}{Pr(O^{(l)} | \Lambda^{(l)})} \quad (\text{C.21})$$

Nótese que $\forall t \in [1, T_l]$ se verifica:

$$\sum_{m=1}^{F_l} \sum_{i \in Q^{(H(m))}} \alpha^{(l)}(m, i, t) \beta^{(l)}(m, i, t) = Pr(O^{(l)} | \Lambda^{(l)}) \quad (C.22)$$

por lo que:

$$\sum_{m=1}^{F_l} \sum_{i \in Q^{(H(m))}} \gamma^{(l)}(m, i, t) = 1 \quad (C.23)$$

Las fórmulas de reestimación de las probabilidades de transición del MOM simple de índice n son:

$$\bar{\pi}_i^{(n)} = \frac{\sum_{l=1}^L \sum_{m=1}^{F_l} \delta(H(m) = n) \sum_{t=1}^{T_l} \frac{A^{(l)}(m, t) \pi_i^{(n)} b_i^{(n)}(O_i^{(l)}) \beta^{(l)}(m, i, t)}{Pr(O^{(l)} | \Lambda^{(l)})}}{\sum_{l=1}^L \sum_{m=1}^{F_l} \delta(H(m) = n) \sum_{t=1}^{T_l} \frac{A^{(l)}(m, t) \sum_{j \in Q^{(n)}} [\pi_j^{(n)} b_j^{(n)}(O_i^{(l)}) \beta^{(l)}(m, j, t)]}{Pr(O^{(l)} | \Lambda^{(l)})}} \quad (C.24)$$

$$\bar{a}_{ij}^{(n)} = \frac{\sum_{l=1}^L \sum_{m=1}^{F_l} \delta(H(m) = n) \sum_{t=1}^{T_l} \gamma^{(l)}(m, i, j, t)}{\sum_{l=1}^L \sum_{m=1}^{F_l} \delta(H(m) = n) \sum_{t=1}^{T_l} \gamma^{(l)}(m, i, t)} \quad (C.25)$$

$$\bar{a}_{iq_F}^{(n)} = \frac{\sum_{l=1}^L \sum_{m=1}^{F_l} \delta(H(m) = n) \sum_{t=1}^{T_l} \frac{\alpha^{(l)}(m, i, t) a_{iq_F}^{(n)} B^{(l)}(m, t)}{Pr(O^{(l)} | \Lambda^{(l)})}}{\sum_{l=1}^L \sum_{m=1}^{F_l} \delta(H(m) = n) \sum_{t=1}^{T_l} \gamma^{(l)}(m, i, t)} \quad (C.26)$$

Fácilmente se comprueba que $\forall n \in [1, M] \quad \sum_{i \in Q^{(n)}} \bar{\pi}_i^{(n)} = 1$, y que $\forall n \in [1, M] \quad \forall i \in Q^{(n)} \quad \sum_{j \in Q^{(n)}} \bar{a}_{ij}^{(n)} + \bar{a}_{iq_F}^{(n)} = 1$.

En cuanto a las probabilidades de emisión, en el caso de MOM discretos, para cada MOM simple $n \in [1, M]$ y para cada representación acústica $r \in [1, R]$, la probabilidad de emitir el símbolo k se reestima como sigue:

$$\bar{b}_{ir}^{(n)}(k) = \frac{\sum_{l=1}^L \sum_{m=1}^{F_l} \delta(H(m) = n) \sum_{t=1}^{T_l} \gamma^{(l)}(m, i, t) \delta(O_t^{(l)}(r) = k)}{\sum_{l=1}^L \sum_{m=1}^{F_l} \delta(H(m) = n) \sum_{t=1}^{T_l} \gamma^{(l)}(m, i, t)} \quad (C.27)$$

En el caso de MOM continuos, en primer lugar se define la función auxiliar $\zeta^{(l)}(m, i, t, r, g)$:

$$\zeta^{(l)}(m, i, t, r, g) = \gamma^{(l)}(m, i, t) \frac{c_{irg}^{(H(m))} \mathcal{N}(O_t^{(l)}(r); \mu_{r, h(H(m), i, r, g)}, \sigma_{r, h(H(m), i, r, g)}^2)}{\sum_{f=1}^G c_{irf}^{(H(m))} \mathcal{N}(O_t^{(l)}(r); \mu_{r, h(H(m), i, r, f)}, \sigma_{r, h(H(m), i, r, f)}^2)} \quad (C.28)$$

Esta función representa la contribución de la observación $O_t^{(l)}(r)$ a la componente g de la mezcla correspondiente a la representación acústica r en el estado i del MOM simple que

ocupa la posición m en el modelo de la muestra l . Una vez calculada esta función, la fórmula de reestimación para el peso $c_{irg}^{(n)}$ es:

$$\bar{c}_{irg}^{(n)} = \frac{\sum_{l=1}^L \sum_{m=1}^{F_l} \delta(H(m) = n) \sum_{t=1}^{T_l} \zeta^{(l)}(m, i, t, r, g)}{\sum_{l=1}^L \sum_{m=1}^{F_l} \delta(H(m) = n) \sum_{t=1}^{T_l} \gamma^{(l)}(m, i, t)} \quad (\text{C.29})$$

Finalmente, la componente c del vector de medias de la gaussiana de índice p en la representación acústica r se reestima según la siguiente expresión:

$$\bar{\mu}_{r,p}(c) = \frac{\sum_{l=1}^L \sum_{m=1}^{F_l} \sum_{i \in Q(H(m))} \sum_{g=1}^G \delta(h(H(m), i, r, g) = p) \sum_{t=1}^{T_l} \zeta^{(l)}(m, i, t, r, g) O_t^{(l)}(r, c)}{\sum_{l=1}^L \sum_{m=1}^{F_l} \sum_{i \in Q(H(m))} \sum_{g=1}^G \delta(h(H(m), i, r, g) = p) \sum_{t=1}^{T_l} \zeta^{(l)}(m, i, t, r, g)} \quad (\text{C.30})$$

y la componente c del vector de varianzas de la gaussiana de índice p en la representación acústica r :

$$\bar{\sigma}_{r,p}^2(c) = \frac{\sum_{l=1}^L \sum_{m=1}^{F_l} \sum_{i \in Q(H(m))} \sum_{g=1}^G \delta(h(H(m), i, r, g) = p) \sum_{t=1}^{T_l} \zeta^{(l)}(m, i, t, r, g) [O_t^{(l)}(r, c)]^2}{\sum_{l=1}^L \sum_{m=1}^{F_l} \sum_{i \in Q(H(m))} \sum_{g=1}^G \delta(h(H(m), i, r, g) = p) \sum_{t=1}^{T_l} \zeta^{(l)}(m, i, t, r, g) - [\bar{\mu}_{r,p}(c)]^2} \quad (\text{C.31})$$

C.4. Algoritmo de entrenamiento por Viterbi.

