LLISTERRI, J. - MACHUCA, M. J.- de la MOTA, C.- RIERA, M.- RÍOS, A. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación*. Barcelona: Ariel (Ariel Lingüística). pp. 209-243. ISBN: 84-344-8255-X

http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

8. Entonación y tecnologías del habla

JOAQUIM LLISTERRI, MARÍA JESÚS MACHUCA, CARME DE LA MOTA, MONTSERRAT RIERA, ANTONIO RÍOS

8.1. Tecnologías del habla y entonación

Al igual que sucede en el campo de la fonética, la entonación y, en general, los aspectos prosódicos de la lengua, son uno de los temas que en la actualidad despiertan un mayor interés entre los especialistas en tecnologías del habla. La síntesis, el reconocimiento y los sistemas de diálogo –es decir, las tres áreas básicas que configuran las **tecnologías del habla**– requieren información prosódica para alcanzar sus objetivos. En el caso de la síntesis y, muy especialmente, de la conversión de texto en habla, la entonación es imprescindible para transformar automáticamente un texto escrito en su equivalente sonoro, mientras que en el reconocimiento es también útil considerar la entonación a la hora de llegar de un enunciado oral a su forma escrita; parece claro, finalmente, que los sistemas de diálogo, que permiten la interacción hablada entre una persona y un ordenador, necesitan procesar la entonación tanto para la mejor comprensión de las intervenciones del usuario como en la generación automática de una respuesta adecuada.

El interés que mencionábamos al principio se pone de manifiesto, por ejemplo, en la organización de encuentros internacionales como los seminarios *Prosody in Speech Recognition and Understanding* (Red Bank, New Jersey, EEUU, octubre de 2001) o Prosody 2000: Speech Recognition and Synthesis (Krakow, Polonia, octubre de 2000), la inclusión de una sesión plenaria sobre prosodia y tecnologías del habla en el congreso *Speech Prosody 2002* (Aix-en-Provence, Francia, abril de 2002), o el hecho de que, en este evento, el 15% de las comunicaciones presentadas abordaran diversos aspectos prosódicos de la síntesis, el reconocimiento o el diálogo persona-máquina. Es interesante también señalar que, en el último congreso de la *European Speech Communication Association*[1] (*Eurospeech 2001*, Aalborg, Dinamarca, septiembre de 2001), un 11% de las sesiones tuvieran como tema monográfico la prosodia.

Esta tendencia muestra también que los expertos en tecnologías del habla, tradicionalmente ligados a la ingeniería de telecomunicación o a la informática y alejados del mundo de la lingüística, son cada vez más conscientes de los beneficios que se derivan de la incorporación de conocimientos lingüísticos a la síntesis, al reconocimiento o al diálogo, especialmente en cuanto se necesita refinar la calidad de las aplicaciones (Pols 2001)[2].

[1] Actualmente *International Speech Communication Association* (ISCA).

[2] La incorporación de conocimientos lingüísticos, especialmente desde la perspectiva de la fonética, a las tecnologías del habla se aborda también, por ejemplo, en Aguilar *et al*. (1999), Llisterri *et al.* (1999) y en Llisterri (2002).

Sin embargo, no deja de ser cierto que los logros alcanzados mediante técnicas estadísticas que parten, en general, del aprendizaje a partir de corpus, pueden en ocasiones reforzar las antiguas concepciones sobre la escasa rentabilidad de los lingüistas en un entorno orientado al desarrollo de productos para el mercado. Esto se hace especialmente patente en el caso de la síntesis, cuando se contrasta la buena calidad que se obtiene mediante los recientes métodos comerciales de conversión de texto en habla fundamentados en el uso de grandes corpus con la menor naturalidad de los resultados que proporcionan las aproximaciones tradicionales basadas en reglas que incorporan conocimiento fonético (Mixdorff 2002).

En lo que se refiere a los productos comerciales de reconocimiento del habla, aunque autores como Mixdorff (2002) han señalado la relativa utilidad de la información prosódica cuando se trata con vocabularios restringidos (por ejemplo, en los sistemas para el reconocimiento de números), es evidente la importancia de este tipo de información en la aplicación del reconocimiento a los sistemas de diálogo, ya que ayuda a identificar actos de habla, focalizaciones, cambios de tema y otros fenómenos pragmáticos propios de la interacción comunicativa.

Parece, pues, que pese al convencimiento general de que un buen conocimiento de la prosodia de las lenguas puede aportar mejoras a las aplicaciones que se desarrollan en el campo de las tecnologías del habla, en la práctica cotidiana los sistemas comerciales se limitan a utilizar aquellas técnicas que, con o sin conocimiento lingüístico, permiten obtener mejores resultados.

Investigadores como Mixdorff (2002) han señalado con acierto que la falta de un modelo prosódico unificado es una de las razones por las que el conocimiento prosódico no se ha incorporado plenamente a los sistemas comerciales de síntesis y, especialmente, a los de reconocimiento. En efecto, una revisión de compilaciones recientes como Horne (2000) o Botinis (2000, 2001), así como este propio volumen, ponen claramente de manifiesto la multiplicidad de enfoques en la descripción prosódica de las lenguas.

Esta situación ha llevado a ver en las tecnologías del habla un banco de pruebas en el que validar diversas concepciones teóricas, y a partir del cual pueden también refinarse las descripciones específicas de cada lengua (Pols 2001). Aunque no se pretende defender aquí que los únicos parámetros para evaluar un modelo prosódico sean la calidad que permite obtener en la síntesis o la mejora que produce en el reconocimiento, no cabe duda de que éstos son criterios empíricos que adquieren pleno sentido cuando se concibe la investigación en prosodia como una actividad encaminada a responder a necesidades del mundo real.

En conjunto, nos hallamos en una situación en la que no deja de manifestarse una cierta disociación entre la teoría y la práctica en lo que a la integración de los conocimientos prosódicos en las tecnologías del habla se refiere. Esta circunstancia, que en principio podría interpretarse como un obstáculo, debería constituir, a nuestro modo de ver, un estímulo para la realización de trabajos que combinen de un modo acertado los planteamientos de las teorías de la entonación con las necesidades derivadas de su implementación en aplicaciones reales.

Tomando como punto de partida las consideraciones anteriores, este capítulo se propone presentar el modo en que el conocimiento prosódico, y en especial el relacionado con la entonación, se ha incorporado a los sistemas de conversión de texto en habla (apartado 8.3), de reconocimiento automático del habla (apartado 8.4) y de diálogo persona-máquina (apartado 8.5). Puesto que para el desarrollo de aplicaciones es importante contar con un corpus adecuadamente anotado, en la primera parte del trabajo se revisan algunos corpus prosódicos así como los principales sistemas de anotación y etiquetado (apartado 8.2). Como conclusión (apartado 8.6), se intentará mostrar que una visión de las tecnologías del habla que pretenda ir algo más allá del producto inmediato debe, necesariamente, plantearse con rigor el desarrollo de modelos de base lingüística, en especial en el ámbito de la prosodia.

8.2. Los corpus prosódicos

8.2.1. CORPUS PROSÓDICOS PARA EL DESARROLLO DE APLICACIONES EN EL ÁMBITO DE LAS TECNOLOGÍAS DEL HABLA

Un **corpus prosódico** puede definirse como un conjunto de realizaciones orales de uno o varios locutores diseñado para estudiar y modelar el comportamiento de los correlatos acústicos de los elementos suprasegmentales (duración, frecuencia fundamental –F0– e intensidad) en relación con determinados fenómenos fonológicos (el acento, la entonación o el ritmo), sintácticos (la modalidad oracional, la agrupación en constituyentes) o pragmáticos (la focalización, el tema del enunciado, la intencionalidad del hablante, el tipo de acto de habla).

Para la conversión de texto en habla, tal como se expone con detalle en el apartado 8.3, es útil disponer de un corpus a partir del cual extraer modelos de duración de cada uno de los alófonos, patrones melódicos que puedan aplicarse en función de las características fonéticas, sintácticas e, idealmente, semánticas y pragmáticas del enunciado, y modelos de variación de intensidad tanto de cada alófono como de todo el enunciado. Suelen utilizarse en general corpus producidos por un único locutor, ya que los datos prosódicos se superponen a las unidades de síntesis, extraídas también de un solo hablante. Simplificando mucho, podría decirse que se trata de reproducir lo más fielmente posible el comportamiento fonético de un hablante tanto en lo segmental como en lo suprasegmental.

En cambio, por la propia naturaleza de sus aplicaciones el reconocimiento del habla requiere que los corpus contengan datos procedentes del mayor número posible de hablantes, de modo que los sistemas puedan enfrentarse a la variación existente entre locutores. Esto se debe a que un reconocedor es, en esencia, un sistema de comparación entre los modelos aprendidos durante la fase de entrenamiento y el análisis que realiza el sistema del enunciado que debe reconocer. Por ello, un corpus prosódico que se utilizara para entrenar un sistema de reconocimiento debería cubrir el mayor número posible de factores de variación y reflejar el comportamiento de un elevado número de locutores. En el caso de reconocedores orientados a una tarea concreta, que puede ir desde el reconocimiento de números de teléfono hasta proporcionar información a un usuario o reservar un billete, el corpus usado en el entrenamiento suele reflejar esta tarea.

Un buen ejemplo de corpus prosódico multilingüe lo constituye el desarrollado en el marco del proyecto MULTEXT (Campione y Véronis 1998), que comprende más de cuatro horas de grabación de textos leídos en cinco lenguas (francés, inglés, italiano, alemán y español[3]) por diez locutores en cada lengua. Los datos recogidos para cada enunciado incluyen la frecuencia fundamental, el contorno melódico estilizado, la anotación de los movimientos tonales mediante el sistema de transcripción prosódica INTSINT (véase el capítulo 5), y la sincronización entre la señal sonora y las fronteras de palabras en la transcripción ortográfica.

En Gurlekian *et al*. (2001) se describe un corpus prosódico del español concebido para obtener los datos que se requieren en un sistema de conversión de texto en habla. El corpus consta de 741 oraciones, diseñadas teniendo en cuenta la modalidad oracional (declarativa e interrogativa), la complejidad sintáctica y la aparición de las palabras más frecuentes de la lengua; se consideró también la aparición de las sílabas más frecuentes, así como su posición en la palabra (inicial, medial y final) y su acentuación. Dos locutores profesionales realizaron la grabación, y cuatro personas convenientemente formadas etiquetaron el corpus usando una adaptación del sistema de transcripción prosódica ToBI (véase el capítulo 7) al español de Argentina. Con los datos procedentes de la transcripción segmental, la anotación en ToBI y la alineación entre las etiquetas para cada uno de los alófonos y su inicio y final en la onda sonora se creó una base de datos a partir de la cual se pueden generar reglas de síntesis o entrenar automáticamente un sistema de conversión de texto en habla.

Para el catalán se ha desarrollado también un corpus prosódico orientado a la extracción de conocimiento fonético aplicable a sistemas de conversión de texto en habla como parte de las actividades del CREL *–Centre de Referència en Enginyeria Lingüística de la Generalitat de Catalunya* (Riera y Jiménez 2000)–. El corpus contiene representados los elementos segmentales del catalán teniendo en cuenta su posición en el enunciado y en la palabra, el contexto anterior y posterior (definido en función del lugar y modo de articulación y de la sonoridad), así como la acentuación en el caso de las vocales, con lo que pueden obtenerse modelos fiables de duración segmental. Para la creación de modelos entonativos, se consideran como variables el número de oraciones por párrafo, la posición de la oración en el párrafo, la complejidad sintáctica, la modalidad y el número de grupos fónicos de cada oración; se tienen igualmente en cuenta la posición en la frase de los grupos de entonación, el tipo de límite sintáctico que les antecede o precede y el número de sílabas que contienen; finalmente, en el diseño del corpus se han considerado también los grupos acentuales, buscando la representatividad en lo que se refiere al número de sílabas y a su posición en el grupo entonativo. En conjunto, el corpus consta de 114 textos leídos por dos locutores. La anotación se ha llevado a cabo marcando el principio y el final de cada uno de los elementos segmentales, así como los límites entre sílabas, entre grupos entonativos y entre grupos acentuales. Los fenómenos tonales se han anotado fundamentalmente mediante INTSINT, aunque una muestra del corpus se ha etiquetado con ToBI.

[3] Posteriormente se desarrolló una versión en catalán, que se presenta en Estruch (2000).

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

8.2.2. ANOTACIÓN PROSÓDICA PARA EL DESARROLLO DE APLICACIONES EN TECNOLOGÍAS DEL HABLA

Para que un corpus sea realmente útil, es un requisito imprescindible que esté adecuadamente anotado. Por "**anotación**" se entiende el marcado mediante etiquetas asociadas a un determinado punto de la señal sonora de aquellos fenómenos que interesan al investigador; este mismo proceso se describe también en ocasiones como "etiquetado" o incluso como "transcripción".

En los capítulos anteriores se han revisado los diversos sistemas utilizados para anotar la curva melódica y, como es lógico, algunos de ellos –ToBI e INTSINT, por ejemplo– se han empleado para señalar los fenómenos prosódicos en corpus orientados al desarrollo aplicaciones en el campo de las tecnologías del habla.

Un primer problema al que se enfrenta la **anotación prosódica** es la ausencia de un sistema considerado estándar y aceptado de modo unánime por la comunidad científica: podría decirse que no existe en prosodia el equivalente a SAMPA, el alfabeto más comúnmente utilizado para la transcripción segmental en tecnologías del habla (Wells 2002). Aunque existe un consenso sobre la necesidad de disponer de una herramienta que facilitara el intercambio y la comparación de los datos anotados mediante los distintos sistemas existentes, la definición de un sistema común no estaría exenta de dificultades, como ponen de manifiesto Quazza y Garrido (1998).

Por este motivo, el método de anotación elegido suele depender de la orientación teórica del investigador y de las propias necesidades del proyecto. Esta es la razón por la que se utilizan, en ocasiones, esquemas de anotación prosódica desarrollados para describir un determinado nivel de análisis lingüístico o para reflejar fenómenos relevantes en determinadas aplicaciones, como pueda ser el caso de la interacción persona-máquina en los sistemas de diálogo.

Por otra parte, dado el tamaño de los corpus requerido para la síntesis y, muy especialmente, para el reconocimiento, la anotación debe poder automatizarse al máximo. En este sentido, INTSINT (véase el capítulo 5), por ejemplo, dispone de un conjunto de herramientas que proporcionan una primera anotación automática de la curva melódica con un alto grado de fiabilidad (Astésano *et al*. 1997, Campione y Véronis 2000 y 2001, Hirst 2002), lo que representa una importante ventaja frente a otros procedimientos que requieren anotación manual.

