Distribución y modelado acústico de fenómenos de habla espontánea

Luis Javier Rodríguez Fuentes, Inés Torres Barañano

Departamento de Electricidad y Electrónica
Facultad de Ciencia y Tecnología
Universidad del País Vasco
luisja@we.lc.ehu.es

Resumen

El desarrollo de las tecnologías de reconocimiento automático del habla está permitiendo acometer tareas cada vez más complejas, que permiten a los usuarios interaccionar de manera espontánea. Precisamente, el habla espontánea contiene fenómenos acústicos, léxicos, sintácticos y pragmáticos que la distinguen de la lengua escrita, y que empeoran el rendimiento del reconocedor. En aplicaciones de diálogo que operan con habla espontánea es imprescindible estudiar la fenomenología asociada a esta modalidad del habla y plantear estrategias que permitan corregir los errores que tales fenómenos introducen, o incluso reconocerlos explícitamente con objeto de mejorar la comprensión de las intervenciones. En este trabajo se presentan, en primer lugar, la distribución de fenómenos en tres bases de datos en castellano, y en segundo lugar, los resultados de decodificación acústico-fonética obtenidos utilizando modelos acústicos explícitos para algunos de los fenómenos de habla espontánea.

1. Introducción

El desarrollo de las tecnologías de reconocimiento automático del habla está permitiendo acometer tareas cada vez más complejas. Entre las aplicaciones que están recibiendo más atención se encuentran los sistemas de diálogo hombre-máquina, bien para para acceder a información, bien para llevar a cabo tareas concretas como comprar un billete de tren o una entrada para un concierto. Estos sistemas operan habitualmente en dominios semánticos restringidos, pero permiten a los usuarios interactuar espontáneamente. Del rendimiento del reconocedor depende que las intervenciones de los usuarios sean comprendidas y por tanto, que el sistema reaccione adecuadamente.

Por otra parte, la lengua hablada espontánea —o simplemente, *habla espontánea*— no es un reflejo de la lengua escrita, sino que dispone de otros recursos, relacionados con la inmediatez de la comunicación y con la necesidad de elaborar y corregir el discurso en tiempo real. Por esta razón el habla 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 comprensión 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 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.

En este trabajo se presenta, en primer lugar, la distribución de fenómenos en tres bases de datos en castellano. Teniendo en cuenta la gran cantidad de fenómenos que, como veremos, aparecen en el habla espontánea, su modelado debería tener un gran impacto en el rendimiento del reconocedor. Dos de las bases de datos constan de diálogos hombre-máquina obtenidos en dos fases distintas del desarrollo de una aplicación de acceso a información sobre viajes en tren. La tercera es una colección de diálogos naturales entre personas, tomados de los medios de comunicación (radio y televisión), que permitirá comparar cuantitativamente el habla espontánea que se produce en interacciones hombre-máquina con la que se produce en circunstancias normales entre personas.

En segundo lugar, se presentan los resultados de decodificación acústico-fonética obtenidos utilizando modelos acústicos explícitos para algunos de los fenómenos de habla espontánea (ruidos, pausas y alargamientos), así como los obtenidos utilizando el conjunto de modelos acústicos de referencia, con objeto de evaluar la mejora que aportan los primeros.

2. Descripción de las bases de datos

2.1. INFOTREN-1

La base de datos que en adelante llamaremos INFOTREN-1 corresponde a una tarea de acceso automático a información —horarios, trayectos y precios de viajes en tren entre ciudades españolas—, mediante una interfaz hablada basada en diálogos abiertos con los usuarios [1]. Lógicamente, el sistema de diálogo no se encontraba operativo en un principio, por lo que INFOTREN-1 fue adquirida sobre un *simulacro* de sistema, lo que se conoce como estrategia del *Mago de Oz*. Esto significa que, independientemente de que determinadas partes del sistema estén realmente automatizadas, es un operador humano el que escucha los requerimientos del usuario, accede a la infor-

Este trabajo ha sido realizado con el apoyo de la Universidad del País Vasco, a través de la subvención a grupos de investigación (9/UPV 00224.310-13566/2001), y del MCYT, a través del proyecto TIC2002-04103-C03-02.

mación, toma las decisiones y define el contenido y hasta cierto punto la forma de la respuesta que ha de llegarle al usuario. Por su parte, 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, interpretando un guión predeterminado pero con absoluta libertad. En ocasiones el operador 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. 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 tienen una calidad aceptable. Las características de INFOTREN-1 aparecen resumidas en la Tabla 1.