Sea $\hat{S}^{(l)} = \{\hat{s}_1^{(l)}, \dots, \hat{s}_{T_l}^{(l)}\}$ la secuencia óptima de estados en el modelo $\Lambda^{(l)}$ de la muestra $O^{(l)}$ —secuencia obtenida mediante el algoritmo de Viterbi. En cada instante t hay un estado óptimo $\hat{s}_t^{(l)}$ en $\Lambda^{(l)}$, que corresponde a un cierto MOM, de índice $\hat{m}_t^{(l)}$, y una sola transición tiene lugar, desde $\hat{s}_t^{(l)}$ hasta $\hat{s}_{t+1}^{(l)}$. Las transiciones que suceden en $S^{(l)}$ simplemente son contadas, acumuladas y finalmente, tras recorrer todas las muestras, normalizadas para reestimar las probabilidades de transición. Por otra parte, cada observación acústica $O_t^{(l)}$ se utiliza para reestimar únicamente los parámetros de emisión en el estado $\hat{s}_t^{(l)}$.

Sea $I^{(l)}(n, i)$ el número de veces que en la secuencia óptima $\hat{S}^{(l)}$ se accede al MOM simple de índice n a través del estado i ; $F^{(l)}(n, i)$ el número de veces que en la secuencia óptima $\hat{S}^{(l)}$ el MOM simple de índice n es abandonado desde el estado i ; y $N^{(l)}(n, i, j)$ el número de veces que en la secuencia óptima $\hat{S}^{(l)}$ tiene lugar la transición entre los estados i y j del MOM simple de índice n . Teniendo en cuenta estas definiciones, las fórmulas de reestimación de las probabilidades de transición se expresan como sigue:

$$\bar{\pi}_i^{(n)} = \frac{\sum_{l=1}^L I^{(l)}(n, i)}{\sum_{l=1}^L \sum_{j \in Q(n)} I^{(l)}(n, j)} \quad (\text{C.32})$$

$$\bar{a}_{ij}^{(n)} = \frac{\sum_{l=1}^L N^{(l)}(n, i, j)}{\sum_{l=1}^L \left\{ \sum_{k \in Q^{(n)}} N^{(l)}(n, i, k) + F^{(l)}(n, i) \right\}} \quad (\text{C.33})$$

$$\bar{a}_{iq_F}^{(n)} = \frac{\sum_{l=1}^L F^{(l)}(n, i)}{\sum_{l=1}^L \left\{ \sum_{k \in Q^{(n)}} N^{(l)}(n, i, k) + F^{(l)}(n, i) \right\}} \quad (\text{C.34})$$

La cantidad $N^{(l)}(n, i) = \sum_{j \in Q^{(n)}} N^{(l)}(n, i, j) + F^{(l)}(n, i)$ es el número de veces que aparece el estado i del MOM simple de índice n en la secuencia óptima $\hat{S}^{(l)}$, es decir, el número de veces que $\hat{m}_t^{(l)} = n$ y $\hat{s}_t^{(l)} = i$. En el caso de MOM discretos, se define $E^{(l)}(n, i, r, k)$ como el número de veces que $O_t^{(l)}(r) = k$ cuando $\hat{m}_t^{(l)} = n$ y $\hat{s}_t^{(l)} = i$. Entonces las probabilidades de emisión se reestiman como sigue:

$$\bar{b}_{ir}^{(n)}(k) = \frac{\sum_{l=1}^L E^{(l)}(n, i, r, k)}{\sum_{l=1}^L N^{(l)}(n, i)} \quad (\text{C.35})$$

En el caso de MOM continuos, en primer lugar se necesita la siguiente definición:

$$Z(x; n, i, r, g) = \frac{c_{irg}^{(n)} \mathcal{N}(x; \mu_{r,h(n,i,r,g)}, \sigma_{r,h(n,i,r,g)}^2)}{\sum_{f=1}^G c_{irf}^{(n)} \mathcal{N}(x; \mu_{r,h(n,i,r,f)}, \sigma_{r,h(n,i,r,f)}^2)} \quad (\text{C.36})$$

donde se han supuesto matrices de covarianza diagonales.

A continuación, los pesos de las gaussianas se reestiman como sigue:

$$\bar{c}_{irg}^{(n)} = \frac{\sum_{l=1}^L \sum_{t=1}^{T_l} \delta(\hat{m}_t^{(l)} = n) \delta(\hat{s}_t^{(l)} = i) Z(O_t^{(l)}(r); n, i, r, g)}{\sum_{l=1}^L \sum_{t=1}^{T_l} \delta(\hat{m}_t^{(l)} = n) \delta(\hat{s}_t^{(l)} = i)} \quad (\text{C.37})$$