En conjunto, la constitución de corpus prosódicos multilingües para la síntesis y el reconocimiento, junto con la creación de herramientas que permitan su anotación de forma automática, es una de las áreas que seguramente experimentará un mayor crecimiento en los próximos años, principalmente a causa de la necesidad, por una parte, de disponer de descripciones fonéticas detalladas y, por otra, de desarrollar de la manera más rápida y económica posible sistemas que puedan incorporarse a diversas aplicaciones comerciales.

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

8.3. La conversión de texto en habla

La conversión de texto en habla es, sin duda, el ámbito de las tecnologías del habla en el que más atención se ha prestado hasta ahora a los aspectos prosódicos. Por ello, en este apartado se presentan, en primer lugar, algunas consideraciones generales sobre el tratamiento de la prosodia (8.3.1), para describir a continuación el uso de diversos modelos entonativos en la conversión de texto en habla (8.3.2), y abordar después aquellos modelos que se han desarrollado específicamente para la síntesis (8.3.3).

8.3.1. PROSODIA Y CONVERSIÓN DE TEXTO EN HABLA

Un sistema de **conversión de texto en habla** (TTS, *text-to-speech system*) tiene como finalidad la transformación automática de cualquier texto escrito y disponible en formato electrónico en su correspondiente realización sonora. Un conversor es una aplicación informática que ha de reproducir el proceso que realizaría el lector de un texto para emitirlo en voz alta; por tanto, se le deberá dotar de toda la información lingüística que sea necesaria para ello[4].

Las investigaciones en informática y en lingüística se han centrado en los últimos años en lograr una mayor naturalidad en el habla sintetizada mediante el perfeccionamiento de dos aspectos esenciales en el proceso de síntesis: la concatenación segmental y la prosodia.

La mejora en la calidad de la síntesis segmental, que inicialmente se realizaba por reglas (Klatt 1980 y 1987), se ha conseguido con el desarrollo de la síntesis del habla concatenada en el dominio del tiempo (*Time Domain Synthesis*). Esta técnica, también conocida como síntesis basada en corpus, requiere la formación de grandes bases de datos fonéticas y un análisis con el que se identifican las partes constituyentes de habla (segmentos, sílabas, palabras). Durante la fase de síntesis, se seleccionan las unidades (difonos, polífonos, palabras enteras) de la base de datos y se concatenan –es decir, se unen– para formar los enunciados. En el modelado prosódico, se modifica la duración de los segmentos y la melodía, y a veces también la intensidad. Mediante este procedimiento se genera un habla sintetizada cercana a la producida por un locutor humano, puesto que las porciones de habla que forman un nuevo enunciado son segmentos de habla natural previamente almacenados.

La importancia de la prosodia para alcanzar una mayor naturalidad en el habla sintetizada se ha puesto de manifiesto en los resultados obtenidos por el proyecto europeo COST 258, en el que participaron 15 países europeos, y cuyos resultados se recogen en Keller *et al*. (2002).

Un conversor de texto en habla está formado por diversos módulos, cada uno de los cuales se ocupa de una etapa en el proceso de transformación de un texto escrito en su forma sonora. El **módulo prosódico** contiene un conjunto de reglas que especifican, fundamentalmente, la

[4] El proceso de conversión de texto en habla se describe detalladamente en Dutoit (1997). Para una presentación general véase, por ejemplo, Llisterri (2001) y Olive (1998).

duración y la intensidad de los sonidos, la colocación y la duración de las pausas, el contorno melódico del enunciado y las modificaciones producidas por el acento.

En lo que se refiere a la **duración** segmental, los estudios fonéticos han puesto de manifiesto que está condicionada por diversos factores: el acento, la longitud de la palabra en que se encuentra el segmento analizado, la consonante o vocal que le sigue, la existencia de una pausa tras el segmento, su posición en el enunciado y la velocidad de elocución del hablante (Klatt 1976). Un modelo de duración segmental para la síntesis debería tener en cuenta, por tanto, todos estos factores, además de contemplar la estructuración rítmica de la lengua.

En cuanto a la **intensidad**, aunque muchos conversores de texto en habla no consideran este aspecto, es sabido que varía según la posición del segmento respecto a las pausas (prepausal/no prepausal), en el enunciado (inicial, medial o final), la acentuación y la longitud de la secuencia (Blecua y Acín 1995).

La psicolingüística ha puesto de relieve la importancia de las **pausas** para la estructuración y comprensión del enunciado. Un sistema de conversión de texto en habla deberá insertar las pausas marcadas ortográficamente, con la duración apropiada según el tipo de signo de puntuación, y las no marcadas haciéndolas coincidir con el límite de un grupo prosódico, como haría un buen lector. Además, se deben determinar los factores fonéticos y sintácticos que determinan la inserción de una pausa no marcada (Puigví *et al*. 1994).

El módulo prosódico de un conversor de texto en habla se ocupa también de establecer los valores de los puntos de F0 que constituyen la **curva melódica**, empleando básicamente tres estrategias (Beaugendre 1996): diseñar un sistema de reglas que partan de un conjunto de símbolos, reproducir un conjunto de patrones melódicos previamente almacenados, y recurrir a métodos estadísticos (Modelos Ocultos de Markov o redes neuronales). Las dos primeras estrategias elaboran modelos entonativos de las lenguas a partir de estudios fonéticos. La aproximación estadística, en cambio, predice la curva entonativa de un enunciado a partir de los datos almacenados en un corpus de habla; evidentemente, para que los fenómenos poco frecuentes aparezcan suficientemente representados se necesita un corpus de gran extensión.

En cualquier caso, para la generación de las curvas entonativas se requiere información sintáctica –por ejemplo, la modalidad oracional–, semántica –por ejemplo, la desambiguación de frases– y pragmática –por ejemplo, la presencia de elementos focalizados–. Hay que tener en cuenta, además, que los movimientos de F0 en el enunciado dependen también de otros fenómenos prosódicos, como las pausas, la duración de los segmentos o la intensidad.

8.3.2. MODELOS ENTONATIVOS APLICADOS A LA CONVERSIÓN DE TEXTO EN HABLA

Como se ha señalado anteriormente, una de las alternativas para mejorar la naturalidad del habla sintetizada ha sido la aplicación de modelos entonativos desarrollados con finalidades diferentes a la conversión de texto en habla. Nos centraremos aquí en tres de los modelos descritos en otros

capítulos de este libro –ToBI, INTSINT y el modelo del IPO–, presentando también la aplicación a la síntesis de los modelos entonativos de Fujisaki, Mertens y Thorsen-Grønnum.

8.3.2.1. ToBI (Tone and Break Indices)

El Modelo de la Secuencia de Tonos (ToBI, *Tone and Break Indices*) fue introducido en 1980 por Pierrehumbert para el inglés americano, aunque el trabajo de Bruce (1977) se considera un antecedente a la propuesta inicial de este modelo. En el estándar ToBI descrito por Silverman *et al*. (1992), el contorno entonativo se concibe en su conjunto como una secuencia de fenómenos tonales discretos (véanse los capítulos 6 y 7). Tal y como indican Kochanski y Shih (2001), a partir de un corpus etiquetado mediante ToBI pueden seguirse diversas estrategias de descodificación y generación de los contornos de F0. Por un lado, existe la opción de formular un conjunto de reglas que permita derivar los valores de frecuencia que darán lugar a un contorno melódico a partir de la información anotada, y por otro, existe también la posibilidad de conseguir los valores requeridos empleando únicamente métodos que permitan la automatización mediante el entrenamiento de los sistemas.

El modelo para la predicción de la F0 cuyo desarrollo inician en Suecia en 1997 Filipsson y Bruce está basado en reglas y, podría decirse, de acuerdo con Frid (1999 y 2001), que su estilo es semejante al de ToBI. Se emplean puntos de inflexión que se representan con niveles altos y bajos, y el contorno de F0 se obtiene interpolando líneas rectas entre los puntos ya establecidos[5].

Por otro lado, en trabajos como el de Anderson *et al*. (1984) o el de Jilka *et al*. (1999), se han propuesto también reglas capaces de describir los contornos de F0 con los valores de frecuencia correspondientes. En el modelo de Jilka, validado para el inglés americano mediante resíntesis y pruebas de percepción, se emplea la interpolación a partir de las etiquetas proporcionadas por la notación ToBI para acentos y marcas de frontera. Se emplean además dos líneas de referencia, una alta y otra baja, que permiten acotar el rango, y que se derivan a partir del locutor mediante un árbol CART, un procedimiento que resulta de ayuda a la hora de aplicar los sistemas de etiquetado a la síntesis[6]. El conjunto de reglas diseñadas se ha implementado finalmente en el sistema Festival, descrito en el apartado 8.3.3.1. También existen otros trabajos en los que se utiliza una versión modificada de ToBI combinada con modelos estocásticos para la predicción de etiquetas entonativas y para la síntesis de la entonación (Black y Hunt 1996, Ross y Ostendorf 1999).

Para poder valorar las posibilidades de tales sistemas de síntesis contamos con los resultados de algunos experimentos. El trabajo de Syrdal *et al.* (1998), en el que se realiza la evaluación de la

---

[5] Frid está actualmente introduciendo mejoras este modelo, en la línea de los trabajos de Black y Hunt (1996) y de Dusterhoff (2000).

[6] CART (*Classification and Regression Tree,* Árbol de Clasificación y Regresión) es una técnica estadística utilizada para clasificar automáticamente un conjunto de datos y obtener una serie de reglas que permitan, a su vez, predecir el comportamiento de nuevos datos.

asignación de las etiquetas de ToBI, nos muestra que "como era de esperar, los modelos que usaban una información de entrada más precisa (etiquetas como en ToBI con el tipo de acento y también con su posición) permitían generar contornos que eran más aceptables para los oyentes que los que sólo empleaban como entrada la posición del acento. Esto no implica que los modelos que únicamente aceptan como entrada la información sobre la posición del acento sean menos útiles [...] Sin embargo, cuando se dispone de la información sobre el tipo de acento y su posición, como es el caso de los sistemas de síntesis a partir de conceptos, parecería razonable y útil usar ambos datos" (Syrdal *et al.* 1998: 6).

En cuanto a futuros trabajos, Wightman (2002), por ejemplo, ha insistido en la necesidad de estandarizar la notación, ya que en algunos estudios se reduce el inventario de índices o se modifican las etiquetas. Sería conveniente también agilizar o automatizar el proceso de etiquetado y comprobar nuevamente los resultados obtenidos en el ámbito de la percepción.

8.3.2.2. INTSINT

En el *Centre National de la Recherche Scientifique* de la *Université de Provence,* en Aix-en-Provence, se ha desarrollado un modelo de entonación que comprende cuatro niveles de representación: acústico, fonético, fonológico superficial y fonológico profundo (véase el capítulo 5). Actualmente, existen para la síntesis dos versiones de este sistema, una basada en reglas (Di Cristo *et al*. 1997 y 2000) y otra estadística (Courtois *et al*. 1997 y Véronis *et al*. 1998), que se han integrado en *ProZed* (Hirst 2000), un editor prosódico multilingüe para la síntesis del habla.

El sistema basado en reglas (SYNTAIX)[7] está constituido por dos módulos, uno lingüístico y otro fonético. El módulo lingüístico consta, a su vez, de dos submódulos: uno de procesamiento de lenguaje natural y otro fonológico o prosódico. El primero ha sido desarrollado principalmente en el marco del proyecto MULTEXT (Véronis *et al.* 1994) y se encarga de tratar el texto de entrada y realizar la división en elementos léxicos, el reconocimiento de enunciados, la asignación de categorías gramaticales a las unidades léxicas, la conversión de grafemas en fonemas, y el análisis sintáctico en el nivel de frase. A continuación, el módulo fonológico o prosódico toma como entrada el texto tratado por el submódulo de procesamiento del lenguaje y crea una representación abstracta del ritmo y de la entonación de los enunciados a partir del cálculo de los niveles de prominencia, de la asignación de los patrones tonales, de la derivación de la representación superficial y de la asignación de los símbolos de duración: N (normal), E (expandido), X (extra expandido), R (reducido).

Una vez tratado el texto en el módulo lingüístico, el módulo fonético asocia los datos acústicos a las etiquetas simbólicas asignadas en el módulo anterior, realizando el cálculo de los valores segmentales de duración, la alineación tonal, el cálculo de los valores de F0 de los puntos de

[7] Una demostración del sistema SYNTAIX se encuentra en http://www.lpl.univ-aix.fr/~roy/cgi-bin/metlpl.cgi

inflexión y el cálculo de la línea de declinación. Finalmente, la salida del sistema, probada en el sintetizador MBROLA (Dutoit *et al.* 1996), consta de una cadena de fonemas asociados a unos valores de duración y a unos valores de F0.

Al igual que el sistema basado en reglas, el sistema estadístico de asignación de contornos entonativos para el francés, basado en INTSINT, se compone de un módulo lingüístico que se encarga de asignar etiquetas prosódicas al texto y de un módulo fonético que toma las etiquetas prosódicas y las asocia a valores acústicos.

La diferencia entre ambos modelos se basa en el método utilizado para la asignación de los datos acústicos: mientras que el primer modelo usa reglas, el segundo se basa en un método probabilístico que permite generar los contornos entonativos a partir del etiquetado de las categorías gramaticales de los textos de un corpus.

8.3.2.3. El modelo del IPO

Uno de los principales objetivos de la llamada Escuela Holandesa (véase el capítulo 4) es tener en cuenta aquellos aspectos que son relevantes para la percepción, puesto que no todos los movimientos tonales observables en la curva melódica son perceptivamente significativos. Durante años el IPO (*Institute for Perception Research*) del *Center for Research on User System Interaction of Technology* de la Universidad Técnica de Eindhoven ha tenido la síntesis como uno de sus principales intereses, y la ha empleado además como apoyo en su trabajo sobre prosodia, tanto para validar el modelo allí desarrollado como para proporcionar una salida oral a los sistemas de diálogo. También se utilizado el modelo del IPO para generar la entonación en varios idiomas, además del holandés, tal como se expone sucintamente a continuación.

Adriaens (1991) presentó un modelo para el alemán en el que los movimientos tonales, que se distribuyen en cinco líneas rectas de declinación, se describen de acuerdo con la posición en la sílaba en la que aparecen, el rango (expresado en semitonos) y la duración. Esta estrategia ha servido a su vez de base para la realización de nuevos modelos (Brindöpke *et al.* 1997).

Para el inglés, Willems *et al.* (1988) desarrollan el sintetizador Step TTS, usado en la generación de secuencias de estímulos para experimentos de percepción, como el de Sanders (1996), que estudia acentos y fronteras entonativas, aunque empleando otros marcos teóricos.

Por su parte, Beaugendre (1994) desarrolla en el *Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur* (LIMSI)[8] un modelo de generación automática de la entonación para el francés en el que la estilización de las curvas melódicas es manual, y la síntesis se realiza mediante predicción lineal (LPC – *Lineal Predictive Coding*). Se obtienen las líneas de referencia superior e inferior y tras la estandarización se proponen nueve movimientos

[8] En http://www.limsi.fr/Individu/cda/sonsCdA97.html se muestran ejemplos sonoros de la conversión de texto en habla desarrollada en el LIMSI.

melódicos. En la Figura 1 se esquematiza el método empleado por la Escuela Holandesa aplicado al francés.