2.2. INFOTREN-2

Esta base de datos se ha adquirido recientemente como continuación y, en cierto modo, mejora de INFOTREN-1. Debido a que no todos los módulos del sistema de diálogo tenían la madurez suficiente como para permitir una grabación completamente automatizada, se ha empleado la misma metodología de adquisición de INFOTREN-1, basada en un operador humano oculto a los usuarios.

INFOTREN-2 presenta dos diferencias importantes con respecto a INFOTREN-1. Por un lado, tanto en la definición de los guiones para los viajes de los usuarios, como en el diseño de la estrategia del operador humano y de la herramienta que éste debía manejar, se han simplificado las capacidades del sistema y se ha restringido el modo en que los usuarios pueden interaccionar, con objeto de aumentar la eficacia de la comunicación. Como consecuencia, los usuarios resuelven sus tareas en menos tiempo, por lo que tanto el número promedio de turnos de usuario por diálogo, como la duración promedio de cada turno, disminuyen con respecto a INFOTREN-1. Por otro lado, INFOTREN-2 contiene un mayor número de hablantes y diálogos, de ahí que el volumen de datos (señal, turnos, palabras y fonemas) sea también significativamente mayor. Sin embargo, tal como se observa en la Tabla 1, el vocabulario es sólo ligeramente mayor: 788 palabras en INFOTREN-1 y 839 en INFOTREN-2. Así, el número promedio de muestras por palabra pasa de 23.77 a 57.5, lo que a priori debería mejorar la robustez del modelo de lenguaje del reconocedor.

2.3. CORLEC-EHU-1

Esta base de datos se ha obtenido del reciclaje de una pequeña parte del Corpus Oral de Referencia de la Lengua Española Contemporánea de la Universidad Autónoma de Madrid [2], de la cual se cortaron las señales, se transformaron las transcripciones originales y se enriquecieron éstas con la anotación de fenómenos de habla espontánea [3]. El subcorpus resultante, al que nos referiremos en adelante como CORLEC-EHU-1, consta de 42 entrevistas, tomadas de radio y televisión mediante una grabadora analógica, y posteriormente digitalizadas a 16 kHz en formato lineal de 16 bits. Las condiciones ambientales y de canal son particularmente adversas. Asimismo, se observa una gran cantidad de solapamientos —fragmentos de señal en los que dos o más voces se superponen—, ya que se trata de conversaciones completamente espontáneas entre personas. Como queda reflejado en la Tabla 1, esta base de datos, aún con menos turnos y hablantes que INFOTREN-2, contiene más tiempo de señal y especialmente más palabras y fonemas. De hecho, el número promedio de turnos por diálogo es de 50.88, y

Tabla 1: Principales características de las tres bases de datos de habla espontánea utilizadas en este trabajo.

	INFOTREN-1	**INFOTREN-2**	**CORLEC-EHU-1**
F_S (kHz)	8	8	16
Entorno	Oficina	Oficina	Estudios de radio/TV
Canal	Línea telefónica	Línea telefónica	Grabadora analógica
# Diálogos	227	900	42
# Hablantes	75	225	105
# Turnos	1657	6277	2137
Duración (horas)	2.50	5.37	5.67
Turnos/Diálogo	7.30	6.97	50.88
Duración/Turno (segundos)	5.43	3.08	9.55
# Palabras	18734	48243	72462
Tamaño Vocabulario	788	839	8253
Muestras/Palabra	23.77	57.50	8.78
# *Fonemas*	75005	193076	277564

la duración promedio de cada turno de 9.55 segundos. Además, al tratarse de conversaciones libres (no enfocadas a una tarea), se observa una gran dispersión léxica, con un vocabulario de 8253 palabras. Esto hace que se tengan tan sólo 8.78 muestras por palabra como promedio, lo cual afectará negativamente a la calidad del modelo de lenguaje.

3. Distribución de los fenómenos de habla espontánea

El trabajo realizado con INFOTREN-1 permitió establecer un primer inventario, adecuado para diálogos hombre-máquina [4], que incluye sobre todo fenómenos acústicos: ruidos (externos o producidos por el propio hablante), pausas de silencio, pausas *habladas* y alargamientos de sonidos; fenómenos léxicos: palabras cortadas o pronunciadas de forma poco ortodoxa; fenómenos sintácticos: reformulaciones o correcciones; y fenómenos pragmáticos: marcadores de discurso. Este mismo inventario, sin modificaciones, ha sido utilizado para anotar INFOTREN-2. Sin embargo, en diálogos espontáneos entre personas la gama de fenómenos es más amplia. Por ello, en el caso de CORLEC-EHU-1 se han añadido afirmaciones guturales, solapamientos y marcas específicas para palabras que no forman parte de la lengua, como siglas (deletreadas total o parcialmente) y palabras extranjeras.