Finalmente, la componente c del vector de medias de la gaussiana de índice p en la representación acústica r se reestima según la siguiente expresión:

$$\bar{\mu}_{r,p}(c) = \frac{\sum_{l=1}^L \sum_{t=1}^{T_l} \sum_{f=1}^G \delta(h(\hat{m}_t^{(l)}, \hat{s}_t^{(l)}, r, f) = p) Z(O_t^{(l)}(r); \hat{m}_t^{(l)}, \hat{s}_t^{(l)}, r, p) O_t^{(l)}(r, c)}{\sum_{l=1}^L \sum_{t=1}^{T_l} \sum_{f=1}^G \delta(h(\hat{m}_t^{(l)}, \hat{s}_t^{(l)}, r, f) = p) Z(O_t^{(l)}(r); \hat{m}_t^{(l)}, \hat{s}_t^{(l)}, r, p)} \quad (\text{C.38})$$

y la componente c del vector de varianzas de la gaussiana de índice p en la representación acústica r :

$$\bar{\sigma}_{r,p}^2(c) = \frac{\sum_{l=1}^L \sum_{t=1}^{T_l} \sum_{f=1}^G \delta(h(\hat{m}_t^{(l)}, \hat{s}_t^{(l)}, r, f) = p) Z(O_t^{(l)}(r); \hat{m}_t^{(l)}, \hat{s}_t^{(l)}, r, p) [O_t^{(l)}(r, c)]^2}{\sum_{l=1}^L \sum_{t=1}^{T_l} \sum_{f=1}^G \delta(h(\hat{m}_t^{(l)}, \hat{s}_t^{(l)}, r, f) = p) Z(O_t^{(l)}(r); \hat{m}_t^{(l)}, \hat{s}_t^{(l)}, r, p)} - [\bar{\mu}_{r,p}(c)]^2 \quad (\text{C.39})$$

Apéndice D

Resultados de DAF sobre INFOTREN con la parametrización del ETSI para entornos ruidosos.

El estándar del ETSI para entornos ruidosos (ETSI, 2003b) ha mostrado en promedio una disminución del 53 % en el error de reconocimiento en distintas condiciones de ruido ambiental, con respecto al estándar para entornos *limpios* (ETSI, 2003a). Sólo muy recientemente, cuando ya se había llevado a cabo toda la experimentación concerniente a los modelos acústicos, se han recibido las especificaciones de este procedimiento de análisis, cuyas líneas básicas quedaron reflejadas en la Tabla 4.1. Por tanto, lo que se plantea en este apéndice es simplemente una rápida constatación de las mejoras que la citada parametrización puede aportar en el reconocimiento del habla espontánea.

INFOTREN no es una base de datos *intrínsecamente ruidosa*. Sin embargo, las señales se adquieren a través de líneas telefónicas, y los hablantes llaman desde oficinas en las que hay ordenadores encendidos, personas que teclean o incluso ciertos ruidos aislados como puertas que se cierran, manejo de papeles, etc. También se producen con cierta frecuencia saturaciones de la señal y soplidos en el auricular. Para comprobar el grado en que las tasas pueden mejorar utilizando una nueva parametrización, se ha aplicado el estándar del ETSI para entornos ruidosos a las señales de INFOTREN, se ha realizado una cuantificación vectorial según el método descrito en el apartado 4.1.2, y se han entrenado MOM discretos y continuos, utilizando el conjunto de unidades subléxicas ampliado. Los resultados de DAF se muestran en la Tabla D.1.

Con MOM continuos las tasas son alrededor de 7 puntos más altas, lo cual implica reducciones relativas del error en DAF de cerca del 17 %. Sin embargo, con MOM discretos se observa un empeoramiento muy notable del rendimiento, de casi 7 puntos absolutos, un 15.6 % peor que con la parametrización estándar del GRFyTH. Así pues, la parametrización del ETSI sólo se puede plantear como alternativa con MOM continuos. Nótese que aunque se utilicen

Tabla D.1: Tasas de DAF sobre INFOTREN, utilizando el estándar de parametrización del ETSI para entornos ruidosos, con MOM discretos y continuos del conjunto ampliado de unidades incontextuales. Se muestran únicamente las tasas obtenidas habiendo filtrado todas las unidades salvo los fonemas.

MOMd	MOMc-8g	MOMc-16g	MOMc-32g	MOMc-64g
48.91	65.42	67.11	67.82	68.24

MOM continuos, sigue siendo necesario entrenar previamente MOM discretos para inicializarlos. Estos mismos resultados demuestran que aunque esta inicialización no sea demasiado precisa, los parámetros de los MOM continuos convergen a valores óptimos durante el proceso de entrenamiento. Nótese también que se está utilizando un conjunto de unidades subléxicas incontextuales, y que, sin embargo, las tasas de DAF con MOM continuos son superiores a las que se obtienen utilizando unidades subléxicas contextuales con la parametrización del GRFyTH.

Bibliografía

- Ahadi, S. M., Woodland, P. C., 1997. Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models. *Computer, Speech and Language* 11 (3), 187–206.
- Ajmera, J., Bourlard, H., Lapidot, I., McCowan, I., 2002. Unknown-multiple speaker clustering using HMM. In: *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*. Denver, Colorado, pp. 573–576.
- Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., Weinert, R., 1991. The HCRC Map Task corpus. *Language and Speech* 34 (4), 351–366, distributed by the Linguistic Data Consortium: <http://www.ldc.upenn.edu>.
- Aubert, X. L., 1999. One pass cross word decoding for large vocabularies based on a lexical tree search organization. In: *Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH)*. Budapest, Hungary, pp. 1559–1562.
- Bahl, L. R., Balakrishnan-Aiyer, S., Bellegarda, J. R., Franz, M., Gopalakrishnan, P. S., Nahamoo, D., Novak, M., Padmanabhan, M., Picheny, M. A., Roukos, S., 1995. Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 41–44.
- Bahl, L. R., de Souza, P. V., Gopalakrishnan, P. S., Nahamoo, D., Picheny, M. A., 1991. Decision trees for phonological rules in continuous speech. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 185–188.
- Ballester, A., Santamaría, C., Marcos-Marín, F. A., 1993. Transcription conventions used for the Corpus of Spoken Contemporary Spanish. *Literary and Linguistic Computing* 8 (4), 283–292.
- Barras, C., Geoffrois, E., Wu, Z., Liberman, M., 2001. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication* 33 (1-2), 5–22.
- Bear, J., Dowding, J., Shriberg, E., Price, P., 1993. A system for labeling self-repairs in speech. Technical Note 522, SRI.
- Beulen, K., Ortmanns, S., Elting, C., 1999. Dynamic programming search techniques for across-word modelling in speech recognition. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Phoenix, Arizona, USA, pp. 609–612.

