ESTILIZACION DE PATRONES MELODICOS DEL ESPAÑOL PARA SISTEMAS DE CONVERSION TEXTO-HABLA

Juan María Garrido Almiñana

Departament de Filologia Espanyola, Facultat de Filosofia i Lletres
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona)

INTRODUCCION

En un conversor texto-habla, la **inteligibilidad** del habla generada se relaciona principalmente con la síntesis de los diferentes elementos segmentales - los sonidos que componen la cadena hablada -, en tanto que la **naturalidad** depende directamente de la correcta superposición de los elementos suprasegmentales - acento y entonación, fundamentalmente - , sobre dicha cadena. Normalmente, el habla generada por los sistemas de conversión texto-habla es más inteligible que natural, dado que la información que se posee sobre el funcionamiento de los fenómenos suprasegmentales es en general menor que para los elementos segmentales.

Por todo esto, en el Laboratorio de Fonética de la Universitat Autónoma de Barcelona se ha iniciado una línea de trabajo, dentro del marco general de la investigación en síntesis de habla ya en marcha desde hace algún tiempo, acerca de las posibilidades de utilización de la información entonativa en sistemas automáticos de síntesis y reconocimiento de habla. En las páginas que siguen se muestran algunos de los resultados obtenidos y su posible aplicación a un sistema de conversión texto-habla.

En estos sistemas, el control de los fenómenos suprasegmentales está a cargo normalmente de un **módulo prosódico**, que tiene como misión, por un lado, generar la curva melódica correspondiente a cada enunciado, y por otro, determinar la **duración** y **amplitud** de los diferentes sonidos.

La generación de curvas melódicas en los módulos prosódicos puede realizarse de acuerdo a diferentes estrategias. En algunos sistemas, como el desarrollado por Pierrehumbert (Pierrehumbert, 81) para el inglés, la curva melódica es el resultado de la yuxtaposición e interpolación de una serie de tonos que se han asignado a cada sílaba, mediante diferentes reglas, según sean éstas tónicas o átonas, y por su posición en la frase. En otros, como el de O´Shaughnessy (O´Shaughnessy, 87), para el inglés, o el de Olabe, para el español (Olabe,83), hay definidos diferentes patrones melódicos básicos para cada tipo de enunciado, que se modifican por regla para incluir la información acentual.

Los patrones melódicos que se emplean en estos sistemas no son curvas melódicas reales, sino **representaciones estilizadas**, líneas esquematizadas en las que sólo se mantienen aquellas variaciones de la Fo que se consideran significativas.

Los trabajos que se presentan a continuación tienen que ver con estas dos cuestiones. Así, en el primer apartado, se describe el desarrollo de un procedimiento de estilización de curvas melódicas adecuado para lenguas como el español o el catalán. En el segundo, se presenta una propuesta de patrones melódicos básicos adecuados para la generación automática de curvas estilizadas en un conversor texto-habla para el español.

ESTILIZACION DE CURVAS MELODICAS

El primer paso en la definición del procedimiento fue la elección de la unidad de análisis. En este sentido, las aproximaciones realizadas hasta el momento al análisis y síntesis de las curvas melódicas han tomado dos caminos diferentes:

a) En unos casos, la curva melódica se ha considerado como una sucesión de tonos independientes, cada uno asociado a una sílaba. Este es el enfoque que subyace en el ya mencionado trabajo de Pierrehumbert, o en el análisis de las curvas melódicas del español presentado en (Quilis, 81). El método de representación utilizado en estos casos se denomina **por niveles**, puesto que las curvas melódicas se presentan como una serie de niveles discretos de tono.

b) En otros, la curva melódica se ha interpretado como una variación continua de la Fo a lo largo de todo el enunciado. En estos estudios los resultados han sido una serie de esquemas que representan las variaciones de la Fo a lo largo de todo un grupo fónico. Es la aproximación realizada por O´Shaughnessy, por Thorsen (Thorsen, 79) para el danés o por Navarro Tomás (Navarro Tomás, 48) para el español. Las representaciones obtenidas en estos casos se denominan **por contornos**, porque la curva se presenta de forma continua, sin saltos bruscos de tono.