En la Tabla 2 se muestran las cuentas absolutas de fenómenos, así como el número promedio de fenómenos por cada 100 palabras. En primer lugar, si se comparan las columnas correspondientes a INFOTREN-1 e INFOTREN-2, se percibe que la proporción de fenómenos es notablemente inferior en este último caso (37.82 fenómenos cada 100 palabras en INFOTREN-1, frente a los 19.66 de INFOTREN-2). La mayor parte de esa diferencia es atribuible a que mientras que en INFOTREN-1 se cuentan 14.27 ruidos aislados por cada 100 palabras, en INFOTREN-2 se cuentan tan sólo 4.26. Sin embargo, hay otros 8 puntos de diferencia más estrechamente ligados al tipo de habla, que sólo pueden explicarse debido a las las características de los respectivos sistemas de adquisición. El sistema de adquisición de INFOTREN-1 tenía más capacidad operativa y cedía una parte de la iniciativa al usuario, quien por su parte se comportaba de forma casi completamente espontánea —teniendo en cuenta que interactuaba con una máquina. En cambio, el sistema de adquisición de INFOTREN-

Tabla 2: Número total de fenómenos anotados (#FHE) y promedio de fenómenos por cada 100 palabras (%FHE), para las tres bases de datos consideradas en este trabajo.

	INFOTREN-1		INFOTREN-2		CORLEC-EHU-1	
Fenómeno	**#FHE**	**%FHE**	**#FHE**	**%FHE**	**#FHE**	**%FHE**
Aspiración	1404	7.49	1160	2.40	3005	4.15
Ruido de labios	600	3.20	378	0.78	161	0.22
Tos	9	0.05	27	0.06	40	0.06
Ruido genérico	661	3.53	491	1.02	530	0.73
Pausa de silencio	753	4.02	1847	3.83	1945	2.68
Pausa hablada /a/	93	0.50	22	0.05	25	0.03
Pausa hablada /e/	546	2.91	881	1.83	800	1.10
Pausa hablada /m/	179	0.96	146	0.30	323	0.45
Pausa hablada sin identificar	210	1.12	108	0.22	616	0.85
Alargamiento	1019	5.44	1640	3.40	3638	5.02
Mala pronunciación	105	0.56	273	0.57	1013	1.40
Palabra cortada	95	0.51	226	0.47	212	0.29
Reformulación	545	2.91	676	1.40	2307	3.18
Marcador de discurso	865	4.62	1606	3.33	2963	4.09
Afirmación gutural	–	–	–	–	295	0.41
Sigla	–	–	–	–	36	0.05
Palabra extranjera	–	–	–	–	187	0.26
Solapamiento	–	–	–	–	1808	2.50

2 se ha diseñado para llevar la iniciativa casi en todo momento, guiando al usuario y formulando preguntas de respuesta inequívoca, lo cual reduce el grado de espontaneidad y, por tanto, la presencia de recursos ligados a esta modalidad del habla, especialmente los de tipo acústico (pausas y alargamientos: casi 15 por cada 100 palabras en INFOTREN-1, menos de 10 en INFOTREN-2) y reformulaciones (2.91 cada 100 palabras en INFOTREN-1, 1.40 en INFOTREN-2). Los fenómenos de tipo léxico (palabras cortadas y mal pronunciadas) tienen frecuencias relativas similares en ambas bases de datos (1.07 y 1.4, respectivamente), ya que no operan como recursos del habla espontánea, sino que aparecen como errores.

En cuanto a CORLEC-EHU-1, la presión cognitiva e interactiva de las entrevistas es mayor que la de los diálogos hombre-máquina, que en gran medida han sido planificados con antelación. Ello debería propiciar, al menos en teoría, la aparición de un mayor número de fenómenos. Sin embargo, en CORLEC-EHU-1 se cuentan 24.25 fenómenos cada 100 palabras —excluyendo solapamientos, afirmaciones guturales, palabras extranjeras y siglas—, lejos de los 37.82 de INFOTREN-1. Como antes, esta diferencia se debe sobre todo a la menor presencia de ruidos (14.27 por cada 100 palabras en INFOTREN-1, tan sólo 5.16 en CORLEC-EHU-1), pero también, y de forma muy significativa, a la menor presencia de fenómenos acústicos (pausas y alargamientos: 14.95 por cada 100 palabras en INFOTREN-1, 10.13 en CORLEC-EHU-1). Por último, la proporción de distorsiones léxicas y reformulaciones es mayor en CORLEC-EHU-1 que en INFOTREN-1 e INFOTREN-2. Esto puede atribuirse a que el habla natural entre personas no está tan planificada como los diálogos hombre-máquina, de manera que se producen más errores y, por tanto, más correcciones. Un análisis más detallado de los fenómenos anotados en CORLEC-EHU-1 puede encontrarse en [3].