- Beyerlein, P., Ullrich, M., Wilcox, P., 1997. Modelling and decoding of crossword context dependent phones in the Philips large vocabulary continuous speech recognition system. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Rhodes, Greece, pp. 1163–1166.
- Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., Liberman, M., 2000. ATLAS: A flexible and extensible architecture for linguistic annotation. In: Proceedings of the International Conference on Language Resources and Evaluation. Paris: European Language Resources Association, France, pp. 1699–1706.
- Bird, S., Liberman, M., 2001. A formal framework for linguistic annotation. *Speech Communication* 33 (1), 23–60.
- Bird, S., Maeda, K., Ma, X., Lee, H., Randall, B., Zayat, S., 2002. TableTrans, MultiTrans, InterTrans and TreeTrans: diverse tools built on the Annotation Graph Toolkit. In: Proceedings of the International Conference on Language Resources and Evaluation. Paris: European Language Resources Association, France, pp. 364–370.
- Bonafonte, A., Aibar, P., Castell, N., Lleida, E., Mariño, J. B., Sanchís, E., Torres, I., 2000. Desarrollo de un sistema de diálogo oral en dominios restringidos. In: Actas de las I Jornadas en Tecnología del Habla. University of Sevilla, Spain, project website: <http://gps-tsc.upc.es/veu/basurde>.
- Bonafonte, A., Estany, R., Vives, E., 1995. Study of subword units for Spanish speech recognition. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Madrid, Spain, pp. 1607–1610.
- Bonafonte, A., Vidal, J., Nogueiras, A., 1996. Duration modeling with expanded HMM applied to speech recognition. In: Proceedings of the International Conference on Speech and Language Processing (ICSLP). Philadelphia, MA, USA, pp. 1097–1100.
- Burnard, L., Sperberg-McQueen, C. M., 1995. TEI Lite: An introduction to Text Encoding for Interchange. Document No. TEI U 5, <http://www-tei.uic.edu/orgs/tei/intros/teiu5.tei>.
- Burshtein, D., 1996. Robust parametric modeling of durations in hidden Markov models. *IEEE Transactions on Speech and Audio Processing* 4 (3), 240–242.
- Calzolari, N., McNaught, J., Zampolli, A., 1996. EAGLES: Expert Advisory Group on Language Engineering Standards. Editor's Introduction. <http://www.ilc.cnr.it/EAGLES96/edintro/edintro.html>.
- Casacuberta, F., 1992. Estimación de los parámetros estadísticos de los modelos ocultos de Markov discretos, continuos y semicontinuos. Informe de investigación DSIC-II/18/92, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.
- Casacuberta, F., García, R., Llisterri, J., Nadeu, C., Pardo, J. M., Rubio, A., 1991. Development of Spanish corpora for speech recognition. In: Proceedings of the Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assessment Methods. Chiavari, Italy.
- Chelba, C., 2000. Exploiting syntactic structure for natural language modeling. Ph.D. thesis, Johns Hopkins University, Baltimore, Maryland, USA.

- Chesta, C., Siohan, O., Lee, C. H., 1999. Maximum A Posteriori Linear Regression for hidden Markov model adaptation. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Budapest, Hungary, pp. 211–214.
- Chou, W., 1999. Maximum A Posteriori Linear Regression with elliptically symmetric matrix variate priors. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Budapest, Hungary, pp. 1–4.
- Chung, G. Y., Seneff, S., 1999. A hierarchical duration model for speech recognition based on the ANGIE framework. *Speech Communication* 27 (2), 113–134.
- Clark, H. H., 2002. Speaking in time. *Speech Communication* 36 (1-2), 5–13.
- Colthurst, T., Kimball, O., Richardson, F., Shu, H., Wooters, C., Iyer, R., Gish, H., 2000. The 2000 BBN Byblos LVCSR System. In: Proceedings of the NIST Speech Transcription Workshop. University of Maryland, <http://www.nist.gov/speech/publications/tw00/>.
- CORLEC, 1992. Corpus Oral de Referencia de la Lengua Española Contemporánea. Corpus website: http://www.lllf.uam.es/corpus/corpus_oral.html.
- Cox, S. J., 1995. Predictive speaker adaptation in speech recognition. *Computer, Speech and Language* 9 (1), 1–17.
- Cunningham, H., 2000. Software Architecture for Language Engineering. Ph.D. thesis, University of Sheffield, <http://gate.ac.uk/sale/thesis/>.
- Cunningham, H., 2002. GATE, a General Architecture for Text Engineering. *Computers and the Humanities* 36, 223–254, GATE url: <http://gate.ac.uk>.
- Davis, S. B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Speech and Audio Processing* 28 (4), 357–366.
- Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society – Series B* 39 (1), 1–38.
- den Os, E., Boves, L., Lamel, L., Baggia, P., 1999. Overview of the ARISE project. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Budapest, Hungary, pp. 1527–1530.
- Digalakis, V. V., Ritchev, D., Neumeyer, L. G., 1995. Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions on Speech and Audio Processing* 3 (3), 357–366.
- DIHANA, 2003-2005. Sistema de diálogo para el acceso a la información mediante habla espontánea en diferentes entornos. Web del proyecto: <http://www.dsic.upv.es/~dihana/>.
- Díaz, J., Peinado, A. M., Rubio, A., Segarra, E., Prieto, N., Casacuberta, F., 1998. ALBAYZIN: a task-oriented Spanish speech corpus. In: Proceedings of the International Conference on Language Resources and Evaluation. Granada, Spain, pp. 497–502.
- Eide, E., Gish, H., 1996. A parametric approach to vocal tract length normalization. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 346–348.