La elección del primer enfoque implica, pues, un análisis a nivel silábico, en tanto que el segundo toma como unidad de análisis la curva de Fo a lo largo de todo el grupo fónico. Dos razones principalmente llevaron a la elección del segundo enfoque para nuestro trabajo:

a) En primer lugar, porque el primer método parece más adecuado para lenguas en las que la forma de la curva es muy dependiente de la posición de las sílabas tónicas y átonas, como en inglés. El segundo, en cambio, se ajustaría más a las características del español o el catalán, lenguas en las que acento implica sólo raras veces una variación de la Fo.

b) En segundo lugar, porque el primer enfoque parece más difícil de formalizar, y por tanto de automatizar, que el segundo. El número de niveles que se van a utilizar y la definición de los límites de cada uno son cuestiones que en principio requieren un estudio acústico amplio de las curvas para su resolución.

La siguiente cuestión que se plantea es la determinación de las variaciones de la Fo que se han de eliminar y las que se han de mantener en la representación estilizada resultante. En este sentido, cabría distinguir entre dos tipos de variaciones:

a) **Variaciones que se dan dentro de la misma curva melódica**. No todas las variaciones de la Fo que se registran en una curva melódica son relevantes. De entrada, las variaciones inferiores a 1,5-3 semitonos en la curva de Fo ya no son percibidas por el oído humano (´t Hart, 74). Por otro lado, algunas de las variaciones de la Fo que se registran a lo largo de la curva se deben a la naturaleza de los elementos segmentales que lo componen. Así, la Fo variará según el sonido sea una vocal o una consonante, o bien según el grado de abertura de las vocales, entre otros factores (Lehiste & Peterson, 61), (Di Cristo, 82). Estas variaciones son dependientes de cada enunciado, y según algunos autores (´t Hart & Collier, 75), tampoco son tenidas en cuenta por los oyentes al analizar perceptivamente las curvas melódicas, por lo que no habrían de mantenerse en una representación estilizada. Desde este punto de vista, el objetivo es obtener una representación en la que sólo se conserven las variaciones perceptivamente relevantes, al estilo de las obtenidas por ´t Hart y su grupo (´t Hart *et al.*, 90) para el holandés.

b) Por otro lado, una curva melódica puede presentar **variaciones inter-locutor**. De todas ellas, quizá la más importante sea las diferencias debidas al fundamental habitual de cada hablante. Por simplicidad y generalidad de las representaciones, es preferible aplicar algún proceso de **normalización frecuencial** que elimine este tipo de diferencias.

De acuerdo con todo esto, se determinó que el método de análisis de este estudio utilizase representaciones de contornos, y no de niveles, que eliminase las variaciones de Fo no relevantes perceptivamente, y que proporcionase representaciones independientes de las características del locutor que las haya emitido. Además, se pretende que las representaciones obtenidas conserven únicamente las variaciones relevantes lingüísticamente, es decir, aquellas que de alguna manera impliquen la transmisión de algún tipo de información al oyente. Este enfoque se encuentra implícito en alguno de los estudios anteriores de la entonación del español - Navarro Tomás, Toledo (Toledo & Gurlekian, 90), Olabe -, pero no se ha llegado a formular de forma explícita.

La base del procedimiento consiste en mantener únicamente los valores de la Fo en los **puntos de inflexión**, que serían aquellos puntos de la curva en los que la pendiente cambia de signo (de positivo a negativo, de positivo a 0, de 0 a negativo, etc). De esta forma se mantienen los valores de los picos y los valles a lo largo de la curva, que luego pueden interpolarse mediante líneas, ya sean rectas o sinusoides. Sin embargo, no todos los cambios de signo a lo largo de una curva han de mantenerse; entre uno y otro punto de inflexión debe haber una variación mínima de Fo (mayor que un **umbral** preestablecido, que para las primeras pruebas se estableció en 10 Hz), de manera que las variaciones no perceptibles o las relacionadas con la micromelodía quedarían fuera de la representación. Posteriormente, se aplica a cada representación obtenida un proceso de **normalización frecuencial**, por el que los valores de los puntos de inflexión se relativizan con respecto al valor de la Fo en el inicio de la curva. Dicho proceso puede formalizarse en la expresión:

$$Fo_n\,nor = Fo_n - Fo_i$$

donde Fo_n es el valor de la Fo en un punto determinado, Fo_i es el valor de la Fo al inicio de la curva, y Fo_n nor es el valor normalizado de Fo correspondiente a Fo_n.