Tabla 3: Ampliación del conjunto de unidades subléxicas para las bases de datos de habla espontánea: codificación interna y descripción del sonido.

Codificación interna	**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 /rr/
S	Alargamiento del sonido /s/
B	Pausa hablada de identidad acústica confusa o inclasificable
X	Sonido gutural de afirmación

4. Experimentos de decodificación acústico-fonética

Los datos presentados en la Tabla 2 muestran que, dependiendo de la base de datos, se producen entre 20 y 40 fenómenos de habla espontánea cada 100 palabras, de los cuales entre 15 y 30 son fenómenos que afectan de uno u otro modo a los modelos acústicos del reconocedor. Esto pone de manifiesto la importancia de modelar explícitamente dichos fenómenos.

4.1. El conjunto de unidades subléxicas

El conjunto básico de unidades subléxicas utilizado por nuestro reconocedor está formado por 23 unidades pseudo-fonéticas incontextuales definidas específicamente para el castellano, más una unidad de silencio. Este conjunto ha mostrado un comportamiento óptimo sobre habla leída frente a otros conjuntos de unidades incontextuales más amplios y detallados [5]. Recientemente se ha propuesto una ampliación del conjunto de unidades subléxicas que cubre los fenómenos de tipo acústico que aparecen en habla espontánea en castellano [6] (véase la Tabla 3).

Esta ampliación consta de los fenómenos acústicos —salvo los silencios, que ya estaban incluidos— y sonidos guturales de afirmación, 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. La solución propuesta en este trabajo trata de aprovechar la potencia de los Modelos Ocultos de Markov (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 por la que implícitamente se opta con el conjunto básico de unidades.

No siempre se tendrán muestras suficientes para entre-

Tabla 4: Particiones de entrenamiento (E) y test (T) definidas para INFOTREN-1 e INFOTREN-2. En el caso de CORLEC-EHU-1 se muestran los datos correspondientes a los tres bloques definidos, dos de los cuales se utilizan como corpus de entrenamiento y el restante como corpus de test, de modo que se aplican sucesivamente tres particiones distintas.

	INFOTREN-1		INFOTREN-2		CORLEC-EHU-1		
	E	T	E	T	C1	C2	C3
# Diálogos	191	36	720	180	14	14	14
# Hablantes	63	12	180	45	37	35	33
# Turnos	1349	308	4928	1349	700	690	690
# Tramos	703719	182722	1517322	415367	680538	683142	672884
# *Fonemas*	61611	13394	150688	42388	89851	91685	96028

nar 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*.

4.2. Definición de los corpus de entrenamiento y test

INFOTREN-1 e INFOTREN-2 han sido divididas cada una en dos bloques independientes, uno de ellos utilizado para estimar (*entrenar*) los modelos acústicos y el otro para llevar a cabo los experimentos de decodificación acústico-fonética (DAF). Por otra parte, CORLEC-EHU-1 se ha dividido en tres bloques de 14 entrevistas (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. Esto es importante en el caso de CORLEC-EHU-1, porque las condiciones ambientales y de canal varían mucho de unas entrevistas a otras. En la Tabla 4 se muestran los datos de cada uno de los bloques definidos.

4.3. Condiciones experimentales

Como parámetros acústicos se han utilizado los coeficientes cepstrales con banco de filtros en escala Mel (*Mel-Filter Cepstral Coefficients*, MFCC), calculados en tramos de 25 milisegundos, cada 10 milisegundos. También se han calculado las primeras y segundas derivadas de los cepstrales (ΔMFCC y Δ^2MFCC), así como la energía (E) de cada tramo y su derivada (ΔE). A continuación se han definido 4 representaciones acústicas distintas: MFCC, ΔMFCC, Δ^2MFC y (E,ΔE). Para poder entrenar modelos acústicos discretos, se ha aplicado un algoritmo de cuantificación vectorial para obtener 4 diccionarios de 256 centroides, y se han etiquetado los vectores acústicos con el índice del centroide más cercano.