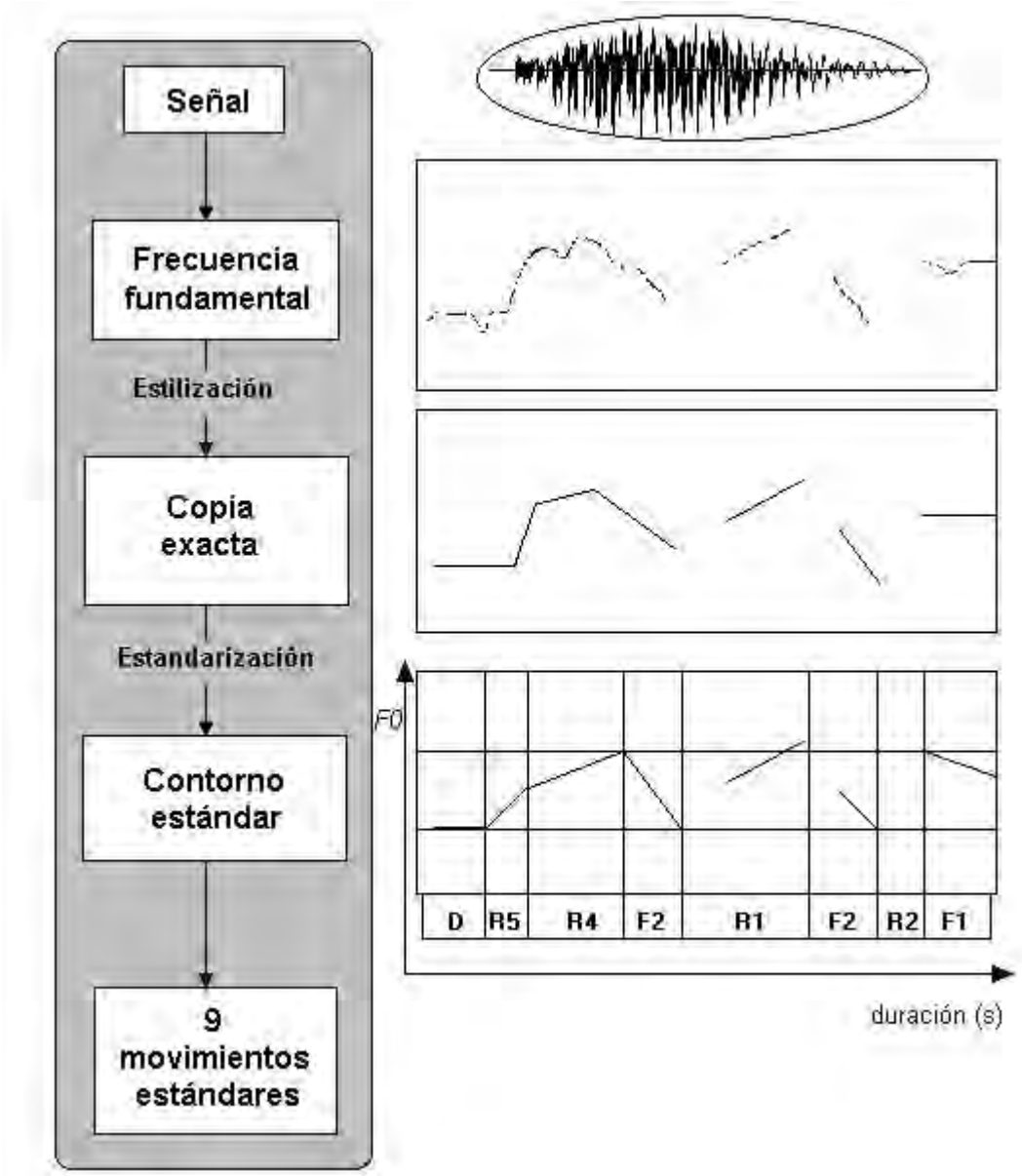

Figura 1. Presentación global del método empleado por la Escuela Holandesa aplicado al francés (Adaptada de Beaugendre 1996[9]).

[9] http://www.bibliotheque.refer.org/parole/beaugend/beaugend.htm

La tesis de Boula de Mareüil, aparecida en 1997 (Boula de Mareüil 1997a), supone también una mejora del sistema del LIMSI, tanto por lo que se refiere a la transcripción fonética como por lo que respecta a la identificación de unidades y límites prosódicos a partir de información eminentemente sintáctica.

El modelo del IPO también se ha utilizado para mejorar sistemas de diálogo; por ejemplo, el sistema de diálogo OVIS, destinado a ofrecer información teléfonica sobre transporte en ferrocarril (Klabbers 2000), o el prototipo SPICOS (*Siemens Philips IPO COntinous Speech*).

8.3.2.4. El modelo de H. Fujisaki

Fujisaki y sus colaboradores han desarrollado en las últimas décadas un modelo prosódico cuantitativo basado en la fisiología, en el que se concibe la melodía como una respuesta al control neuromotor de la vibración de las cuerdas vocales (Fujisaki y Nagashima 1969). Es, además, un modelo de superposición en el que diversos componentes actúan e interaccionan, ya que un contorno de F0 se determina mediante la suma de un valor básico de F0 para el hablante (F_{min}), de unos componentes de frase y de unos componentes de acento (expresados en una escala logarítmica). Cada uno de estos componentes se justifica además por la fisiología propia de los mecanismos de espiración. La generación de un contorno sintetizado con este método requiere además información sobre el tipo de locutor y sobre la lengua a la que se aplica.

Para imitar un contorno natural de F0 se modelan separadamente dos tipos de unidades: un contorno entonativo global bajo-alto-bajo que comprende toda la unidad prosódica y una serie de contornos bajo-alto-bajo de ámbito silábico. Se produce así un movimiento global que se inicia con un ascenso y que sigue con un descenso lento hasta el final. Paralelamente, con los contornos locales se generan picos sobre las sílabas acentuadas, atendiendo a los parámetros de duración y amplitud. El contorno obtenido al fin tras los cálculos es el resultado de la superposición de ambos componentes.

El análisis por síntesis permite una comparación del contorno natural con el generado, así que la F0 se puede descomponer en los componentes del modelo y decidir qué valores van a tomar los parámetros mediante sucesivas aproximaciones de los contornos simulados a los naturales. Los resultados del análisis por síntesis pueden observarse, para una oración del español, en la Figura 2.

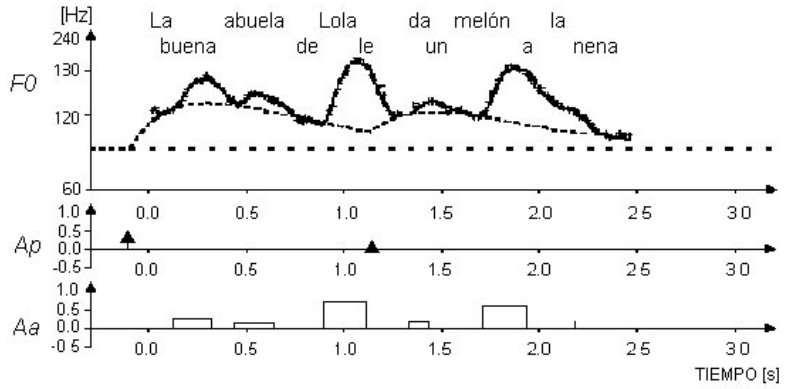

Figura 2. Ejemplo de análisis por síntesis del contorno de F0 de la oración "La buena abuela de Lola le da melón a la nena" (Fujisaki *et al.* 1998).

La representación de la entonación que proporciona el modelo es, por tanto, paramétrica. La Figura 3 muestra un diagrama en el que se esquematiza el proceso de extracción automática de parámetros.

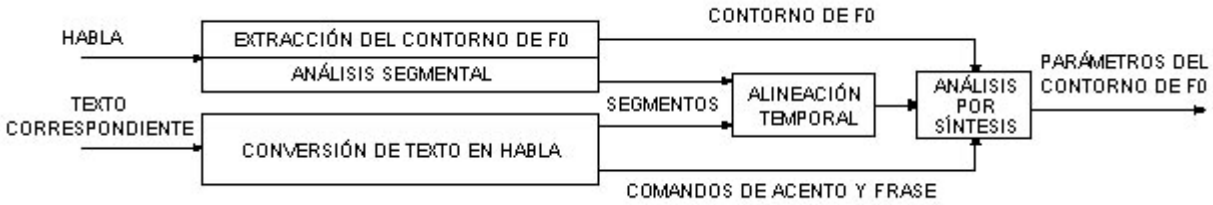

Figura 3. Extracción automática de los parámetros del contorno de F0 (Adaptada de Fujisaki y Ohno 1996).

A pesar de que, en principio, el modelo se diseñó para dar cuenta de los contornos de F0 en japonés (Fujisaki y Nagashima 1969), ha sido extendido y aplicado, aunque con distintas modificaciones, a varias lenguas, como el chino (Fujisaki *et al.* 1987 y Fujisaki *et al.* 1990), el inglés (Fujisaki y Ohno 1995 y Fujisaki *et al.* 1998), el alemán (Möbius 1993, Mixdorff 1997 y Mixdorff y Fujisaki 1994), el griego (Fujisaki *et al.* 1997), el coreano (Fujisaki 1996), el español (Fujisaki *et al.* 1994 y Gutiérrez *et al*. 2001), el sueco (Fujisaki *et al.* 1994), el vasco[10] (Navas *et al.* 2000) y el italiano (Salvo Rossi *et al.* 2002).

Con el método de superposición de Fujisaki puede darse cuenta de los fenómenos prosódicos que originan modulaciones en un ámbito local, como es el caso del acento en lenguas como el inglés, que incide en el contorno de la sílaba. De hecho, aunque en la aplicación original del modelo al japonés únicamente se generaban contornos de ámbito silábico para las sílabas acentuadas, la

[10] En http://bips.bi.ehu.es/ahoweb/ puede escucharse el resultado de la aplicación del modelo de Fujisaki a la conversión de texto en habla en vasco.

implementación del modelo en lenguas como el francés ha requerido generar estos contornos para todas las sílabas del enunciado. Así se ha conseguido dar cuenta de fenómenos locales, a la par que se ha ganado en naturalidad en la síntesis (Keller *et al*. 1997).

En cuanto a la predicción del contorno global, la generación de oraciones enunciativas ha dado resultados positivos en muy distintas lenguas. Sin embargo, una de las limitaciones de la versión inicial que más se ha discutido (Taylor 1994 y Keller 1997) es la que concierne a la predicción de contornos en los que existe un ascenso final, como en el caso de algunas oraciones interrogativas, dado que el modelo predice un contorno global que tiende a descender. El problema se ha solventado en versiones posteriores al introducir también ascensos (Mixdorff 1997).

Aunque se trata de un método de base matemática y fisiológica, se persigue una vinculación entre la acción del modelo y los factores lingüísticos que justifican las variaciones prosódicas. De acuerdo con Möbius (1993), una descripción cuantitativa de la entonación permite obtener mejores resultados si el modelado de los contornos de F0 y la extracción de los parámetros pertinentes se realiza básicamente siguiendo criterios lingüísticos y prosódicos, dejando en segundo término el criterio matemático de la aproximación óptima. De hecho, uno de los retos actuales es la incorporación a la síntesis de las variaciones debidas al uso de distintos estilos de habla. Higuchi *et al.* (1997), por ejemplo, han presentado la modelización mediante reglas de cuatro estilos distintos, y Tams y Tatham (2000) han considerado también la aplicación del modelo en sistemas que pretenden reflejar la variación característica del habla natural.

8.3.2.5. El modelo de P. Mertens

El sistema de Mertens (1999), concebido inicialmente para el francés, toma la sílaba como unidad básica de entonación, y describe las curvas melódicas como una secuencia de tonos asociados a las sílabas. En este modelo se distinguen cuatro niveles tonales básicos, dos clases de acentos y tres dominios prosódicos.

Los niveles tonales básicos son: *High* (simbolizado mediante H, h), Low (simbolizado mediante L, l), *Extrahigh* (simbolizado mediante H+, h+) y *Extralow* (simbolizado mediante L+, l+), que se asocian a cualquier tipo de sílaba, ya sea acentuada (simbolizada mediante H, L, H+, L+) o inacentuada (simbolizada mediante h, l, h+, l+). Las dos clases de acentos consideradas son el acento final (AF) y el acento inicial (AI); y los tres dominios prosódicos, el grupo acentual, el grupo de entonación y el "paquete entonativo" (*paquet intonatif*).

En francés, un grupo acentual consiste en una palabra con acento léxico y todas las palabras átonas que dependen de ella. Estos grupos acentuales se utilizan para determinar los grupos de entonación, que se definen por la presencia de una sílaba tónica del tipo AF. Este acento final puede estar precedido por una secuencia de una o más sílabas átonas (NA), por una sílaba tónica de tipo AI, o por una sílaba tónica AI precedida y/o seguida de una serie de sílabas átonas. La estructura del grupo entonativo (IG) del francés se esquematiza del siguiente modo (los paréntesis indican partes opcionales): IG → ((NA) AI) (NA) AF (NA).

Cada tono de tipo AF constituye un límite prosódico de modo que una secuencia de grupos de entonación implica, por tanto, una secuencia de límites. Los grupos de entonación se organizan recursiva y jerárquicamente dentro de un paquete entonativo. El mecanismo de agrupación es específico de cada lengua. Las reglas de agrupación del francés se explican en Mertens (1990).

La figura 4 ilustra algunos ejemplos de grupos de entonación.

			AF	oui HL-
		NA	AF	sûrement ll HL-
	AI		AF	assez H L- L-
	AI	NA	AF	précisément H h.l L-L-
NA	AI	NA	AF	dans ma situation 1....1 H ll HH
NA	AI		AF	pour ce projet 1.....1 H L-L-

Figura 4. Ejemplos de grupos de entonación del francés según el modelo de Mertens (2002).

En este modelo, antes de aplicar el módulo de generación de la secuencia tonal, se lleva a cabo un análisis morfológico y sintáctico que realiza una representación jerárquica del enunciado en una cadena de sílabas, reagrupadas posteriormente en palabras, que a su vez se agrupan en constituyentes sintácticos; también se representan las relaciones sintácticas entre los constituyentes y se asocian marcas enunciativas (por ejemplo, incisos prosódicos, focalizaciones y modalidad entonativa) a ciertas partes de la representación.