- Eklund, R., 2001. Prolongations: a dark horse in the disfluency stable. In: Proceedings of the ISCA Workshop on Disfluency in Spontaneous Speech. University of Edinburgh, Scotland, pp. 5–8.
- Emori, T., Shinoda, K., 2001. Rapid vocal tract length normalization using maximum likelihood estimation. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Aalborg, Denmark, pp. 1649–1652.
- ETSI, September 2003a. European Telecommunications Standards Institute, ETSI Standard ES 201 108 V1.1.3. Speech Processing, Transmission and Quality Aspects (STQ). Distributed Speech Recognition. Front-End Feature Extraction Algorithm. Compression Algorithms. Reference RES/STQ-00044. Web page: <http://www.etsi.org>.
- ETSI, November 2003b. European Telecommunications Standards Institute, ETSI Standard ES 202 050 V1.1.3. Speech Processing, Transmission and Quality Aspects (STQ). Distributed Speech Recognition. Advanced Front-End Feature Extraction Algorithm. Compression Algorithms. Reference RES/STQ-00056. Web page: <http://www.etsi.org>.
- Forney, G. D., 1973. The Viterbi algorithm. Proceedings of the IEEE 61 (3), 268–278.
- Gales, M., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Computer, Speech and Language 12 (2), 75–98.
- Gales, M., 2000. Cluster adaptive training of hidden Markov models. IEEE Transactions on Speech and Audio Processing 8 (4), 417–428.
- Gales, M., Woodland, P. C., 1996. Mean and variance adaptation within the MLLR framework. Computer, Speech and Language 10 (4), 249–264.
- Gao, Y., Padmanabhan, M., Picheny, M., 1997. Speaker adaptation based on pre-clustering training speakers. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Rhodes, Greece, pp. 2091–2094.
- García, P., Vidal, E., 1990. Inference of k-testable languages in the strict sense and application to syntactic pattern recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (9), 920–925.
- Gauvain, J. L., Lamel, L., Adda, G., Adda-Decker, M., 1994. The LIMSI continuous speech dictation system: Evaluation on the ARPA Wall Street Journal task. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Adelaide, Australia, pp. 557–560.
- Gauvain, J. L., Lee, C. H., 1994. Maximum A Posteriori estimation for multivariate gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing 2 (2), 291–298.
- Godfrey, J. J., Holliman, E. C., McDaniel, J., 1992. SWITCHBOARD: Telephone speech corpus for research and development. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vol. 1. San Francisco (CA), USA, pp. 517–520, distributed by the Linguistic Data Consortium: <http://www.ldc.upenn.edu>.
- Goronzy, S., Kompe, R., 1999. A MAP-like weighting scheme for MLLR speaker adaptation. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Budapest, Hungary, pp. 5–8.
- Hain, T., Woodland, P. C., Evermann, G., Povey, D., 2000. The CU-HTK March 2000 Hub5E Transcription System. In: Proceedings of the NIST Speech Transcription Workshop. University of Maryland, <http://www.nist.gov/speech/publications/tw00/>.

- Hazen, T. J., 1998. The use of speaker correlation information for automatic speech recognition. Ph.D. thesis, Massachusetts Institute of Technology.
- Hazen, T. J., 2000. A comparison of novel techniques for rapid speaker adaptation. *Speech Communication* 31 (1), 15–33.
- Heeman, P. A., 1997. Speech repairs, intonational boundaries and discourse markers: modeling speakers' utterances in spoken dialog. Ph.D. thesis, University of Rochester.
- Heeman, P. A., 1999. Modeling speech repairs and intonational phrasing to improve speech recognition. In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*. Keystone, Colorado.
- Heeman, P. A., Allen, J. F., 1995a. The TRAINS 93 dialogues. TRAINS Technical Note 94-2, The University of Rochester, Computer Science Department, Rochester, New York.
- Heeman, P. A., Allen, J. F., 1995b. The TRAINS spoken dialog corpus. Distributed by the Linguistic Data Consortium: <http://www ldc.upenn.edu>.
- Heeman, P. A., Allen, J. F., 1999. Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics* 25 (4), 527–571.
- Hermansky, H., Morgan, N., 1994. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing* 2 (4), 578–589.
- Huang, X. D., Ariki, Y., Jack, M. A., 1990. Hidden Markov Models for Speech Recognition. Edinburgh Information Technology Series, Edinburgh University Press, Edinburgh, UK.
- Ide, N., 1994. Encoding standards for large text resources: the Text Encoding Initiative. Department of Computer Science. Vassar College. Poughkeepsie, New York (USA), <http://www.cs.vassar.edu/~ide/papers/tei10.ps>.
- Ide, N., Bonhomme, P., Romary, L., 2000. XCES: an XML-based encoding standard for linguistic corpora. In: *Proceedings of the International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association, France, pp. 825–830.
- Ide, N., Priest-Dorman, G., 2000. CES: Corpus Encoding Standard - Document CES 1. Version 1.5. <http://www.cs.vassar.edu/CES/>.
- Ide, N., Véronis, J., 1994a. Background and context for the development of a Corpus Encoding Standard. EAGLES Technical Report, Istituto di Linguistica Computazionale, Universite di Pisa, Italy, 30 pages.
- Ide, N., Véronis, J., 1994b. MULTEXT: Multilingual Text Tools and Corpora. In: *Proceedings of the International Conference on Computational Linguistics*. Vol. I. Kyoto, Japan, pp. 588–592.
- Jeanrenaud, P., Eide, E., Chaudhari, U., McDonough, J., Ng, K., Siu, M., Gish, H., 1995. Reducing word error rate on conversational speech from the Switchboard corpus. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 53–56.
- Jelinek, F., 1990. Self-Organized Language Modeling for Speech Recognition. In *Readings in Speech Recognition*, A. Waibel and K. F. Lee (Eds.), chapter 8.1, Morgan Kaufmann Publishers, San Mateo, California, USA, pp. 450–506.