Cada unidad subléxica se ha representado mediante un MOM izquierda-derecha de 3 estados. La probabilidad de emisión en cada estado del MOM y en cada instante t se calcula como el producto de las probabilidades obtenidas en las 4 representaciones acústicas. Los MOM discretos se han estimado utilizando el algoritmo de Baum-Welch general, a partir de modelos equiprobables. Los MOM continuos se han inicializado a partir de los MOM discretos óptimos y se han reestimado utilizando el algoritmo de Viterbi (véase [7]).

Tabla 5: Tasas de DAF obtenidas sobre INFOTREN-1, INFOTREN-2 y CORLEC-EHU-1 (promedio para las 3 particiones), utilizando los conjuntos básico (CB) y ampliado (CA) de unidades subléxicas, con MOM discretos y MOM continuos 32 gaussianas.

	MOMd		MOMc32g	
	CB	CA	CB	CA
INFOTREN-1	51.98	55.81	56.86	61.34
INFOTREN-2	50.89	52.09	57.22	59.26
CORLEC-EHU-1	45.60	46.78	50.83	52.06

Durante la búsqueda de la decodificacion óptima no se han aplicado restricciones fonológicas. La probabilidad de transitar de un modelo acústico a cualquier otro (salvo a sí mismo) es $w = 1/(H-1)$, donde H es el número de modelos acústicos. 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 y las unidades específicas del habla espontánea. 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:

$$\%DAF = \frac{a}{a+s+b+i} * 100 \ . \qquad (1)$$

4.4. Resultados

En la Tabla 5 se muestran los resultados de DAF obtenidos para las tres bases de datos consideradas en este trabajo. En todos los casos se observa cómo la modelización explícita de los fenómenos de habla espontánea produce mejoras en la tasa de reconocimiento, especialmente en el caso de INFOTREN-1. En el caso de INFOTREN-2 y CORLEC-EHU-1, las mejoras son más pequeñas. Esto es debido, por un lado, a que los fenómenos del conjunto ampliado se producen en estas bases de datos en una proporción mucho menor que en INFOTREN-1 (véase la Tabla 2), lo cual limita la mejora que potencialmente puede suponer la introducción de modelos explícitos para este tipo de fenómenos. Por otra parte, en el caso de CORLEC-EHU-1, 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.

5. Conclusiones

El estudio de las cuentas absolutas y relativas de fenómenos de habla espontánea en tres bases de datos de diálogos en castellano permite concluir que es necesario integrar este tipo de fenómenos en el diseño del reconocedor, tanto en lo que se refiere a los modelos acústicos como en lo que respecta al modelo de lenguaje. Se han contabilizado más de 37 fenómenos cada 100 palabras en INFOTREN-1, casi 20 en INFOTREN-2 y más de 24 en CORLEC-EHU-1. De ellos, más de 15 (más de 30 en el caso de INFOTREN-1) afectan directa o indirectamente a los modelos acústicos. Se han contabilizado, además, entre 1.4 y 3.2 reformulaciones cada 100 palabras, que pueden afectar seriamente a la estructura sintáctica y, por tanto, a la correcta

comprensión de las intervenciones. Por último, se han modelado explícitamente los fenómenos de tipo acústico y se ha comprobado que su utilización en experimentos de decodificación acústico-fonética mejora el rendimiento de forma significativa.

6. Referencias

[1] A. Bonafonte, P. Aibar, N. Castell, E. Lleida, J.B. Mariño, E. Sanchís, I. Torres, "Desarrollo de un sistema de diálogo oral en dominios restringidos", I Jornadas en Tecnología del Habla, Sevilla, 6-10 de noviembre de 2000.

[2] Corpus Oral de Referencia de la Lengua Española (1992), http://www.lllf.uam.es /corpus/corpus_oral.html.

[3] Luis Javier Rodríguez, Inés Torres, Amparo Varona, "Anotación de disfluencias en un corpus de habla espontánea no específico", II Jornadas en Tecnología del Habla, Granada, 16-18 de diciembre de 2002.

[4] Luis Javier Rodríguez, Inés Torres, Amparo Varona, "Annotation and analysis of disfluencies in a spontaneous speech corpus in Spanish", Proceedings of the ISCA Workshop on Disfluency in Spontaneous Speech, pp. 1–4, University of Edinburgh, Scotland, August 29-31, 2001.

[5] I. Torres, "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, 1992.

[6] Luis Javier Rodríguez, Inés Torres, Amparo Varona, "Evaluation of sublexical and lexical models of acoustic disfluencies for spontaneous speech recognition in Spanish", Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH), Aalborg (Denmark), September 2-7, 2001.

[7] Luis Javier Rodríguez, Inés Torres, "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, pages 847-857, Berlin Heidelberg, 2003.