La aplicación de este método de estilización a una serie de curvas extraídas de un *corpus* de oraciones sencillas del español (para una descripción más detallada del mismo, ver apartado siguiente) dio como resultado un conjunto de representaciones como la que aparece en la figura 1. El análisis de estas representaciones muestra que, en general, se obtiene la forma de la curva descrita en estudios anteriores para cada tipo de frase analizado. Sin embargo, en algunos casos, el umbral establecido inicialmente para la eliminación de las variaciones demasiado pequeñas no eliminó algunas de las inflexiones debidas claramente a la micromelodía, por lo que deberá ser revisado en posteriores estudios.

La validez de este método de representación debe ser refrendada mediante tests de percepción, por lo que actualmente se está desarrollando, en colaboración con Francesc Gudayol, ingeniero de Telecomunicaciones, un sistema que permitirá extraer representaciones estilizadas con este método de forma automática, y aplicarlas a los enunciados originales. Con él se podrá evaluar hasta qué punto dichas representaciones mantienen la información lingüística contenida en la curva original. Las representaciones también podrán ser manipuladas, lo cual permitirá realizar otros estudios que sirvan para refinar el método de estilización actual.

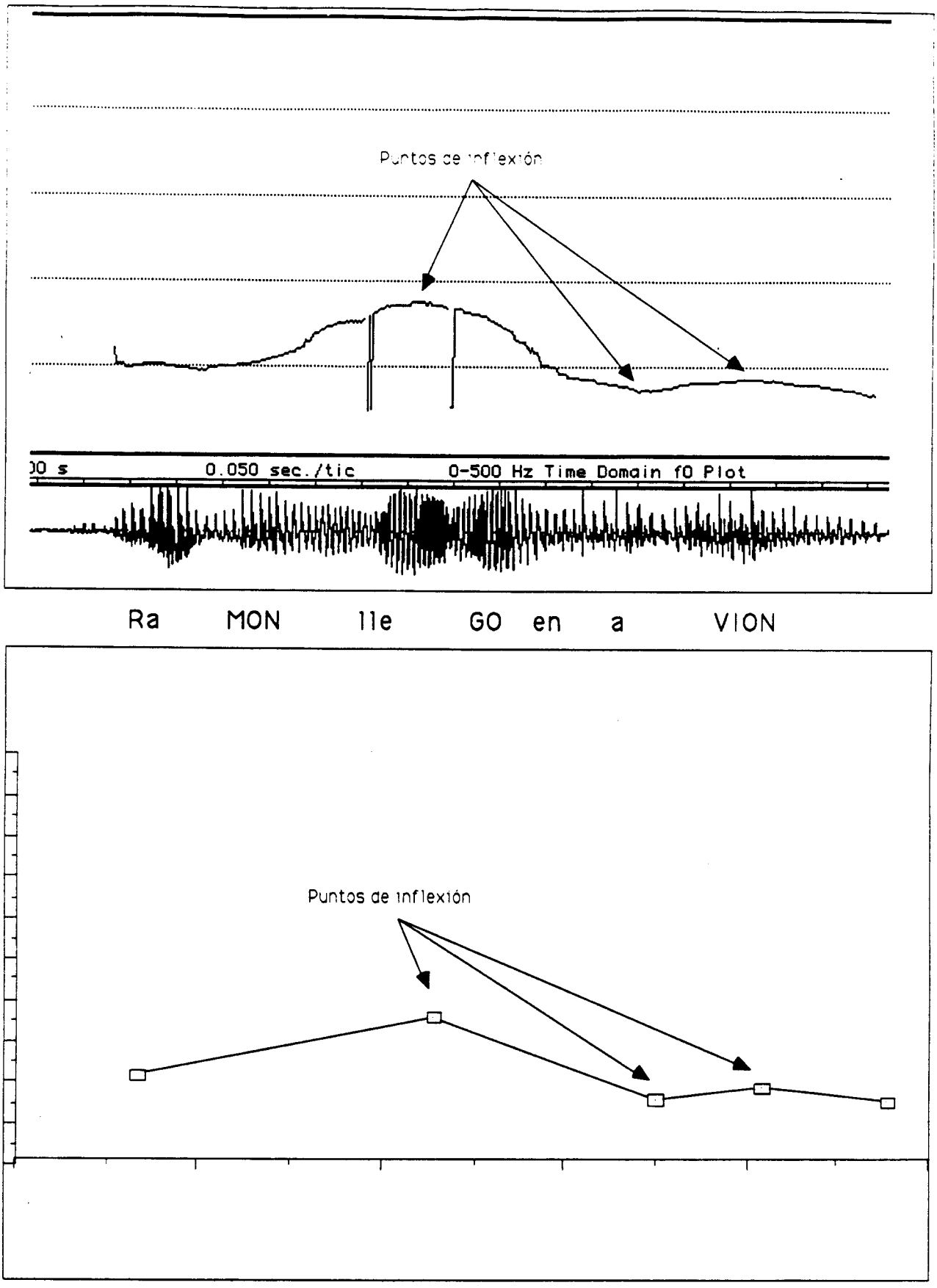

Figura 1: ejemplo de representación estilizada con su correspondiente curva original