- Jiang, L., Huang, X., 2000. Subword-dependent speaker clustering for improved speech recognition. In: Proceedings of the International Conference on Speech and Language Processing (ICSLP). Vol. 4. Beijing, China, pp. 137–140.
- Jin, H., Kubala, F., Schwartz, R., 1997. Automatic speaker clustering. In: Proceedings of the DARPA Speech Recognition Workshop. Chantilly, VA, pp. 108–111.
- Johnson, S. E., Woodland, P. C., 1998. Speaker clustering using direct maximisation of the MLLR-adapted likelihood. In: Proceedings of the International Conference on Speech and Language Processing (ICSLP). pp. 1775–1779.
- Juang, B. H., Rabiner, L. R., 1990. The segmental k-means algorithm for estimating parameters of Hidden Markov Models. *IEEE Transactions on Acoustics, Speech and Signal Processing* 38 (9), 1639–1641.
- Juang, B. H., Rabiner, L. R., Wilpon, J. G., 1987. On the use of bandpass liftering in speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* 35 (7), 947–954.
- Katz, S. M., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing* 35 (3), 400–401.
- Kosaka, T., Sagayama, S., 1994. Tree-structured speaker clustering for fast speaker adaptation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 245–248.
- Kuhn, R., Junqua, J. C., Nguyen, P., Niedzielski, N., 2000. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing* 8 (6), 695–707.
- Kuhn, R., Lazarides, A., Normandin, Y., Brousseau, J., 1995. Improved decision trees for phonetic modeling. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 552–555.
- Lee, K. F., 1989. *Automatic Speech Recognition: The Development of the SPHINX System*. Kluwer Academic Publishers, Boston, Massachusetts.
- Lee, K. F., Hayamizu, S., Hon, H. W., Huang, C., Swartz, J., Weide, R., 1990. Allophone clustering for continuous speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Albuquerque, NM, USA, pp. 749–752.
- Lee, L., Rose, R. C., 1996. Speaker normalization using efficient frequency warping procedures. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 353–356.
- Leggetter, C. J., Woodland, P. C., 1995a. Flexible speaker adaptation using Maximum Likelihood Linear Regression. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Madrid, Spain, pp. 1155–1158.
- Leggetter, C. J., Woodland, P. C., 1995b. Maximum Likelihood Linear Regression for speaker adaptation of Continuous Density Hidden Markov Models. *Computer, Speech and Language* 9 (2), 171–185.
- Levelt, W. J. M., 1989. *Speaking: From Intention to Articulation*. The MIT Press. ACL-MIT Press Series in Natural-Language Processing, Cambridge, Massachusetts.

- Lickley, R. J., 1998. HCRC Disfluency Coding Manual. HCRC Technical Report 100, Human Communication Research Centre, University of Edinburgh.
- Linde, Y., Buzo, A., Gray, R. M., 1980. An algorithm for vector quantizer design. *IEEE Transactions on Communications* 28 (1), 84–95.
- Liu, D., Nguyen, L., Matsoukas, S., Davenport, J., Kubala, F., Schwartz, R., 1998. Improvements in spontaneous speech recognition. In: *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. Lansdowne, Virginia.
- Ljolje, A., Hindle, D. M., Riley, M. D., Sproat, R. W., 2000. The AT&T LVCSR-2000 System. In: *Proceedings of the NIST Speech Transcription Workshop*. University of Maryland, <http://www.nist.gov/speech/publications/tw00/>.
- López de Ipiña, K., Rodríguez, L. J., Varona, A., Torres, I., 1999. Decision tree based context dependent sublexical units for Spanish continuous speech recognition tasks. In: *Proceedings of the Spanish Symposium on Pattern Recognition and Image Analysis*. Vol. 1. Bilbao, Spain, pp. 53–59.
- López de Ipiña, K., Torres, I., Oñederra, L., Varona, A., Rodríguez, L. J., 2000. Selection of sublexical units for continuous speech recognition of Basque. In: *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*. Vol. 1. Beijing, China, pp. 544–547.
- MADCOW, 1992. Multi-site data collection for a spoken language corpus. In: *Proceedings of the DARPA Speech and Natural Language Workshop*. Harriman NY, pp. 7–14.
- Maeda, K., Bird, S., Ma, X., Lee, H., 2002. Creating annotation tools with the Annotation Graph Toolkit. In: *Proceedings of the International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association, France, pp. 1914–1921.
- Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., 1994. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19 (2), 313–330.
- Mariño, J. B., Nogueiras, A., Bonafonte, A., 1997. The demiphone: an efficient subword unit for continuous speech recognition. In: *Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH)*. Rhodes, Greece, pp. 1215–1218.
- Mariño, J. B., Nogueiras, A., Pachès-Leal, P., Bonafonte, A., 2000. The demiphone: and efficient contextual subword unit for continuous speech recognition. *Speech Communication* 32 (3), 187–197.
- McKelvie, D., Isard, A., Mengel, A., Moeller, M. B., Grosse, M., Klein, M., 2001. The MATE workbench - an annotation tool for XML coded speech corpora. *Speech Communication* 33 (1-2), 97–112.
- McLaughlin, J., Reynolds, D., Singer, E., O’Leary, G. C., 1999. Automatic speaker clustering from multi-speaker utterances. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 2. pp. 817–820.
- Mengel, A., Dybkjaer, L., Garrido, J., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A., Soria, C., 1999. MATE Dialogue Annotation Guidelines. Deliverable D2.1, MATE: Multilevel Annotation Tools Engineering: LE Telematics Project LE4-8370.
- Merhav, N., Ephraim, Y., 1991. Hidden Markov modeling using a dominant state sequence with application to speech recognition. *Computer, Speech and Language* 5, 327–339.