La primera implementación de este modelo está descrita en los trabajos de Malfrère, Dutoit y Mertens (1998a, 1998b y 1998c), que utilizan el sintetizador MBROLA (véase el apartado 8.3.3.2.) para su evaluación. En la versión actual (Mertens 2002), se utiliza, además, un modelo de contornos constituido por una secuencia de valores de F0 (*pitch targets*) asociados a un sonido vocálico determinado. El modelo permite controlar los siguientes parámetros de forma independiente:

1. El intervalo melódico entre los niveles tonales bajo y alto (*pitch range*).
2. El tono más bajo del rango del hablante, que normalmente aparece al final de una oración enunciativa (*floor*).
3. La pendiente (*slope*) de los niveles tonales altos y bajos en una escala temporal, que determina el efecto de declinación o inclinación.
4. El tono más alto del rango del hablante (*ceiling*).

Tal como señala Mertens (2002), la implementación en el sistema de conversión de texto en habla MINGUS (*Modular Intonation Generation Using Syntax*) constituye un procedimiento para validar este y otros modelos prosódicos, dada la flexibilidad con la que permite controlar los diversos parámetros que configuran la curva melódica[11].

8.3.2.6. El modelo de N. Thorsen-Grønnum

Al igual que Fujisaki (véase el apartado 8.3.2.4), Thorsen-Grønnum (1979, 1980, 1983, 1985, 1986, 1995 y 1998) propone un modelo de superposición, pensado en principio para el danés estándar, pero aplicable a otras lenguas. Se trata de un modelo creado inicialmente para generar la curva de entonación de frases simples constituidas por un máximo de cuatro grupos acentuales. En él se tiene en cuenta la relación entre los movimientos de F0 y la tonicidad de las sílabas. Cada sílaba, átona o tónica, lleva asociado un movimiento de F0, modificado por propiedades intrínsecas de los sonidos que aparecen en la secuencia. En este modelo se consideran los siguientes componentes: el texto, en el que podemos encontrar un contorno de entonación global; la oración, que presenta un contorno de entonación propio; los patrones de los grupos acentuales; la sílaba en la que recae la diferencia tonal (*stød*)[12]; y los valores intrínsecos de la frecuencia del fundamental de cada sonido (componente microprosódico).

La curva de entonación y los patrones de los grupos acentuales son específicos de cada lengua y están controlados por el hablante, mientras que el componente microprosódico es independiente del locutor.

El componente prosódico más importante es el grupo acentual, que consta de una sílaba tónica más todas las sílabas átonas que le siguen dentro del mismo contorno de entonación. El límite del grupo prosódico se coloca antes de la sílaba tónica, independientemente del número y tipo de fronteras sintácticas. Thorsen, basándose en la observación de la curva de F0, describe el grupo acentual del danés mediante un patrón constante: una F0 descendente (*Low*) asociado a las sílabas tónicas, seguido por un patrón ascendente-descendente (*High-falling*) asociado a las sílabas átonas.

[11] Ejemplos sonoros del papel de la sintaxis en MINGUS se encuentran en http://bach.arts.kuleuven.ac.be/pmertens/prosody/mingus.html

[12] El *stød* se define como una irregularidad en la vibración de las cuerdas vocales que se da en ciertas rimas de las sílabas en determinadas condiciones (Fischer-Jørgensen, 1989).

HÅNDBOLDSPIL ER meget anstrengende
(jugar a balonmano es agotador)
baNANERNE I kassen er rådne
(las bananas de la caja se han roto)
allerGI ER EN INfam sygdom
(la alergia es una enfermedad infame)

Figura 5. Representación del patrón tonal del grupo acentual correspondiente a las secuencias en mayúsculas (extraído de Thorsen 1983: 189).

El patrón del grupo acentual puede variar en función de su posición dentro de la oración y del número de sílabas átonas, que modifica además el intervalo de tiempo entre las sílabas tónicas. Si, por ejemplo, no se dan sílabas átonas, no existe la posibilidad dentro del grupo fónico de que tenga lugar un ascenso y un descenso de F0, y se da un truncamiento.

Se considera también la variación de F0 entre hablantes, puesto que el grado de descenso de F0 después de las sílabas tónicas no es el mismo en todos los locutores.

Dada la recurrencia y la constancia de este patrón acentual, el contorno de entonación se define sólo a partir de las sílabas tónicas –el pico se alinea con la primera sílaba postónica– y de la modalidad oracional. Esto no quiere decir que las sílabas átonas no sean importantes para la identificación del contorno entonativo, pero, según Thorsen (1980), son redundantes. Estos contornos de entonación son suficientes para distinguir los diferentes tipos de oraciones simples en danés: enunciativas, interrogativas con partícula interrogativa e interrogativas sin partícula. En frases cortas, el contorno de entonación está constituido por una línea cuya pendiente depende de la modalidad oracional: una enunciativa, por ejemplo, presenta una pendiente mayor que una interrogativa. La diferencia entre una enunciativa y una interrogativa no reside en el patrón del grupo acentual final del enunciado, sino en la declinación del enunciado.

Para las frases complejas enunciativas (Thorsen 1983) se sigue el mismo procedimiento que hemos explicado para las frases simples. La figura 6 muestra el contorno melódico (línea continua) y los patrones entonativos de los grupos acentuales (línea discontinua) en una frase enunciativa larga. Los círculos grandes muestran las sílabas tónicas; los pequeños, las sílabas átonas. En este tipo de frases, se descompone el enunciado en frases más cortas, obteniendo contornos de frase menores, cada uno con su propia declinación; todos ellos unidos describen el contorno global.

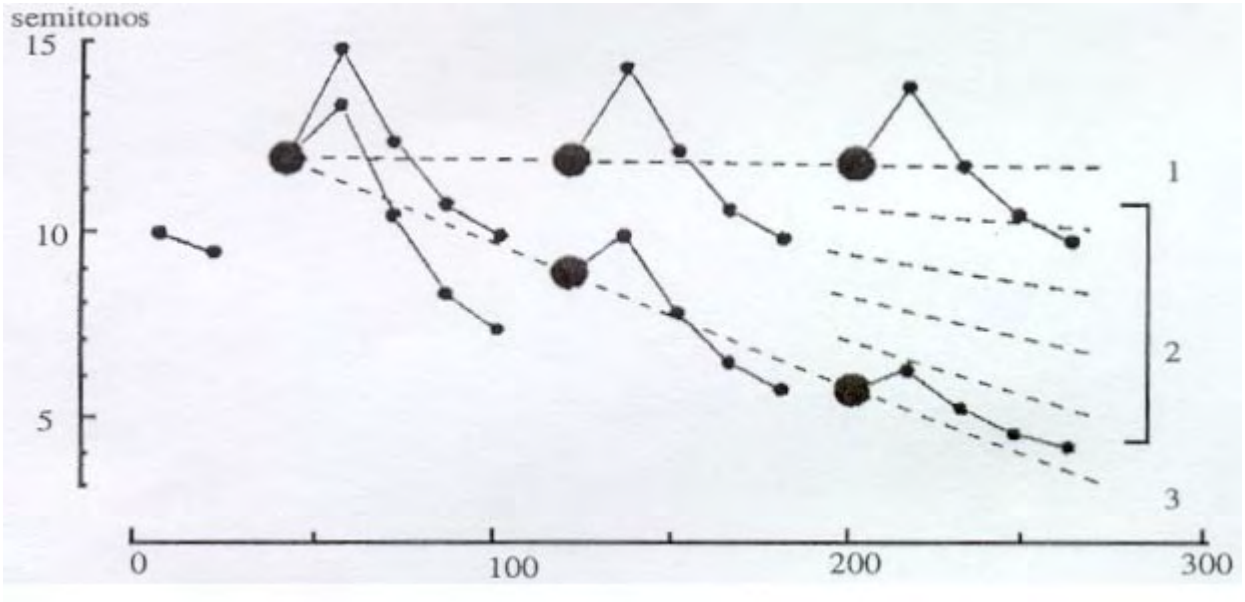

Figura 6. Patrones entonativos de los grupos acentuales y contorno melódico formado por la unión de las sílabas tónicas, marcadas en la figura mediante círculos grandes (Grønnnum 1998: 134).

Finalmente, en el texto se da la misma tendencia que en las frases largas; sin embargo, para obtener el contorno de un texto se debe tener en cuenta el número de oraciones que lo componen, además de la longitud de estas oraciones (Thorsen 1985 y 1986). Brøndsted (1997) señala que este modelo no define la entonación de las oraciones constituidas por un solo grupo acentual.

El modelo de Thorsen ha sido usado por Brøndsted (1996) en el módulo prosódico de un sistema de información horaria de trenes en danés.

8.3.3. HERRAMIENTAS PARA LA GENERACIÓN DE UN MODELO ENTONATIVO EN LA CONVERSIÓN DE TEXTO EN HABLA

También existen sistemas de síntesis del habla en los que la entonación no se genera a partir de los modelos mencionados en el apartado anterior, sino que el modelo entonativo se desarrolla a propósito para un determinado conversor. Son ejemplos de ello Festival, MBROLA y los desarrollados por algunos grupos especializados en tecnologías del habla; dado el gran número de trabajos realizados, en este último caso nos referiremos únicamente a los llevados a cabo en España.

8.3.3.1. Festival

Festival es un sistema de conversión de texto en habla multilingüe (incluye inglés británico, inglés americano, español y galés) desarrollado en el *Centre for Speech Technology Research*

(CSTR) de la Universidad de Edimburgo[13]. La arquitectura del sistema permite la manipulación de los parámetros utilizados en los distintos módulos del conversor.

Este sistema permite generar la entonación para la síntesis del habla partiendo de modelos de entonación, como por ejemplo ToBI (véase el apartado 8.3.2.1), pero también ofrece la posibilidad de crear otros modelos de entonación a partir de la predicción de los acentos (y/o los tonos finales) y de los valores de F0.

Con estos parámetros, Festival ofrece, por defecto, el modelo más simple, que es el resultado de interpolar el valor de F0 inicial y el valor de F0 final –generalmente más bajo que el inicial– del enunciado que se desea generar.

Otra manera de generar la entonación en Festival es utilizar un procedimiento basado en árboles de regresión binaria. Aunque este modelo es el resultado de decisiones estadísticas y no de decisiones lingüísticas, la generación de las curvas melódicas es rápida y sencilla, y el investigador no necesita un conocimiento exhaustivo del funcionamiento de la lengua en la que está trabajando. Es un modelo basado en los parámetros comentados -acentos y F0- en el que se considera además el valor medio de F0 de un hablante y la desviación típica de los valores de frecuencia del mismo.

Para la extracción de la curva de F0 se consideran el inicio y el final del enunciado, además de tres puntos para cada sílaba tónica (inicio, medio y final). La asignación del valor de F0 inicial del enunciado se toma del valor medio de la frecuencia del fundamental del hablante. El valor de F0 final del enunciado se obtiene de la diferencia entre el valor medio de la frecuencia del fundamental del hablante y su desviación típica. Los valores iniciales y finales de cada sílaba acentuada corresponden a los de la línea de base. Finalmente, el valor medio de cada sílaba acentuada se obtiene a partir de la suma del valor de la línea de base y del valor de la desviación típica.

A modo de ejemplo, si tenemos un hablante con una F0 media de 110 Hz. y una desviación típica de 25 Hz., el valor inicial del enunciado será de 110 Hz. y el valor final será de 85 Hz. (110 Hz. – 25 Hz.). En la figura 7 se muestra la línea de base correspondiente a estos valores.

[13] Se accede a demostraciones de Festival en http://www.cstr.ed.ac.uk/projects/festival/

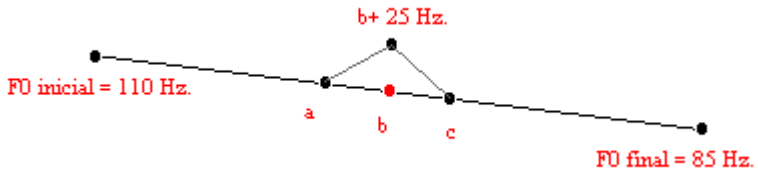

Figura 7. Ejemplo de la generación de la curva de F0 de un enunciado con una sola sílaba tónica utilizando el modelo basado en árboles de regresión binaria (CART).

Basándose también en los árboles de regresión, Festival incluye un modelo de árboles de entonación a partir de los tres valores de F0 de la sílaba tónica (inicio, medio y final) y de la asignación de los tonos de frontera; este modelo utiliza para su implementación las etiquetas de entonación del sistema ToBI (Black y Hunt 1996), aunque se podría utilizar con etiquetas procedentes de otros modelos.

8.3.3.2. MBROLA

MBROLA 2.00 es un herramienta de conversión de texto en habla basada en la concatenación de difonemas desarrollada en la *Faculté Polytechnique de Mons* (Bélgica) que funciona en más de 20 lenguas: entre otras, portugués de Brasil, bretón, inglés británico, holandés, francés, alemán, español y sueco[14].

Este programa permite dar información sobre los valores de duración y de frecuencia del fundamental de cada alófono considerado, y acepta hasta un máximo de 20 valores de F0 para cada alófono. Sin embargo, aunque estos valores de F0 se pueden utilizar para dibujar una curva melódica, esto no implica la existencia de un modelo prosódico.

MBROLA ha sido usado, por ejemplo, por Malfrère, Dutoit y Mertens (1998a, 1998b y 1998c) con el fin de evaluar los resultados obtenidos en sus trabajos sobre generación automática de prosodia para la conversión de texto en habla (véase el apartado 8.3.2.5).

8.3.2.3. El tratamiento de la entonación en algunos sistemas de conversión de texto en habla desarrollados en España

El modelo de entonación diseñado para el conversor de texto en habla del *Grup de Processament de la Parla* del *Departament de Teoria del Senyal i Comunicacions* de la Universidad Politécnica de Cataluña[15] es el resultado de la interacción de diversos niveles, aunque en este momento únicamente usa los niveles de enunciado y de grupo tónico. Los patrones entonativos del

[14] Los resultados obtenidos con MBROLA en diversas lenguas pueden escucharse en http://tcts.fpms.ac.be/synthesis/mbrola.html

[15] Pueden encontrarse demostraciones en http://gps-tsc.upc.es/veu/veu.html

enunciado se generan a partir de una serie de puntos de inflexión unidos mediante líneas rectas, procedimiento con el que se describe la evolución de la F0 a lo largo del tiempo. Los puntos de inflexión que dan lugar al patrón entonativo del enunciado se representan de una forma paramétrica, de modo que el número de puntos y los valores temporales y frecuenciales de cada punto pueden ajustarse al modelo entonativo de la lengua y a las características idiosincrásicas del hablante. Los patrones básicos que se consideran en este sistema son los de la declarativa, exclamativa, interrogativa y oración inacabada (Bonafonte *et al.* 1998).