PATRONES MELODICOS BASICOS DEL ESPAÑOL

Los patrones melódicos del español pueden clasificarse en dos grupos principales:

a) patrones terminales: aparecen en grupos fónicos situados al final de oraciones, o en grupos fónicos que incluyen una oración completa. Sirven, además de para indicar el final de las mismas, para transmitir información acerca del tipo de oración de que se trata: una afirmación,una pregunta, una exclamación, una orden... Este tipo de información es lo que en términos lingüísticos se denomina **modalidad oracional.**

b) patrones no terminales: aparecen en grupos fónicos situados en el interior de una oración, de manera que contienen normalmente frases subordinadas o sintagmas, nunca oraciones completas. Contienen información que permite al oyente deducir que la oración no ha terminado, y en algunos casos información acerca del tipo de relación sintáctica que se establece con los grupos fónicos anterior o posterior.

Los estudios de la entonación del español han dedicado más atención al primer grupo de patrones que al segundo, y de hecho no se ha realizado una distinción muy clara entre ambos tipos. Así ocurre, por ejemplo, en el trabajo ya mencionado de Navarro Tomás, el más completo sobre los patrones melódicos del español realizado hasta la fecha. Estudios como el de Quilis (Quilis, 81) sí analizan patrones de ambos tipos, aunque el inventario que presentan no es ni mucho menos exhaustivo.

Por otro lado, Navarro Tomás únicamente describe los diferentes tipos de curvas melódicas que pueden encontrarse en un determinado tipo de frase. Varios de los esquemas que se presentan para distintos tipos de frases guardan bastantes semejanzas. Sin embargo, no se ha hecho hasta ahora ningún intento de definir una serie de patrones melódicos básicos a partir de esta descripción.

Por tanto, la información de que se dispone acerca de los patrones melódicos del español es, con vistas a su incorporación a un sistema de síntesis, incompleta y poco sistematizada. El ya citado estudio de Olabe, realizado para el desarrollo del módulo de entonación del conversor texto-habla de la ETSIT de Madrid, de alguna manera llena este vacío, pero analiza únicamente los patrones descritos por Quilis en (Quilis, 81).

Otra de las tareas que se han emprendido en el Laboratorio de Fonética es, por tanto, una descripción de los diferentes esquemas melódicos del español, en un formato que sea utilizable para su aplicación en sistemas automáticos de reconocimiento o síntesis. El primer paso ha sido el estudio de los patrones terminales descritos en (Navarro Tomás, 48) y su reducción a una serie de reglas y patrones básicos, cuyos resultados se presentan a continuación.

El procedimiento de estilización descrito en el apartado anterior se aplicó a un *corpus* de frases pronunciadas por diferentes locutores. Dicho *corpus* contenía frases enunciativas, tres tipos diferentes de interrogaciones, dos tipos de exclamaciones, ruegos y mandatos, siguiendo la clasificación hecha por Navarro Tomás, aunque con algunas simplificaciones. Las oraciones estaban constituidas por un solo grupo fónico, para evitar la aparición de patrones no terminales. Contenían exclusivamente sonidos sonoros, que nos permitieran obtener los contornos completos, y fueron incluidas en diálogos para conseguir una realización más natural.

El análisis de las representaciones obtenidas incluyó diversos estudios estadísticos sobre la altura tonal y la posición de los puntos de inflexión. Tras dicho análisis se definieron **tres esquemas melódicos básicos** y una serie de recursos secundarios o **formas superpuestas**, que los locutores utilizaron para modificar estos esquemas básicos.