- Meteer, M., Taylor, A., MacIntyre, R., Iyer, R., 1995. Disfluency Annotation Stylebook for the Switchboard Corpus. Technical Report (revised by Ann Taylor, June 1995), Published by the Linguistic Data Consortium, University of Pennsylvania, Department of Computer and Information Science.
- Milde, J. T., Gut, U., 2002. The TASX-environment: an XML-based toolset for time aligned speech corpora. In: Proceedings of the International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain, pp. 1922–1927, <http://tasxforce.lili.uni-bielefeld.de>.
- Mohri, M., Riley, M., Hindle, D., Ljolje, A., Pereira, F., 1998. Full expansion of context-dependent networks in large vocabulary speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seattle, Washington, USA, pp. 665–668.
- Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J. B., Nadeu, C., 1993. ALBAYZIN speech database: Design of the phonetic corpus. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Berlin, Germany, pp. 175–178.
- Nadeu, C., Macho, D., Hernando, J., 2001. Time and frequency filtering of filter bank energies for robust HMM speech recognition. *Speech Communication* 34 (2), 93–114.
- Naito, M., Deng, L., Sagisaka, Y., 2002. Speaker clustering for speech recognition using vocal tract parameters. *Speech Communication* 36 (3-4), 305–315.
- Nakatani, C. H., Hirschberg, J., 1994. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America* 95 (3), 1603–1616.
- Nedel, J. P., Singh, R., Stern, R. M., 2000. Phone transition acoustic modeling: application to speaker independent and spontaneous speech systems. In: Proceedings of the International Conference on Speech and Language Processing (ICSLP). Vol. 4. Beijing, China, pp. 572–575.
- NIST, 1998. A Universal Transcription Format (UTF) annotation specification for evaluation of spoken language technology corpora. Version 1.0, http://www.nist.gov/speech/hub4_98/utf-1.0-v2.ps.
- Odell, J. J., 1995. The use of context in large vocabulary speech recognition. Ph.D. thesis, University of Cambridge.
- O’Shaughnessy, D., 1999. Better detection of hesitations in spontaneous speech. In: Proceedings of the ISCA Workshop on Disfluency in Spontaneous Speech. U.C. Berkeley, pp. 39–42, Satellite Meeting of ICPHS99.
- Oualline, S., 2001. Vi Improved – Vim. New Riders Publishing, Indianapolis, Indiana, versión PDF: <http://www.truth.sk/vim/vimbook-OPL.pdf>.
- Oviatt, S., 1995. Predicting spoken disfluencies during human-computer interaction. *Computer, Speech and Language* 9 (1), 19–35.
- Padmanabhan, M., Bahl, L. R., Nahamoo, D., Picheny, M. A., 1996. Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 701–704.
- Paul, D. B., Baker, J. M., 1992. The design for the Wall Street Journal-based CSR corpus. In: Proceedings of the DARPA Speech and Natural Language Workshop. Harriman, NY, pp. 357–362.

- Picone, J. W., 1993. Signal modeling techniques in speech recognition. *Proceedings of the IEEE* 81 (9), 1215–1247.
- Pitz, M., Molau, S., Schlüter, R., Ney, H., 2001. Vocal tract normalization equals linear transformation in cepstral space. In: *Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH)*. Aalborg, Denmark, pp. 2653–2656.
- Pusateri, E. J., Hazen, T. J., 2002. Rapid speaker adaptation using speaker clustering. In: *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*. Denver, Colorado, pp. 61–64.
- Quilis, A., 1981. *Fonética acústica de la lengua española*. Gredos, Madrid.
- Rabiner, L. R., 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE* 77 (2), 257–286.
- Rodríguez, L. J., 1999. Anotación de corpora para diálogo. Technical Report BS12AV02, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad del País Vasco.
- Rodríguez, L. J., 2002. Cortado de señales y anotación de disfluencias en el Corpus Oral UAM. Documento interno, Grupo de Reconocimiento de Formas y Tecnologías del Habla, Departamento de Electricidad y Electrónica, Facultad de Ciencias, Universidad del País Vasco.
- Rodríguez, L. J., Torres, I., 2003. Comparative study of the Baum-Welch and Viterbi training algorithms applied to read and spontaneous speech recognition. In *Pattern Recognition and Image Analysis (IbPRIA 2003)*, F.J. Perales, A.J.C. Campilho, N. Pérez de la Blanca and A. Sanfeliú (Eds.), Springer-Verlag, Lecture Notes in Computer Science, LNCS 2652, Berlin Heidelberg, pp. 847–857.
- Rodríguez, L. J., Torres, I., Varona, A., 2000. Manual para el etiquetado de disfluencias. Technical Report BS12BV30, Proyecto TIC98-0423-C06: Sistema de diálogo para habla espontánea en un dominio semántico restringido, Universidad del País Vasco.
- Rose, R. C., Riccardi, G., 1999. Modeling disfluency and background events in ASR for a natural language understanding task. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1709–1712.
- Rosenberg, A. E., Lee, C. H., Soong, F., 1994. Cepstral channel normalization techniques for HMM-based speaker verification. In: *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*. pp. 1835–1839.
- Rosenfeld, R., Agarwal, R., Byrne, B., Iyer, R., Shriberg, M. L. L., Unverferth, J., Vergyri, D., Vidal, E., 1995. Error Analysis and Language Modeling for Conversational Speech: Team Report. In: *Proceedings of the Johns Hopkins University Language Modeling Summer Workshop*.
- Sankar, A., Beaufays, F., Digalakis, V., 1995. Training data clustering for improved speech recognition. In: *Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH)*. Madrid, Spain, pp. 503–506.
- Schiel, F., Burger, S., Geumann, A., Weilhammer, K., 1998. The Partitur Format at BAS. In: *Proceedings of the International Conference on Language Resources and Evaluation*. Granada, Spain, pp. 1295–1301, <http://www.phonetik.uni-muenchen.de/Bas/BasFormatseng.html>.