En el conversor del Grupo de Tecnología del Habla del Departamento de Ingeniería Electrónica de la Universidad Politécnica de Madrid[16] el contorno de F0 se estiliza mediante líneas rectas trazadas entre picos y valles. Cada grupo fónico se divide en tres partes: una inicial, que va desde el principio del grupo hasta la primera sílaba acentuada; una central, que va desde la primera sílaba acentuada hasta la última sílaba acentuada; y la final, que va desde la última sílaba acentuada hasta el final del grupo fónico. En la parte inicial, se asigna un valor de F0 a la primera sílaba dependiendo del tipo de pausa del final del grupo fónico. En la parte central, se calculan los valores de F0 de los picos y los valles en función del número de sílabas acentuadas, y se interpolan líneas rectas entre ellos. En la parte final se tienen en cuenta la modalidad oracional y la posición de la última sílaba acentuada para obtener los valores de F0. Finalmente, los picos, los valles y las reglas de declinación se calculan a partir de valores estadísticos extraídos de una base de datos de habla continua (Pardo *et al.* 1995). Actualmente, se está desarrollando un sintetizador capaz de generar habla con emociones (Montero *et al.* 1998 y 1999).

El modelado prosódico propuesto por el Grupo de Aplicaciones del Procesado de Señal del Departamento de Señales, Sistemas y Radio de la Universidad Politécnica de Madrid se realiza en tres etapas: determinación de las pausas (y por tanto de los grupos fónicos), asignación de un patrón entonativo a cada grupo fónico resultante, y cálculo de la curva de entonación. En este sistema, cuya unidad básica es la sílaba, los constituyentes silábicos se agrupan en grupos tónicos y éstos en grupos fónicos. El contorno prosódico de cada núcleo silábico está representado por cinco parámetros: dos valores de duración y tres de F0 (inicial, medio y final). Para generar la estructura prosódica se concatenan las sílabas, evitando entre ellas una gran variación de valores de F0 (López *et al*. 2002).

El Grupo de Tecnologías de las Comunicaciones, Procesado de Voz y Audio de la Universidad de Zaragoza ha desarrollado también un conversor de texto en habla[17], en cuyo módulo de análisis lingüístico se extrae el tipo de frase para aplicarle el patrón entonativo que le corresponda. Las frases se dividen en cuatro grupos en función del signo de puntuación del texto: enunciativa terminada, enunciativa inacabada, interrogativa y exclamativa.

[16] Están disponibles demostraciones en http://www-gth.die.upm.es/index-e.html

[17] Una demostración se encuentra en http://www.gtc.cps.unizar.es/

El sistema Departamento de Informática de la Universidad de Valladolid[18] ofrece la entonación del texto según dos modelos distintos: un módulo de entonación básico para frases enunciativas, interrogativas y exclamativas, y un módulo de entonación avanzado para frases enunciativas. Se pueden ajustar parámetros del sintetizador, como la velocidad de lectura, la duración de las pausas y el valor de la F0. En el módulo de entonación avanzado se generan los patrones entonativos asignando un tipo de patrón a cada grupo acentual, para después concatenar los patrones de grupos acentuales consecutivos.

Finalmente, cabe mencionar los conversores de texto en habla desarrollados para el español por empresas privadas como Telefónica I+D (Castejón *et al*. 1994, Rodríguez *et al*. 1993, Rodríguez 1997), Loquendo (Garrido *et al*. 2000), Élan Speech, Natural Vox, Babel Technologies, AT&T, Bell Labs o ScanSoft [19].

8.4. El reconocimiento automático del habla

Si la finalidad de la conversión de texto en habla es, como hemos visto, la transformación de un texto escrito en su equivalente leído, el **reconocimiento automático del habla** tiene como objeto conseguir una representación escrita o simbólica del mensaje presente en una onda sonora[20]. Precisamente la aplicación a la que tal vez se dedican más esfuerzos en el reconocimiento es el dictado automático, disponible ya en la actualidad en varias lenguas. Entre los elementos que deben extraerse de la onda sonora y que pueden contribuir a su adecuada interpretación se cuenta, como es natural, la prosodia.

Mientras que, como se ha expuesto en el apartado anterior, los sistemas de conversión de texto en habla incorporan, en mayor o menor medida, un modelo prosódico que incluye, al menos, duración segmental y patrones melódicos, no es este el caso del reconocimiento automático del habla (Pagel 1999, Batliner *et al*. 2001a, Mixdorff 2002), pese a que, como acertadamente menciona Pagel (1999), en el trabajo pionero de Lea (1980) se consideraba que los elementos suprasegmentales podían mejorar substancialmente el reconocimiento.

Suelen aducirse diversas causas que explican que la información prosódica no se haya aún incorporado plenamente a los sistemas de reconocimiento (Wang 2001): la complejidad de la manifestación acústica de fenómenos lingüísticos como el acento, relacionado con tres

[18] La demostración del sistema está disponible en en http://logos.dcs.fi.uva.es/index.html

[19] Pueden escucharse ejemplos en http://www.loquendo.com/ (Loquendo), http://www.elan.fr/ (Elan Speech), http://www.natvox.es/ (Natural Vox), http://www.babeltech.com/ (Babel Technologies), http://www.research.att.com/projects/tts/ (AT&T), http://www.bell-labs.com/project/tts/ (Bell Labs) y en http://www.scansoft.com/realspeak/demo/ (ScanSoft). En http://www.ldc.upenn.edu/ltts/ se ofrece la posibilidad de comparar varios sistemas de conversión de texto en habla.

[20] Para una presentación general del reconocimiento del habla véase, por ejemplo, Bersntein y Franco (1996), Kurzweil (1998) o Levinson y Liberman (1981).

parámetros –duración, frecuencia fundamental y energía/intensidad– que no pueden modelarse independientemente sin pérdidas importantes de información, y que dependen, además, de factores intrínsecos y de factores contextuales; la dificultad de reconocer automáticamente las unidades utilizadas en las descripciones fonológicamente orientadas de la entonación, que constituyen un nivel intermedio –que algunos autores consideran excesivamente reduccionista y otros de una complejidad innecesaria– entre la onda sonora y la función sintáctica o semántica; la importante variación interlocutor en la manifestación acústica de la entonación y el acento, tanto léxico como de frase; las dificultades en la detección automática de F0, –agravadas en entornos ruidosos o cuando el reconocimiento debe realizarse a través del teléfono– que pueden llevar a errores en el reconocimiento.

Por otra parte, diversos autores coinciden en que los Modelos Ocultos de Markov en los que se basan buena parte de los sistemas de reconocimiento no constituyen el método óptimo para modelar aquellos elementos prosódicos que se manifiestan en el nivel de la sílaba o de unidades mayores. La estrecha correlación entre la prosodia y la sintaxis hace también que el reconocimiento de determinados fenómenos como las fronteras prosódicas o los acentos de frase dependan de un análisis sintáctico y semántico a menudo no disponible.

Finalmente, se ha sugerido también (Batliner *et al*. 2001a) –y seguramente éste es un elemento esencial que ayuda a comprender la situación actual de las tecnologías del habla– la diferencia entre la "cultura" de las humanidades y la de la ingeniería como una de las razones que explican el escaso uso en los sistemas reales de reconocimiento de los datos prosódicos acumulados hasta el momento. El carácter excesivamente teórico de muchos modelos entonativos –en algunos casos, como ToBI, centrados en una concepción muy específica de la fonología y, por ello, poco integradores– y la discusión entre escuelas sobre aspectos genéricos hacen difícil llevarlos a la práctica. Los comentarios de Batliner *et al*. (2001a), especialmente relevantes por referirse a un proyecto emblemático como Verbmobil, en lo que se refiere al debate sobre niveles o movimientos en la descripción de la entonación son ilustrativos al respecto: "a un reconocedor de habla le es indiferente si se ha entrenado con niveles o con movimientos mientras la base de datos sea lo suficientemente grande y las etiquetas se hayan anotado correctamente. Después de todo, lo que sube debe bajar: no importa si es un H* a 200 Hz. al que sigue un L* a 100 Hz. o si se trata de un movimiento entre 200 y 100 Hz.".

Los trabajos orientados a obtener modelos prosódicos para el reconocimiento del habla se han centrado principalmente en dos niveles: el léxico, intentando modelar los correlatos acústicos del acento en la palabra, y el oracional, generalmente abordando los correlatos prosódicos de las fronteras entre constituyentes sintácticos.

El **acento léxico** ha sido uno de los aspectos que más atención ha recibido en el campo del modelado prosódico para el reconocimiento del habla (Wang 2001)[21]. Los primeros trabajos con

[21] Puede verse, por ejemplo, sobre el español Caminero *et al*. (1999) y Rubio y Milone (2002).

sistemas de reconocimiento de palabras aisladas se centraron en la determinación automática del patrón acentual con el fin de reducir el número de palabras que acústicamente pueden confundirse en sistemas con vocabularios muy amplios, o para diferenciar entre pares mínimos que se distinguen únicamente por el acento. Los estudios dedicados al reconocimiento del habla continua proponen el uso de modelos acústicos diferenciados de segmentos acentuados y no acentuados con el fin de recoger conjuntamente las variaciones debidas a los cambios inducidos por el acento en la duración, la energía y la F0, o la distinción automática entre sílabas tónicas y átonas. En el caso del reconocimiento del habla espontánea, las dificultades son aún mayores, pues no todas las sílabas léxicamente acentuadas en palabras aisladas lo son en el discurso continuo y el acento léxico puede superponerse al acento de frase, que a menudo es impredecible a partir solo de la representación ortográfica.

Los resultados obtenidos al incorporar al reconocimiento información sobre el acento léxico parecen indicar que su principal utilidad reside, al menos en el caso de las lenguas de acento libre como el inglés, en la reducción de errores causados por el módulo de análisis acústico del reconocedor, ya que se reduce el número de hipótesis sobre las posibles palabras que el sistema debe reconocer en un determinado punto del enunciado. Tal como lo resume Wang (2001: 109): "las mejoras derivadas del uso de modelos prosódicos parecen conseguirse principalmente eliminando hipótesis que no son plausibles más que distinguiendo las diferencias de detalle entre varios niveles de acento y varias clases segmentales; por ello, no encontramos una mejora adicional utilizando un modelado más refinado".

En lo que se refiere a la detección automática de indicios prosódicos en el **ámbito oracional**, el trabajo de Pagel (1999) muestra una mayor fiabilidad en la detección de acentos que preceden el final de un grupo fónico que en la detección de acentos en el interior de grupos si se utilizan únicamente parámetros acústicos; de estos resultados deduce el autor la necesidad de un modelo de lenguaje o, preferiblemente, de un analizador sintáctico. Esto explicaría el motivo por el cual, como sostienen también Batliner *et al*. (2001b), "en el campo del reconocimiento del habla, la comunidad científica centra más sus esperanzas en los progresos realizados en módulos como el modelo de lenguaje que en la prosodia" (Pagel 1999: 127).

También Batliner *et al*. (2001a) insisten en la relevancia de las fronteras prosódicas dado su papel funcional en la delimitación de unidades, del acento y de la modalidad oracional.

En una breve exposición del papel de la prosodia en el reconocimiento automático del habla, Cassidy (2001) centra igualmente la utilidad de la información prosódica en niveles postléxicos, como la localización de fronteras entre constituyentes para decidir entre posibles análisis sintácticos o la detección del acento de frase en relación con la probabilidad de aparición de cierto tipo de palabras en esta posición. El entrenamiento de sistemas de reconocimiento con corpus prosódicos anotados utilizando etiquetas similares a las de ToBI tiene, en opinión de este autor, el problema de la ambigüedad que presentan las diferentes clases de tonos y acentos tonales. Su conclusión no es mucho más optimista que la de los trabajos anteriormente presentados: "los indicios prosódicos pueden ser útiles en diversas aplicaciones del

reconocimiento del habla, pero es claro que, por el momento, no sabemos cómo extraer información de la señal acústica de un modo fiable" (Cassidy 2001).

A pesar de las dificultades mencionadas, es previsible que en los próximos años los indicios prosódicos tengan un papel relevante en el reconocimiento del habla. Los sistemas de diálogo, tal como se describe en el apartado 5, incorporan necesariamente un reconocedor que puede beneficiarse del uso de información prosódica. Por otra parte, aplicaciones con un importante interés comercial como la extracción automática de información a partir de fuentes documentales sonoras se basan también en sistemas de reconocimiento que podrían considerar, por ejemplo, el acento léxico como portador de información semántica relacionada con el tema principal de un documento (Crestani 2001) o los indicios prosódicos asociados a un cambio de tema (Schriberg y Stolcke 2001); también otras aplicaciones como el resumen automático de los mensajes recibidos en un contestador telefónico pueden basarse parcialmente en correlatos prosódicos del énfasis puesto por el emisor en las palabras clave del mensaje (Koumpis y Renals 2001); finalmente, se han realizado trabajos orientados a la detección automática de una considerable intoxicación alcohólica a partir del análisis de características acústicas relacionadas con la F0, la intensidad y la duración en el interior de constituyentes prosódicos (Levit *et al*. 2001).

8.5. Los sistemas de diálogo

Los **sistemas de diálogo** tienen como finalidad que un usuario, generalmente a través del teléfono y sin la intervención de un operador humano, pueda acceder automáticamente a una determinada información o llegue a realizar una transacción. Dado que se emplea la lengua oral como medio de interacción, la gestión del diálogo entre una persona y una máquina requiere un sistema de reconocimiento –para interpretar el mensaje recibido del usuario– y un sistema de síntesis –para proporcionar la información solicitada–. En la figura 8 podemos observar los principales módulos en los que se organiza un sistema de diálogo[22].

[22] Los sistemas de diálogo se describen, por ejemplo, en Gibbon *et al*. (2000) o en Minker y Benaceff (2001). Una presentación general puede encontrarse en Zue (1999).

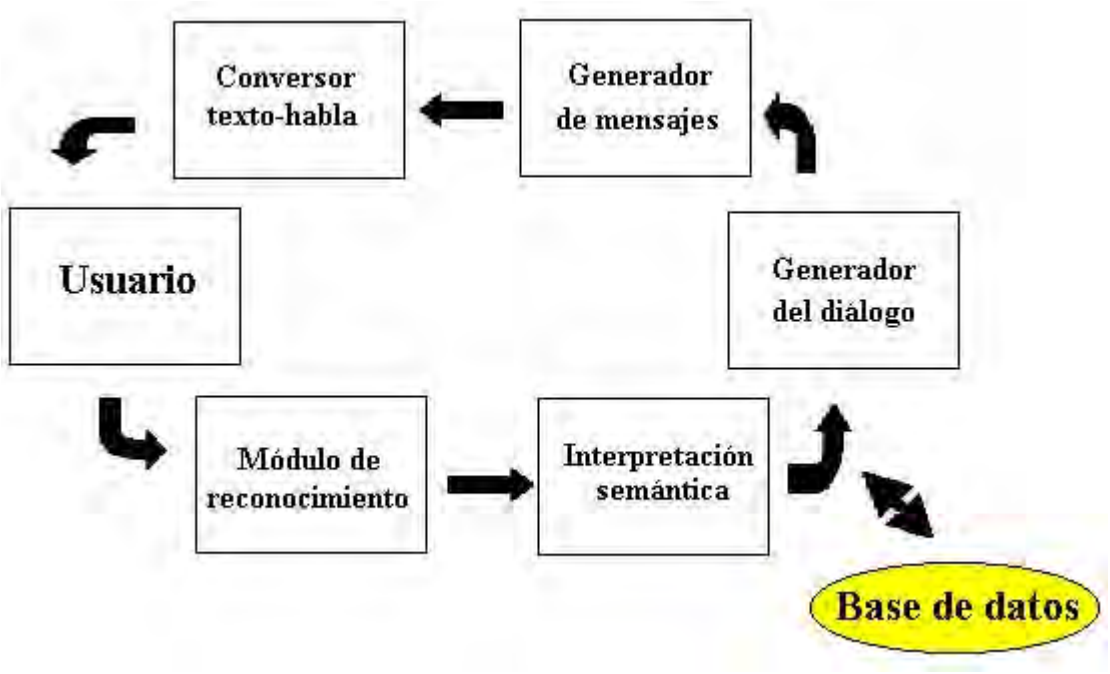

Figura 8. Principales módulos de un sistema de diálogo.