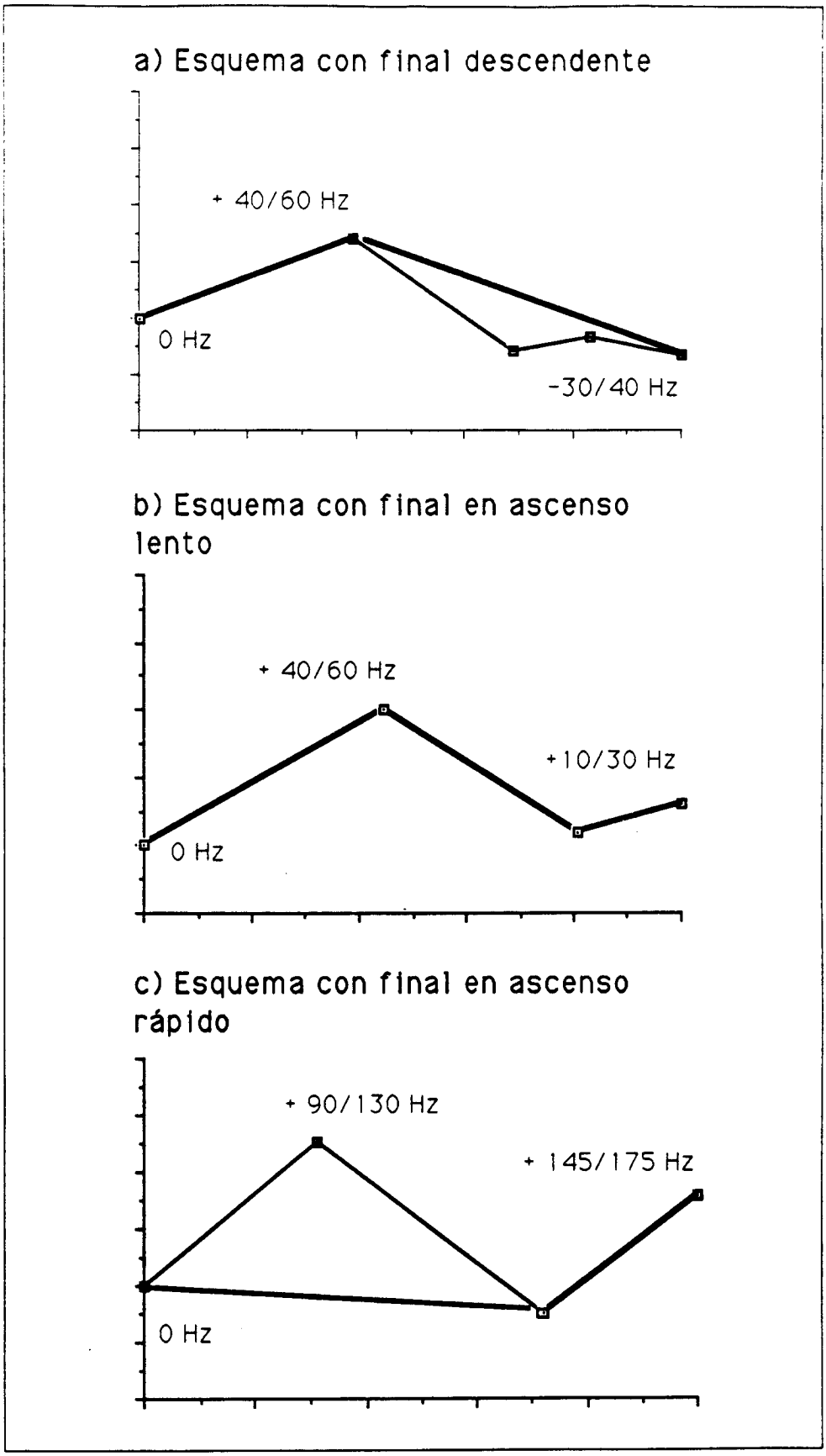

Figura 2: patrones melódicos básicos

Los tres esquemas melódicos básicos serían (ver figura 2):

1) Un esquema con final descendente, que se utilizaría en las frases enunciativas, en algunas realizaciones de interrogaciones parciales (las que exigen una respuesta diferente de "sí" o "no") y en algunas oraciones exclamativas,de mandato o de ruego. Sus rasgos fundamentales serían:

a) Un primer pico de Fo en la primera sílaba tónica o en la sílaba posterior a la misma. Este primer pico se elevaría unos 60 Hz por término medio sobre el nivel de Fo al principio de la curva.

b) Según la longitud del grupo fónico, otros picos secundarios, de menor altura que el primero, y que no necesariamente deben coincidir con sílabas tónicas. Los diferentes picos de la oración podrían unirse con una línea imaginaria descendente o de **declinación**, al estilo de las definidas para las curvas del inglés (Cooper & Sorensen, 81).

c) Un segmento final descendente (en algún caso excepcional, ascendente), cuyo final se situaría por debajo del valor de la Fo al inicio de la curva (unos 30 o 40 Hz), y que comenzaría normalmente en la última o penúltima sílaba tónica, o en la sílaba posterior.