- Segura, J. C., 1991. Modelos de Markov con cuantización dependiente para reconocimiento de voz. Ph.D. thesis, Departamento de Electrónica y Tecnología de los Computadores. Facultad de Ciencias. Universidad de Granada.
- Senia, F., 1997. Specification of speech database interchange format. Technical Report LE2-4001-SD1.3.1, SpeechDat, CSELT, Italy.
- Shinoda, K., Lee, C. H., 1997. Structural MAP speaker adaptation using hierarchical priors. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding. pp. 381–388.
- Shriberg, E. E., 1994. Preliminaries to a theory of speech disfluencies. Ph.D. thesis, University of California at Berkeley.
- Shriberg, E. E., 1999. Phonetic consequences of speech disfluency. In: Proceedings of the International Congress of Phonetic Sciences. Vol. 1. pp. 619–622.
- Shriberg, E. E., Bates, R. A., Stolcke, A., 1997. A prosody-only decision-tree model for disfluency detection. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Vol. 5. pp. 2383–2386.
- Shriberg, E. E., Stolcke, A., 1996. Word predictability after hesitations: A corpus-based study. In: Proceedings of the International Conference on Speech and Language Processing (ICSLP). pp. 1868–1871.
- Siohan, O., Myrvoll, T. A., Lee, C. H., 2002. Structural Maximum A Posteriori Linear Regression for fast HMM adaptation. *Computer, Speech and Language* 16 (1), 5–24.
- Sjölander, K., Beskow, J., 2000. WaveSurfer - An Open Source Speech Tool. In: Proceedings of the International Conference on Speech and Language Processing (ICSLP). Beijing, China.
- Sperberg-McQueen, C. M., Burnard, L. (Eds.), 1994. Guidelines for Electronic Text Encoding and Interchange (TEI P3). Vol. I. ACH/ACL/ALLC, Chicago, Illinois, USA, Ch. 2: A Gentle Introduction to SGML, pp. 13–36, online version: <http://etext.virginia.edu/bin/tei-tocs?div=DIV1&id=SG>.
- Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Rao Gadde, V. R., Plauché, M., Richey, C., Shriberg, E., Sönmez, K., Weng, F., Zheng, J., 2000. The SRI March 2000 Hub-5 Conversational Speech Transcription System. In: Proceedings of the NIST Speech Transcription Workshop. University of Maryland, <http://www.nist.gov/speech/publications/tw00/>.
- Stolcke, A., Shriberg, E. E., 1996. Statistical language modeling for speech disfluencies. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 405–408.
- Stolcke, A., Shriberg, E. E., Hakkani-Tür, D., Tür, G., 1999. Modeling the prosody of hidden events for improved word recognition. In: Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH). Vol. 1. pp. 307–310.
- Sundaram, R., Ganapathiraju, A., Hamaker, J., Picone, J., 2000. ISIP 2000 Conversational Speech Evaluation System. In: Proceedings of the NIST Speech Transcription Workshop. University of Maryland, <http://www.nist.gov/speech/publications/tw00/>.
- Thompson, H. S., McKelvie, D., 1997. Hyperlink semantics for standoff markup of read-only documents. In: SGML Europe. Barcelona, Spain, <http://www.ltg.ed.ac.uk/~ht/sgmleu97.html>.

- Tokhura, Y., 1987. A weighted cepstral distance measure for speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* 35 (10), 1414–1422.
- Torres, I., 1992. Selección de unidades subléxicas para la decodificación acústico-fonética del habla en castellano. Informe de investigación, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.
- Torres, I., Casacuberta, F., 1993a. Multiple codebook Spanish phone recognition using semicontinuous hidden Markov models. In: *Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH)*. pp. 1543–1546.
- Torres, I., Casacuberta, F., 1993b. Spanish phone recognition using semicontinuous hidden Markov models. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 515–518.
- Torres, I., Varona, A., 2000. An Efficient Representation of k -TSS Language Models. *Computación y Sistemas (Revista Iberoamericana de Computación)* 3 (4), 237–244.
- Torres, I., Varona, A., 2001. k -TSS Language Models in Speech Recognition Systems. *Computer, Speech and Language* 15 (2), 127–149.
- Torres, I., Varona, A., Casacuberta, F., 1994. Automatic segmentation and phone model initialization in continuous speech recognition. In: *Proceedings of the CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology*. Munich, Germany, pp. 286–289.
- Varona, A., 2000. Modelos k -explorables en sentido estricto integrados en un sistema de reconocimiento automático del habla. Ph.D. thesis, Departamento de Electricidad y Electrónica. Facultad de Ciencias. Universidad del País Vasco.
- Varona, A., Torres, I., 2003. Integrating High and Low Smoothed LMs in a CSR System. In *Progress in Pattern Recognition, Speech and Image Analysis*, A. Sanfeliú and J. Ruiz-Shulcloper (editors), Springer-Verlag, Lecture Notes in Computer Science, LNCS 2905, pp. 236–243.
- Varona, A., Torres, I., Casacuberta, F., 1995. Discriminative-transitional/steady units for Spanish continuous speech recognition. In: *Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH)*. pp. 1471–1474.
- Wegmann, S., McAllaster, D., Orloff, J., Peskin, B., 1996. Speaker normalization on conversational telephone speech. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 339–341.
- Welling, L., Kanthak, S., Ney, H., 1999. Improved methods for vocal tract normalization. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 761–764.
- Witten, I. H., Bell, T. C., 1991. The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory* 37 (4), 1085–1094.
- Woodland, P. C., 2001. Speaker Adaptation for Continuous Density HMMs: A Review. In: *Proceedings of the ISCA Workshop on Adaptation Methods for Speech Recognition*. Sophia Antipolis, France, pp. 11–19.

- XML, 2000. Edited by Tim Bray (Textuality and Netscape), Jean Paoli (Microsoft), C. M. Sperberg-McQueen (University of Illinois at Chicago and Text Encoding Initiative), and Eve Maler (Sun Microsystems, Inc.), Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation 6-October-2000, <http://www.w3.org/TR/2000/REC-xml-20001006>.
- XSLT, 1999. Edited by James Clark, Extensible Stylesheet Language Transformations (XSLT). Versión 1.0. W3C Recommendation 16-November-1999, <http://www.w3.org/TR/1999/REC-xslt-19991116>.
- Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 1999. The HTK Book (version 2.2). Entropic Speech Technology, <http://htk.eng.cam.ac.uk>.
- Young, S. J., Odell, J. J., Woodland, P. C., 1994. Tree-based state tying for high accuracy acoustic modeling. In: Proceedings of the ARPA Workshop on Human Language Technology. pp. 307–312.
- Zhan, P., Westphal, M., 1997. Speaker normalization based on frequency warping. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1039–1042.
- Zhang, Z., Furui, S., 2000. An online incremental speaker adaptation method using speaker-clustered initial models. In: Proceedings of the International Conference on Speech and Language Processing (ICSLP). Vol. 3. Beijing, China, pp. 694–697.