El módulo de **reconocimiento** debe procesar la información acústica que contiene el mensaje que recibe del usuario para poder interpretarlo semánticamente. En esta fase aparecen en ocasiones errores que son difíciles de detectar. Por ejemplo, el módulo de reconocimiento de un sistema automático de información de trenes puede tener problemas para distinguir *Palencia* de *Valencia*. Mediante el módulo de interpretación semántica, el sistema analiza la información solicitada por el usuario y la interpreta; así, la palabra *mañana* no se puede interpretar de la misma forma en la oración *yo quiero saber el horario de los trenes que salen hacia Valencia mañana* que en la oración *yo quiero saber el horario de los trenes que salen hacia Valencia por la mañana*. En estos casos, el módulo semántico debe ser capaz de identificar el significado de un determinado tipo de expresiones relacionadas con la aplicación para la que se ha diseñado un determinado sistema.

El módulo de **gestión de diálogo** se encarga, como su propio nombre indica, de dirigir el diálogo, bien para obtener más información del usuario bien para decidir que la información que éste le ha proporcionado ya es suficiente. En los sistemas actuales, el gestor de diálogo utiliza dos tipos de confirmaciones: la confirmación explícita y la confirmación implícita. En la confirmación explícita el sistema pregunta directamente al usuario, mientras que en la confirmación implícita el sistema formula la pregunta de forma indirecta, recuperando la información que ha recibido previamente. Una pregunta del tipo *¿De qué ciudad desea información?* es un ejemplo de confirmación directa, frente a *Para ir a Valencia...,* en la que el sistema comienza la pregunta a partir de la pregunta del usuario. En este caso, el usuario puede rectificar si el módulo de reconocimiento ha fallado. Finalmente, una vez que el sistema ha

interpretado qué información se desea, la busca en la base de datos y genera la respuesta adecuada. El módulo de conversión de texto en habla transforma esta respuesta en un mensaje oral, facilitando de este modo la información solicitada.

Una parte esencial de un sistema de diálogo es el **modelado de la prosodia**, porque éste debe aplicarse a dos módulos: el de síntesis y el de reconocimiento. Para el módulo de síntesis el problema se soluciona desarrollando un modelo prosódico dependiente de la aplicación o utilizando alguno de los descritos en el apartado 3, pero para el de reconocimiento la tarea es mucho más difícil, ya que un cambio en la entonación de un enunciado por parte del usuario implica que el sistema debe realizar interpretaciones semánticas diferentes. En un servicio de información de trenes, por ejemplo, la respuesta ante mensajes que desde el punto de vista segmental son iguales debe ser, en ocasiones, distinta: *No, quiero viajar por la mañana / No quiero viajar por la mañana*. Pensemos que, en el primer caso, el hablante puede realizar una juntura únicamente con una inflexión tonal, complicando aún más el proceso que conduce a la interpretación semántica que debe llevar a cabo el sistema. Otro problema propio de la lengua oral que debe resolverse en el módulo de reconocimiento es la identificación del realce de información contrastiva y el consiguiente debilitamiento prosódico de la información propia del trasfondo discursivo (de la Mota 1995 y 1997).

Actualmente, existen proyectos de investigación[23] cuyo objetivo es analizar aquellos aspectos prosódicos que tienen especial relevancia en la interpretación correcta de los mensajes que debe procesar un sistema de diálogo: la duración de las pausas y su función en el diálogo, los turnos de palabra, la clasificación de los diálogos en función de los diferentes actos de habla, la información nueva y la información conocida, la focalización de algunas partes del enunciado, los marcadores del discurso oral, la función de las palabras comodín en el discurso espontáneo y la correferencia (Hansson 1999, Goto *et al*. 1999).

También es patente una reciente preocupación por establecer la relación entre la prosodia y el gesto (Cassell 1999a y 1999b, Poggi 2001). Esta información se utiliza en aquellos sistemas de diálogo que se desarrollan mediante agentes animados. Entre el usuario y el sistema se establece un tipo de relación parecida a una conversación espontánea en la que los interlocutores tienen una relación directa cara a cara, de modo que el agente dialoga con el usuario para ofrecerle los servicios que le puede proporcionar el sistema.

8.6. Conclusiones

La incorporación de los modelos entonativos a la síntesis, al reconocimiento y a los sistemas de diálogo presentados en este capítulo pone de manifiesto, como señalan Botinis *et al*. (2001), la

[23] Por ejemplo, *Swedish Dialogue Systems Project* : http://www.ida.liu.se/~nlplab/sds/; NITE - *Natural Interactivity Tools Engineering*: http://mate.nis.sdu.dk/; ISLE-HLT - *International Standards for Language Engineering*: http://isle.nis.sdu.dk/; MATE- *Multilevel Annotation, Tools Engineering*: http://mate.nis.sdu.dk/

existencia de una relación circular entre modelos y aplicaciones: por una parte, el desarrollo de nuevas aplicaciones en el campo de las tecnologías del habla y la necesidad de mejorar las ya existentes requiere contar con investigaciones de naturaleza básica como las que se llevan a cabo en la lingüística; por otra, se dispone de una gran cantidad de conocimientos que no se han puesto todavía en práctica y que, cuando se integran en una aplicación, llevan a replantear las teorías o a buscar nuevos datos.

Los trabajos revisados muestran también que, en el estudio de la entonación, queda aún un largo camino por recorrer para que los datos y los modelos de que disponemos encuentren plenamente su lugar en las tecnologías del habla. En primer lugar, son necesarias descripciones entonativas con diversos niveles de abstracción que puedan aplicarse a sistemas de conversión de texto en habla y a sistemas de reconocimiento. Para ello, es imprescindible disponer de corpus prosódicos anotados en el mayor número de lenguas posible que reflejen, además, la variedad de estilos de habla propia de las diversas situaciones de interacción comunicativa que pueden darse tanto entre personas como entre personas y sistemas informáticos. Dadas las dificultades que señalábamos, debidas a la falta de un estándar de anotación prosódica, sería preciso, al menos, asegurar una mínima compatibilidad entre sistemas de anotación diferentes que facilite el uso compartido de los recursos existentes. Por otro lado, la implementación de un modelo prosódico en un conversor o en un reconocedor debería considerar simultáneamente los aspectos computacionales y los lingüísticos, contribuyendo así a establecer de un modo fructífero la relación circular a la que aluden Botinis *et al*. (2001).

La "tensión" entre las teorías lingüísticas de la entonación y la necesidad, en el ámbito de las tecnologías del habla, de llevarlas a la práctica en productos que respondan a necesidades reales –en otras palabras, la diferencia entre las dos "culturas" (Batliner *et al*. 2001a), o la "tensión" entre ciencia y tecnología (Greenberg 2001)– no debe interpretarse, al menos desde nuestro punto de vista, como un factor que dificulte el progreso, sino como un estímulo para avanzar conjuntamente, integrando adecuadamente el conocimiento lingüístico con el conocimiento tecnológico, de modo que podamos disponer en el futuro de teorías más explicativas y de aplicaciones más eficaces.

Bibliografía[24]

Adriaens, Léon (1991). *Ein Modell deutscher Intonation. Eine experimentell-phonetische Untersuchung nach den perzeptiv relevanten Grundfrequenzänderungen im gelesenen Text*. Tesis doctoral, Technische Universität, Eindhoven.

Aguilar, Lourdes, Juan M. Garrido y Joaquim Llisterri (1997). Incorporación de conocimientos fonéticos a las tecnologías del habla. En Enrique Serra, Beatriz Gallardo, Montserrat Veyrat, Daniel Jorques y Amparo Alcina (eds.) *Panorama de la investigació lingüística a l'Estat*

[24] La validez de las direcciones en Internet citadas en la biblografía y en las notas a pie de página se ha verificado en diciembre de 2002.

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

Espanyol. Actes del I Congrés de Lingüística General. Volum III. Comunicacions: Fonètica i Fonologia. Semàntica i Pragmàtica. València: Universitat de València. 5-13.
http://liceu.uab.es/~joaquim/publicacions/valencia_94.html

Anderson, Mark, Janet Pierrehumbert y Mark Liberman (1984). Synthesis by rule of English intonation patterns. En *Proceedings of ICASSP 1*, San Diego, CA. 2.8.1-2.8.4.

Astesano, Corine, Robert Espesser, Daniel Hirst, y Joaquim Llisterri. (1997). Stylisation automatique de la fréquence fondamentale: une évaluation multilingue. En *Actes du 4ème Congrès Français d'Acoustique*, Marsella, Francia, vol. 1. 441-443.
http://www.lpl.univ-aix.fr/~corine/articles/Cfa96.PDF

Batliner, Anton, Bernd Möbius, Gregor Möhler, Antje Schweitzer, Elmar Nöth (2001a). Prosodic models, automatic speech understanding, and speech synthesis: toward the common ground. En Paul Dalsgaard, Børge Lindberg, Henrik Benner y Zheng-Hua Tan (eds.) (2001) *Eurospeech 2001. Proceedings of the 7th European Conference on Speech Communication and Technology*, vol. 4. Aalborg, Denmark. 2285-2288.
http://www5.informatik.uni-erlangen.de/literature/ps-dir/2001/Batliner01:PMA.ps.gz

Batliner, Anton, Elmar Nöth, Jan Buckow, Richard Huber, Volker Warnke, y Heinrich Niemann (2001b). Whence and whither prosody in automatic speech understanding: A case study. En Bacchiani, Michiel, Julia Hisrschberg, Diane J. Litman y Mari Ostendorf (eds.) *Proceedings of the ISCA Tutorial and Research Workshop on Prosody and Speech Recognition*. Red Bank, N.J. 3-12.
http://www5.informatik.uni-erlangen.de/literature/ps-dir/2001/Batliner01:WAW.ps.gz

Beaugendre, Frédéric (1994). *Une étude perceptive de l'intonation du français, développement d'un modèle et application à la génération automatique de l'intonation pour un système de synthèse à partir du text*e. Tesis doctoral, Université de Paris XI, Notes et Documents LIMSI n° 94-25.

Black, A. y Hunt, A. (1996). Generating FO contours from ToBI labels using linear regression. En H. Timothy Bunnell y William J. Idsardi (Eds.) *Proceedings of the 1996 International Conference on Spoken Language Processing*, vol 3. Philadelphia: Penn. 1385-1388.

Blecua, Beatriz y Victoria Acín (1995). Propuesta de un modelo de intensidad vocálica del castellano y el catalán aplicable a un sistema de conversión de texto en habla, *Procesamiento del Lenguaje Natural,* 17. 257-271.

Bonafonte, Antonio, Ignasi Esquerra, Albert Febrer, José A. R. Fonollosa y Francesc Vallverdú (1998). The UPC Text-to-Speech System for Spanish and Catalan. En Robert H. Mannell y Jordi Robert-Ribes (eds.) *Proceedings of the 1998 International Conference on Spoken Language*

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

Processing, vol. 5. Sydney: Australian Speech Science and Technology Association, Incorporated (ASSTA). 1967-1970.
http://gps-tsc.upc.es/veu/research/pubs/download/Bon98c.pdf

Botinis, Antonis (2001). Special Issue on Intonation, *Speech Communication* 33, 4.

Botinis, Antonis (ed.) (2000). *Intonation. Analysis, Modelling and Technology.* Dordrecht: Kluwer Academic Publishers (Text, Speech and Language Technology, 15).

Botinis, Antonis, Björn Granström, y Bernd Möbius, B. (2001). Developments and paradigms in intonation research, *Speech Communication* 33, 4: 263-296.

Boula de Mareüil, Philippe. (1997a). *Étude linguistique appliquée à la synthèse de la parole à partir du texte*, Tesis doctoral,Université Paris XI, Orsay.

Brindöpke, Christel y Brigitte Schaffranietz (1997). Evaluation of an Intonation Model for German Spontaneous Speech. En Antonis Botinis, Georgios Kouroupetroglou y George Carayiannis (eds.) *Intonation: Theory, Models and Applications. Proceedings of an ESCA Workshop*, Atenas: European Speech Communication Association. 51-54.

Brøndsted, Tom (1996). Adapting a Prosody Classification Module from German to Danish. A Contrastive Analysis. En Yifan Gong (ed.) *Multi-lingual Spontaneous Speech Recognition in Real Environments*, Nancy: Spontaneous Interlingua Network (SPIN). 1-6.

Brøndsted, Tom (1997). Intonation Contours "distorted" by Tone Patterns of Stress Groups and Word Accents. En Antonis Botinis, Georgios Kouroupetroglou y George Carayiannis (eds.) *Intonation: Theory, Models and Applications. Proceedings of an ESCA Workshop*, Atenas: European Speech Communication Association. 55-58.

Bruce, Gösta (1977). *Swedish Word Accents in Sentence Perspective*. Lund: Gleerup.

Caminero, Javier, Eduardo López, Laura Docío y Luis A. Hernández (1999). On the use of fundamental frequency in a connected numbers recognition system. En *Proceedings of Eurospeech'99. 6th European Conference on Speech Communication and Technology,* Budapest, Hungary.

Campione, Estelle y Jean Véronis (1998). A Multilingual Prosodic Database. En Robert H. Mannell y Jordi Robert-Ribes (eds.) *Proceedings of the 1998 International Conference on Spoken Language Processing*, vol. 7. Sydney: Australian Speech Science and Technology Association, Incorporated (ASSTA). 3163-3166.
http://www.up.univ-mrs.fr/~veronis/pdf/1998icslp-database.pdf

Campione, Estelle y Jean Véronis (2000). Une évaluation de l'algorithme de stylisation mélodique MOMEL, *TIPA, Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence* 19. 27-44.