2) Un esquema con final en ascenso lento, que se utilizaría en algunas frases exclamativas, de mandato o de ruego. Sus características principales serían:

a) Un primer pico, que se situaría normalmente en la primera sílaba tónica o en la sílaba posterior a la primera tónica, como en el esquema anterior. La altura también sería semejante.

b) Una serie de picos secundarios, cuyo número variará según la longitud del grupo, y que en principio seguirían también la línea de declinación descrita para el primer esquema.

c) Un segmento final ligeramente ascendente, cuyo final se situaría algo por encima del valor de la F0 en el inicio de la curva, y que se iniciaría en la última sílaba tónica o en la sílaba posterior.

3) Un esquema con final en ascenso rápido, que sería el propio de las oraciones interrogativas en general, y que se caracterizaría por:

a) La presencia, en las frases con más de una sílaba tónica, de un primer pico de Fo en la primera sílaba tónica o en la sílaba posterior a la misma. La altura de este pico sería superior (unos 30 o 40 Hz) a la del primer pico del esquema anterior.

b) Un segmento final ascendente, que comenzaría al final de la penúltima sílaba de la oración o al principio de la última, un poco por debajo del nivel inicial de Fo, y que ascendería rápidamente, hasta alcanzar al final de la curva valores superiores a los 100 Hz sobre el valor inicial de la Fo.

Como puede comprobarse, un mismo patrón puede corresponder a modalidades diferentes. Sobre estos esquemas básicos se aplicarían una serie de recursos secundarios o formas superpuestas para acabar de definir el tipo de frase (ver figuras 3a y 3b):

1) Una **elevación de la altura del primer pico** en el primer esquema, hasta una altura semejante a la del segundo esquema, en los casos de interrogaciones parciales con final descendente.

2) La colocación de un último pico sobre el primer esquema, de mayor amplitud que el resto de picos de la curva en el fragmento correspondiente a la última sílaba del grupo. Este recurso, que ha sido etiquetado como **esquema circunflejo** por algunos autores (Navarro Tomás, 48), se utilizaría en oraciones exclamativas, de mandato o de ruego como recurso diferenciador frente a las oraciones enunciativas.

3) El **aumento del número de picos** en el primer o tercer esquema, también como un recurso para marcar la presencia de una oración con cierto contenido expresivo (exclamativa, de orden o ruego).

4) Una **elevación del rango frecuencial** (la diferencia entre el pico más alto y el más bajo de la curva) en cualquiera de los tres esquemas, como una marca general de expresividad.

Una vez obtenidos estos patrones y reglas, el siguiente paso es realizar una **validación perceptiva** de los mismos. En concreto, está pendiente:

a) un análisis de la importancia de la información de niveles superiores al fonético (léxica y semántica, especialmente) en la identificación de los distintos tipos de frases: así, por ejemplo, habría que comprobar hasta qué punto influye la presencia de una partícula interrogativa en la identificación de una interrogativa parcial, cuando esta presenta una curva melódica con final descendente; o la presencia de un verbo en forma imperativa, para el etiquetado de una oración como imperativa.

b) una comprobación perceptiva de las curvas melódicas generadas mediante estas reglas y patrones.

A nivel acústico, está pendiente también un análisis de habla espontánea que permita comprobar la presencia de estos esquemas en oraciones "reales",
y un estudio sobre los patrones no terminales del español.

Este trabajo ha sido realizado con el soporte de una ayuda a investigadores jóvenes de la CIRIT de la *Generalitat de Catalunya*, y de una beca de formación de personal investigador concedida al Departamento de Filología Española de la *Universitat Autònoma de Barcelona*.