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

Campione, Estelle y Jean Véronis (2001). Semi-automatic tagging of intonation in French spoken corpora. En Rayson, Paul, Andrew Wilson, Tony McEnery, Andrew Hardie y Shereen Khoja (eds.) *Proceedings of the Corpus Linguistics'2001 Conference.* Lancaster, U.K.: Lancaster University, UCREL. 90-99.
http://www.up.univ-mrs.fr/~veronis/pdf/2001-lancaster-intonation.pdf

Cassell, Justine y Matthew Stone (1999a). Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems. *Proceedings of the AAAI 1999 Fall Symposium on Psychological Models of Communication in Collaborative Systems*. 34-42.
http://gn.www.media.mit.edu/groups/gn/publications/cassell-stone_AAAI99.pdf

Cassell, Justine, David McNeill y Karl-Erik McCullough (1999b). Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Non-Linguistic Information. *Pragmatics and Cognition* 7, 1. 1-33.

Cassidy, Steve (2001). Prosody and speech recognition. *SLP806: Speech Recognition*. Master in Science and Speech Language Processing, Department of Linguistics, Macquarie University, Sydney.
http://www.shlrc.mq.edu.au/masters/806/slp806/prosody.html

Castejón, Federico, Gregorio Escalada, Luis Monzón, Miguel A. Rodríguez y P. Sanz Velasco, (1994). Un conversor texto-voz para el español, *Comunicaciones de Telefónica I+D*, 5, 2.114-131.
http://www.tid.es/presencia/publicaciones/comsid/esp/articulos/vol52/artic8/8.html

Courtois, Fabienne, Philippe Di Cristo, Benoît Lagrue y Jean Véronis (1997). Un modèle stochastique des contours intonatifs en français pour la synthèse à partir de textes. *4ème Congrès Français d'Acoustique*, Marsella (Francia). 373-376.

Crestani, Fabio (2001). Towards the use of prosodic information for spoken document retrieval. *SIGIR'01, 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. New Orleans, Louisiana. 420-421.
http://www.cs.strath.ac.uk/~fabioc/papers/01-sigir.pdf

Di Cristo, Albert, Philippe Di Cristo y Jean Véronis (1997). A metrical model of rhythm and intonation for French text-to-speech synthesis, En Antonis Botinis, Georgios Kouroupetroglou y George Carayiannis (eds.) *Intonation: Theory, Models and Applications. Proceedings of an ESCA Workshop*, Atenas: European Speech Communication Association. 83-86.

Di Cristo, Albert, Philippe Di Cristo, Estelle Campione y Jean Véronis (2000). A Prosodic Model for Text-to-speech Synthesis in French. En Antonis Botinis (ed.) *Intonation: Models and Theories*. Dordrecht: Kluwer Academic Publishers. 321-356.
http://www.up.univ-mrs.fr/~veronis/pdf/2000DiCristo.pdf

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

Dusterhoff, Kurt (2000). *Synthesizing Fundamental Frequency Using Models Automatically Trained from Data*. Tesis Doctoral, University of Edinburgh.
http://www.cstr.ed.ac.uk/publications/new/2000/Dusterhoff_2000_a.ps

Dutoit, Thierry (1997). *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers.

Dutoit, Thierry, Vincent Pagel, Nicolas Pierret, Olivier van der Vreken y François Bataille (1996). The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. En H. Timothy Bunnell y William J. Idsardi (Eds.) *Proceedings of the 1996 International Conference on Spoken Language Processing*, vol 3. Philadelphia: Penn. 1393-1397.
http://tcts.fpms.ac.be/publications/papers/1996/icslp96_tdvpnpfbovdv.zip

Estruch, Mònica. (2000). Évaluation de l'algorithme de stylisation mélodique MOMEL et du système de codage symbolique INTSINT avec un corpus de passages en catalan, *TIPA -Travaux Interdisciplinaires du laboratoire Parole et langage d'Aix-en-Provence* 19. 54-61.

Filipsson, Marcus y Bruce, Gösta (1997). "LUKAS - a preliminary report on a new Swedish speech synthesis", *Working Papers* 46:45-56, Department of Linguistics, Lund University.

Fischer-Jørgensen, Eli (1989). Phonetic analysis of the stød in standard Danish, *Phonetica* 46. 1-59.

Frid Johan (1999). An environment for testing prosodic and phonetic transcriptions. En John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville, y Ashlee C. Bailey (eds.) *Proceedings of ICPhS 1999. Proceedings of the XIVth International Congress of Phonetic Sciences*. University of California, Berkeley. vol 3: 2319.

Frid, Johan (2001). Prediction of intonation patterns of accented words in a corpus of read Swedish news. *Working Papers 49*, 42-45. Lund, Sweden: Department of Linguistics, Lund University.

Fujisaki Hiroya, Sumio Ohno y Takashi Yagi (1997). Analysis and modelling of fundamental frequency contours of Greek utterances. *Eurospeech'97. Proceedings of the 5th European Conference on Speech Communication and Technology*, vol. 1. 465–468.

Fujisaki, Hiroya (1996). Analysis and modeling of fundamental frequency contours of Korean utterances. A preliminary study. *Phonetics and Linguistics. In honour of Prof. H. B. Lee*. 640–657.

Fujisaki, Hiroya y Shin-ichi Nagashima (1969). A model for the synthesis of pitch contours of connected speech," *Annual Report of Engineering Research Institute, University of Tokyo*, vol. 28. 53–60.

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

Fujisaki, Hiroya y Sumio Ohno (1995). Analysis and modeling of fundamental frequency contours of English utterances. *Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Technology*, vol. 2. 985–988.

Fujisaki, Hiroya y Sumio Ohno (1996). Prosodic parameterization of spoken Japanese based on a model of the generation process of F0 contours. En H. Timothy Bunnell y William J. Idsardi (Eds.) *Proceedings of the 1996 International Conference on Spoken Language Processing*, vol 4. Philadelphia: Penn. 2439-2442.

Fujisaki, Hiroya, Keikichi Hirose, P. Hall´e y Hsi Lei (1990). Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese. *Proceedings of the 1990 Int'l Conf. Spoken Language Processing*, vol. 2. 841–844.

Fujisaki, Hiroya, P. Hall´e y Hsi Lei (1987). Application of F0 contour command-response model to Chinese tones. *Reports of the Autumn Meeting, Acoustical Society of Japan*, vol. 1. 197–198.

Fujisaki, Hiroya, Sumio Ohno y Changfu Wang (1998). A command-response model for F0 contour generation in multilingual speech synthesis. En *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, NSW, Australia, November 1998. 299-304.

Fujisaki, Hiroya, Sumio Ohno, Kei-ichi Nakamura, Miguelina Guirao y Jorge Gurlekian (1994). Analysis of accent and intonation in Spanish based on a quantitative model," *Proceedings of the 1994 International Conference on Spoken Language Processing*, vol. 1. 355–358.

Fujisaki, Hiroya, Sumio Ohno, Takashi Yagi y Takeshi Ono (1998). Análisis and interpretation of fundamental frequency contours of British English in terms of a command-response model. En Robert H. Mannell y Jordi Robert-Ribes (eds.) *Proceedings of the 1998 International Conference on Spoken Language Processing*, vol. 5. Sydney: Australian Speech Science and Technology Association, Incorporated (ASSTA).

Garrido, Juan María, Isabel Ortín, Silvia Quazza, Pierluigi Salza y Franca Mancini (2000). Desarrollo de un módulo de asignación de parámetros prosódicos para la versión en español del sistema de conversión texto-habla ACTOR®, *Procesamiento del Lenguaje Natural*, 26. 183-190.

Garrido, Juan María, Joaquim Llisterri, Rafael Marín, Carme de la Mota y Antonio Ríos (1995). Prosodic markers at syntactic boundaries in Spanish. En Kjiell Elenius y Peter Branderud (eds.) *ICPhS 95, Proceedings of the XIIIth International Congress of Phonetic Sciences*. Estocolmo, Suecia, vol. 2. 370-373.
http://liceu.uab.es/~joaquim/publicacions/Stockholm_95/stockholm_95.html

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

Gibbon, Dafydd, Inge Mertins, Roger K. Moore (eds.) (2000). *Handbook of multimodal and spoken dialogue systems*. Dordrecht: Kluwer Academic Publishers.

Goto, Masataka, Katunobu Itou y Satoru Hayamizu (1999). A Real-time Filled Pause Detection System for Spontaneous Speech Recognition. En Géza Gordos y Géza Németh (eds.) *Eurospeech'99. Proceedings of the 6th European Conference on Speech Communication and Technology*, vol. 1. Budapest: European Speech Communication Association. 277-230.

Greenberg, Steven (2001). From here to utility - Melding phonetic insight with speech technology. En Paul Dalsgaard, Børge Lindberg, Henrik Benner y Zheng-Hua Tan (eds.) *Eurospeech 2001. Proceedings of the 7th European Conference on Speech Communication and Technology*, vol 4. 2485-2488.
http://www.icsi.berkeley.edu/~steveng/PDF/Utility.pdf

Grønnum, Nina (1995). Superposition and subordination in intonation – a non linear approach. En Kjiell Elenius y Peter Branderud (eds.) *ICPhS 95, Proceedings of the XIIIth International Congress of Phonetic Sciences*. Estocolmo, Suecia, vol. 2. 124-131.

Grønnum, Nina (1998). Intonation in Danish. en Daniel Hirst y Albert Di Cristo (eds.) *Intonation Systems*. Cambridge University Press: Cambridge. 131 – 151.

Gurlekian, Jorge A., Hernán Rodríguez, Laura Colantoni y Humberto Torres (2001). Development of a prosodic database for an Argentine Spanish text to speech system. En *Proceedings of the IRCS Workshop on Linguistic Databases.* University of Pennsylvania, Philadelphia.
http://www.ldc.upenn.edu/annotation/database/papers/Gurlekian_etal/33.3.pdf

Gutiérrez, Juana Mª, Juan Manuel Montero, David Saiz y José M. Pardo (2001). New Rule-Based and Data-Driven Strategy to Incorporate Fujisaki´s F0 Model to a Text-To-Speech System in Castillian Spanish. En *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*. 821-824.

Hansson, Petra (1999). Prosodic Correlates of Discourse Markers in Dialogue. En *Proceedings of the ESCA Tutorial and Research Workshop on Dialogue and Prosody*, Holanda. 99-104.

Higuchi, Norio, Toshio Hirai y Yoshinori Sagisaka (1997). Effects of Speaking Style on Parameters of Fundamental Frequency Contour. En Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, y Julia Hirschberg (eds.) *Progress in Speech Synthesis*. Springer-Verlag. 417-428.

Hirst, Daniel (2000). ProZed: a multilingual prosody editor for speech synthesis. En Chambers (ed.) *Proceedings of the IEE Workshop on the State of the Art in Speech Synthesis*. Londres.

Hirst, Daniel (2002). Automatic analysis of prosody for multilingual speech corpora. En Eric Keller, Gérard Bailly, Alex Monaghan, Jacques Terken y Mark Huckvale (eds.) *Improvements in Speech Synthesis. Cost 258: The Naturalness of Synthetic Speech*. Chichester: John Wiley & Sons. 320-327.

Horne, Merle (ed.) (2000). *Prosody: Theory and Experiment. Studies presented to Gösta Bruce.* (*Text, Speech and Language Technology*. 14). Dordrecht: Kluwer Academic Publishers.

Jilka, Matthias, Gregor Möhler y Grzegorz Dogil (1999). Rules for the generation of ToBI-based American English intonation. *Speech Communication*. 28. 83-108.

Keller, Eric, Brigitte Zellner y Stefan Werner (1997). Improvements in prosodic processing for speech synthesis. En George Kokkinakis, Nikos Fakotakis y Evangelos Dermatas (eds.) *Proceedings of the COST Speech Technology in the Public Telephone Network: Where are we Today?* Rodas, Grecia. 73-84.

Keller, Eric, Gérard Bailly, Alex Monaghan, Jacques Terken y Mark Huckvale (eds.) (2002). *Improvements in Speech Synthesis. COST 258. The Naturalness of Synthetic Speech*. Chichester: Wiley & Sons.

Klabbers, Esther (2000). *Segmental and Prosodic Improvements to Speech Generation*. Tesis doctoral, Eindhoven University of Technology.

Klatt, Dennis (1976). "Linguistic uses of segmental duration in English: acoustic and perceptual evidence". *Journal of the Acoustical Society of America* 59, 5. 1208-1221. En Raymon Kent, Bishnu Atal y Joanne Miller (eds.) (1991) *Papers in Speech Communication: Speech Production*. Nueva York: Acoustical Society of America. 503-516.

Klatt, Dennis (1980). "Software for a Cascade/Parallel Formant Synthesizer", *Journal of the Acoustical Society of America* 67, 3: 971-995. En Raymon Kent, Bishnu Atal y Joanne Miller (eds.) (1991) *Papers in Speech Communication: Speech Production*. Nueva York: Acoustical Society of America. 765-789.

Klatt, Dennis (1987). "Review of Tex-to-Speech Conversion for English". *Journal of the Acoustical Society of America* 82,3. 737-793. En Raymon Kent, Bishnu Atal y Joanne Miller (eds.) (1991) *Papers in Speech Communication: Speech Production*. Nueva York: Acoustical Society of America. 57-114.

Kochanski, Greg y Chilin Shih (2001). *Prosody modelling with soft templates*.. *Speech Communication* 39. 311-352.

Koumpis, Konstantinos y Steve Renals (2001). The role of prosody in a voicemail summarization system. En Michiel Bacchiani, Julia Hirschberg, Diane Litman y Mari Ostendorf

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

(eds.) *Proceedings of the ISCA Tutorial and Research Workshop on Prosody and Speech Recognition*. Red Bank, N.J: International Speech Communication Association.
http://www.dcs.shef.ac.uk/~sjr/pubs/2001/pros01-vm.html

Kurzweil, Raymond (1998). When Will HAL Understand What We Are Saying? Computer Speech Recognition and Understanding. En David G. Stork(ed.) *Hal's Legacy: 2001's Computer as Dream and Reality.* Cambridge, MA: The MIT Press.
http://mitpress.mit.edu/e-books/Hal/chap7/seven1.html

Lea, Wayne A. (1980). Prosodic aids in speech recognition. En Wayne A. Lea (ed.) *Trends in Speech Recognition*. Englewood Cliffs, N.J.: Prentice-Hall. 166-205.

Levinson, Stephen E. y Mark Liberman (1981). Speech Recognition by Computer. *Scientific American* 244. 64-76. Trad. cast. de Ramón Cerdà: "Reconocimiento del habla por medio de ordenadores", *Investigación y Ciencia,* 1981. 38-51. En Joaquim Agulló (ed.) (1989) *Acústica musical.* Barcelona: Prensa Científica (Libros de Investigación y Ciencia). 106-121.

Levit, Michael, Richard Huber, Anton Batliner y Elmar Noeth (2001). Use of prosodic characteristics for automated detection of alcohol intoxication. En Michiel Bacchiani, Julia Hirschberg, Diane Litman y Mari Ostendorf (eds.) *Proceedings of the ISCA Tutorial and Research Workshop on Prosody and Speech Recognition*. Red Bank, N.J.: International Speech Communication Association. 103 - 106.
http://www5.informatik.uni-erlangen.de/literature/ps-dir/2001/Levit01:UOP.ps.gz

López Gonzalo, Eduardo, José Ignacio Villar Navarro y Luis Antonio Hernández Gómez (2002). Automatic Prosody Modeling of Galician and its Application to Spanish. En Keller, Eric, Gérard Bailly, Alex Monaghan, Jacques Terken y Mark Huckvale (eds.) *Improvements in Speech Synthesis. COST 258. The Naturalness of Synthetic Speech*. Chichester: Wiley & Sons. 218-227.

Llisterri, Joaquim (2001). La conversión de texto en habla. *Quark. Ciencia, Medicina, Comunicación y Cultura* 21. 79-89.
http://liceu.uab.es/~joaquim/publicacions/Quark2001/CTH_Quark_01.pdf

Llisterri, Joaquim (2002). Las tecnologías del habla: Entre la ingeniería y la lingüística, *Congreso Internacional La Ciencia ante el Público. Cultura humanística y desarrollo científico y tecnológico*. Universidad de Salamanca, Salamanca. 51-74.
http://liceu.uab.es/~joaquim/publicacions/TecnolHab_Salamanca_02.pdf

Llisterri, Joaquim, Lourdes Aguilar, Juan M. Garrido, María Machuca, Rafael Marín, Carme de la Mota y Antonio Ríos (1999). Fonética y tecnologías del habla. En José Manuel Blecua, Gloria Clavería, Carlos Sánchez y Joan Torruella (eds.) *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona - Editorial Milenio. 449-479.
http://liceu.uab.es/~joaquim/publicacions/Fonetica_TecnolHabla.pdf

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

Malfrère, Fabrice, Thierry Dutoit y Piet Mertens (1998a). Automatic prosody generation using suprasegmental unit selection. En *Proceedings of the 3rd ESCA/COCSADA Workshop on Speech Synthesis*, Jenolan Caves, Australia. 323-328.
http://tcts.fpms.ac.be/publications/papers/1998/jenolan98_fmtdpm.zip

Malfrère, Fabrice, Thierry Dutoit y Piet Mertens (1998b). Fully Automatic Prosody Generator for Text-to-Speech Synthesis. En Robert H. Mannell y Jordi Robert-Ribes (eds.) *Proceedings of the 1998 International Conference on Spoken Language Processing*, vol. 4. Sydney: Australian Speech Science and Technology Association, Incorporated (ASSTA). 1395-1398.
http://tcts.fpms.ac.be/publications/papers/1998/icslp98_fmtdpm.zip

Malfrère, Fabrice, Thierry Dutoit y Piet Mertens (1998c). Un générateur de prosodie "tout automatique". En *Actes des XXII Journées d'Etude sur la Parole*, Martigny, Suisse. 147-150.
http://tcts.fpms.ac.be/publications/papers/1998/jep98_fmtdpm.zip

Mertens, Piet (1990). L'Intonation. En Claire Blanche Benveniste, Mireille Bilger, Christine Pouget y Karel van den Eynden (eds.) *Le français parlé, études gramaticales*, cap. IV. Paris: Éditions du CNRS. 159-176.

Mertens, Piet (1999). Un algorithme pour la génération de l'intonation dans la parole de synthèse. *Actes Conférence TALN 99*, Cargèse. 233-242.
http://bach.arts.kuleuven.ac.be/~piet/papers/taln99.pdf

Mertens, Piet (2002). Synthesizing Elaborate Intonation Contours in Text-to-Speech for French. En Bernard Bel y Isabelle Marlien (eds.) *Proceedings of the Speech Prosody 2002 Conference*. Aix-en-Provence: Laboratoire de la Parole et du Langage. 499-502.
http://www.lpl.univ-aix.fr/sp2002/pdf/mertens.pdf

Minker, Wolfgang y Samir Bennacef (2001). *Parole et dialogue homme-machine.* Paris: Étidions Eyrolles - Éditions du CNRS (Sciences et techniques de l'ingénieur).

Mixdorff, Hansjörg (1997). Production of Broad and Narrow Focus in German - A Study Applying a Quantitative Model. En Antonis Botinis, Georgios Kouroupetroglou y George Caryannis (eds.) *Theory, Models and Applications. Proceedings of an ESCA Workshop.* Atenas: European Speech Communication Association, 239-242.

Mixdorff, Hansjörg (2002). Speech Technology, ToBI and Making Sense of Prosody. En Bernard Bel y Isabelle Marlien (eds.) *Proceedings of the Speech Prosody 2002 Conference*. Aix-en-Provence: Laboratoire de la Parole et du Langage. 31-38
http://www.lpl.univ-aix.fr/projects/aix02/sp2002/pdf/mixdorff.pdf

Mixdorff, Hansjörg y Hiroya Fujisaki (1994). Analysis of voice fundamental frequency contours of German utterances using a quantitative model. En Hiroya Fujisaki (ed.) *Proceedings of the 1994 International Conference on Spoken Language Processing*, vol. 4. 2231–2234.

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

Möbius, Bernd (1993). *Ein quantitatives Modell der deutschen Intonation - Analyse und Synthese von Grundfrequenzverläufen*, Tübingen: Niemeyer.

Montero, Juan Manuel, Juana María Gutiérrez-Arriola, José Colás, Emilia Enríquez, José Manuel Pardo (1999). Analysis and Modeling of Emotional Speech in Spanish. En John Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville y Ashlee C Bailey (eds.) *Proceedings of the XIVth International Congress of Phonetic Science*, vol. II. Berkeley: UC Berkeley, Dept of Linguistics. 957-960.

Montero, Juan Manuel, Juana María Gutiérrez-Arriola, Sira Palazuelos, Emilia Enríquez, Santiago Aguilera, José Manuel Pardo (1998). Emotional Speech Synthesis: from Speech Database to TTS. En Robert H. Mannell y Jordi Robert-Ribes (eds.) *Proceedings of the 5th International Conference on Spoken Language Processing*, vol. 3. Sydney: Australian Speech Science and Technology Association, Incorporated (ASSTA). 923-926.

Mota, Carme de la (1995). *La representación gramatical de la información nueva en el discurso.* Tesis doctoral, Universitat Autònoma de Barcelona.

Mota, Carme de la (1997). Prosody of Sentences with Contrastive New Information in Spanish. En Antonis Botinis, Georgios Kouroupetroglou y George Caryannis (eds.) *Theory, Models and Applications. Proceedings of an ESCA Workshop.* Atenas: European Speech Communication Association. 75-78.

Navas, Eva, Inma Hernáez, Ana Armenta, Borja Etxebarria y Jasone Salaberria (2000). Modelling Basque intonation using Fujisaki's models and CARTs. En Chambers (ed.) *Proceedings of the IEE Workshop on the State of the Art in Speech Synthesis*. Londres. 3/1-3/6.
http://bips.bi.ehu.es/ahoweb/files/publicaciones/SAISS2000.pdf

Olive, Joseph (1998). 'The Talking Computer': Text to Speech Synthesis. En David G. Stork (ed.) *Hal's Legacy: 2001's Computer as Dream and Reality.* Cambridge, MA: The MIT Press.
http://mitpress.mit.edu/e-books/Hal/chap6/six1.html

Pagel, Victor (1999). *De l'utilisation d'informations acoustiques suprasegmentales en reconnaissance de la parole continue.* Tesis doctoral. Université Henri Poincaré, Nancy.
http://vincent.pagel.free.fr/THESE/

Pardo, José Manuel, Francisco M. Giménez de los Galanes, José A. Vallejo, Miguel A. Berrojo, Juan M. Montero, Emilia Enríquez y Ángeles Romero (1995) Spanish text to speech: from prosody to acoustic. En *Proceedings of the International Conference on Acoustics*, vol. III. 133-136.

Pierrehumbert, Janet (1980). *The Phonology and Phonetics of English Intonation*. Tesis Doctoral, MIT.

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

Poggi, Isabella (2001). The Lexicon and the Alphabet of Gesture, Gaze and Touch. En *Proceedings of Third International Workshop,* Madrid, IVA 2001. 235-236.

Pols, Louis (2001). Acquiring and implementing phonetic knowledge. En Paul Dalsgaard, Børge Lindberg, Henrik Benner y Zheng-Hua Tan (eds.) *Eurospeech 2001. Proceedings of the 7th European Conference on Speech Communication and Technology*. Aalborg, Denmark. Vol 1. K3-K6.

Puigví, David, D. Jiménez y Josep M. Fernández (1994). Parametrización de las pausas ortográficas en castellano. Aplicación a un conversor de texto en habla". En *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Córdoba, 20-22 de julio de 1994.

Quazza, Silvia y Juan M. Garrido (1998). Prosody. En Marion Klein (Ed.) (1998) *Supported Coding Schemes*. MATE Deliverable D1.1. LE Telematics Project LE4 – 8370.
http://tsc.uvigo.es/~carmen/bego/data/D1.1-MATE-excerpt.pdf

Riera Masjoan, Montserrat y Elena Jiménez Ojea (2000). Corpus prosòdic. En *I Jornades del Centre de Referència en Enginyeria Lingüística (CREL)*. Institutd'Estudis Catalans, Barcelona.
http://liceu.uab.es/~joaquim/teaching/Language_resources/spoken_res/CREL/SFI_UAB_Corpus_Prosodic.pdf

Rodríguez Crespo, Miguel Ángel (1997). Introducción a la conversión texto-voz, *Philologia Hispalensis* 11, 2. 177-192.

Rodríguez Crespo, Miguel Ángel, Gregorio Escalada Sardina, Alejandro Macarrón Larumbe y Luis Monzón Serrano (1993). AMIGO: Un conversor texto-voz para el español, *Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural* 13. 389-400.

Ross, Ken y Mari Ostendorf (1999) A dynamical system model for generating fundamental frequency for speech synthesis, *IEEE Transactions on Speech and Audio Processing*. 7, 3. 295-309.

Rubio Ayuso, Antonio J. y Diego H. Milone (2002). Información prosódica y acentual para el reconocimiento automático del habla. En *Actas del II Congreso de Fonética Experimental*. Sevilla. 54-75.

Salvo Rossi, Pierluigi, Francesco Palmieri, Francesco Cutugno (2002). A method for automatic extraction of Fujisaki-model parameters. En Bernard Bel y Isabelle Marlien (eds.) *Proceedings of the Speech Prosody 2002 Conference*. Aix-en-Provence: Laboratoire de la Parole et du Langage. 615-618.
http://www.lpl.univ-aix.fr/sp2002/pdf/salvorossi-etal.pdf

Shriberg, Elisabeth y Andreas Stolcke (2001). Prosody Modeling for Automatic Speech Understanding: An Overview of Recent Research at SRI. En Michiel Bacchiani, Julia

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

Hirschberg, Diane Litman y Mari Ostendorf (eds.), *Proceedings of the ISCA Tutorial and Research Workshop on Prosody and Speech Recognition and Understanding*. Red Bank, N.J.: International Speech Communication Association.13-16.
http://www.speech.sri.com/papers/prosody2001-overview.ps.gz

Silverman, Kim, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert y Julia Hirschberg (1992). ToBI: a standard for labelling English prosody. En Robert H. Mannell y Jordi Robert-Ribes (eds.) *Proceedings of the 1998 International Conference on Spoken Language Processing*, vol. 2. Sydney: Australian Speech Science and Technology Association, Incorporated (ASSTA). 867-870.

Syrdal, Ann, Gregor Moehlerand, Kurt Dusterhoff, Alistair Conkie y Alan W Black (1998). Three methods of intonation modeling. En *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, NSW, Australia, November 1998, Paper J.3 (R54).

Tams, Andy y Mark Tatham, (2000). Intonation for synthesis of speaking styles. En *IEE Seminar on State-of-the-Art in Speech Synthesis* (Ref. No. 2000/58).
http://www.iee.org/oncomms/pn/humanfactors/download.cfm?ID=583F723F-8731-4520-9DCE1392DB9491AB

Taylor, Paul (1994). The Rise/Fall/ Connection Model of Intonation. *Speech Communication* 15. 169-186.

Thorsen, Nina (1979). Interpreting raw fundamental frequency tracings in Danish. *Phonetica 36*. 57-78.

Thorsen, Nina (1980). A study of the perception of sentence intonation. Evidence from Danish. *Journal of the Acoustical Society of America* 67. 1014-1030.

Thorsen, Nina (1983). Standard Danish sentence intonation-Phonetic data and their representation. *Folia linguistica* 17. 187-220.

Thorsen, Nina (1985). Intonation and text in Standard Danish. *Journal of the Acoustical Society of America* 77. 1205-1216.

Thorsen, Nina (1986). Sentence intonation in textual context – supplementary data. *Journal of the Acoustical Society of America 80*. 1041-1047.

Véronis, Jean, Daniel Hirst, Robert Espesser y Nancy Ide (1994). NL and speech in the MULTEXT project. En Paul McKevitt (ed.) *Proceedings of the 1994 American Association for Artificial Intelligence (AIII-94) Workshop on Integration of Natural Language and Speech Processing*. Seattle. 72-78.

Véronis, Jean, Philippe Di Cristo, Fabienne Courtois y Cédric Chaumette (1998). A stochastic model of intonation for text-to-speech synthesis. *Speech Communication, 26*, 4. 233-244.

LLISTERRI, J. - MACHUCA, M.J.- de la MOTA, C.- RIERA, M.- RÍOS, M. (2003) "Entonación y tecnologías del habla", in PRIETO, P. (Ed.) *Teorías de la entonación.* Barcelona: Ariel (Ariel Lingüística). pp. 209-243.
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicaciones.pdf

Wang, Chao (2001). *Prosodic Modeling for Improved Speech Recognition and Understanding*. Tesis doctoral, MIT.
http://www.sls.lcs.mit.edu/sls/publications/2001/Wang_phd_thesis.pdf

Wells, John (2002). *SAMPA Computer Readable Phonetic Alphabet.*
http://www.phon.ucl.ac.uk/home/sampa/home.htm

Wightman, Colin (2002). "ToBI or not ToBI?". En Bernard Bel y Isabelle Marlien (eds.) *Proceedings of the Speech Prosody 2002 Conference*. Aix-en-Provence: Laboratoire de la Parole et du Langage.
http://www.lpl.univ-aix.fr/sp2002/pdf/wightman.pdf

Willems, Nico, Collier, René y Johan t'hart (1988). A synthesis scheme for British English intonation. *Journal of the Acoustical Society of America*, 84. 1250-1261.

Zue, Victor (1999). Talking with your computer., *Scientific American*, August 1999. 40-41.
http://www.sciam.com/article.cfm?articleID=0009D2B7-F2E6-1C72-9B81809EC588EF21&catID=2