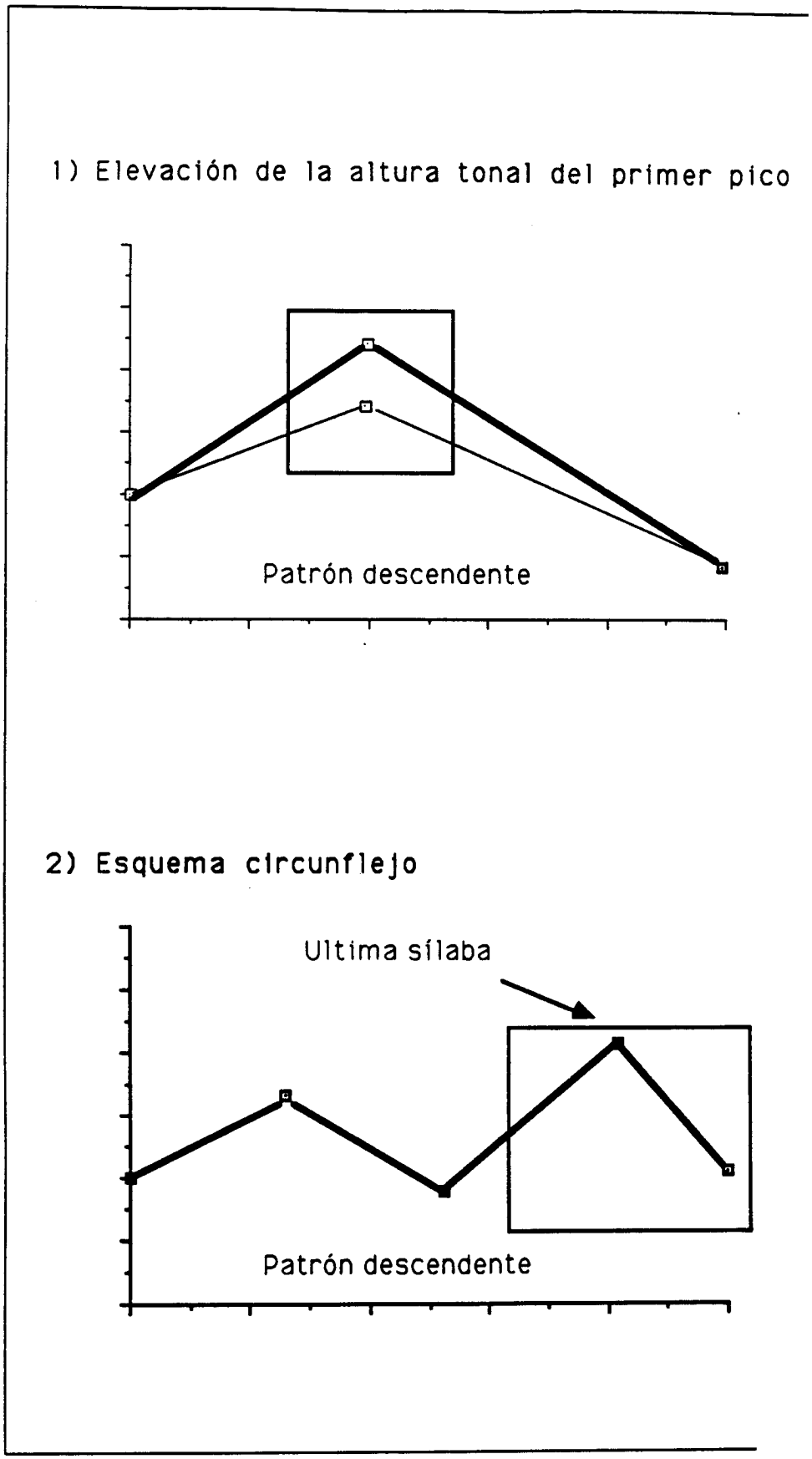

Figura 3a: esquemas superpuestos

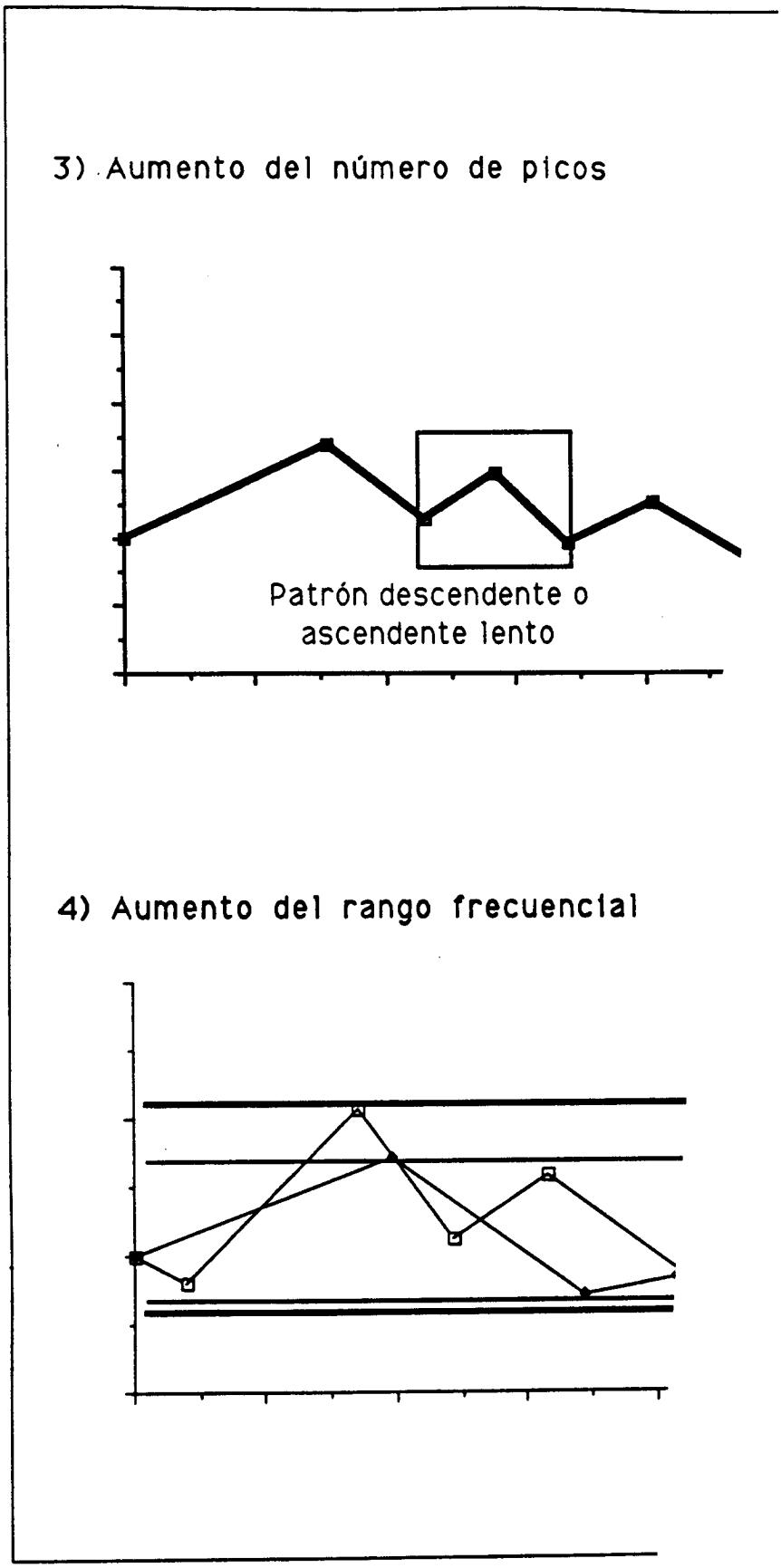

Figura 3b: esquemas superpuestos

REFERENCIAS

(Cooper & Sorensen, 81).- COOPER, W.E. - SORENSEN, J.M. (1981).- *Fundamental Frequency in Sentence Production*, New York: Springer Verlag.

(Di Cristo, 82).- DI CRISTO, A. (1982).- *Prolegomènes à l'étude de l'intonation. Micromélodie*, Paris: Ed. du CNRS.

(Lehiste & Peterson, 61).- LEHISTE, I. - PETERSON, G. (1961).- "Some basic considerations in the analysis of intonation", *Journal of the Acoustical Society of America*, 33, pp. 419-425.

(Navarro Tomás, 48).- NAVARRO TOMAS, T. (1948).- *Manual de entonación española*, Madrid: Guadarrama (4ª ed.).

(O'Shaughnessy, 87).- O' SHAUGHNESSY, D. (1987).- "The fundamental frequency generator", en ALLEN et al. (1987).- *From Text to Speech: the MITalk System*, Cambridge, MA: Cambridge University Press, pp. 100-107.

(Olabe, 83).-OLABE, J.C. (1983).- *Sistema para la conversión de un texto ortográfico a hablado en tiempo real*, Tesis doctoral, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid.

(Pierrehumbert, 81).- PIERREHUMBERT, J. (1981).- "Synthesizing intonation", *Journal of the Acoustical Society of America*, 70, 4, pp. 985-995.

(Quilis, 81).- QUILIS, A. (1981).- *Fonética acústica de la lengua española*, Madrid, Gredos.

(Thorsen, 79).- THORSEN, N. (1979).- "Interpreting Raw Fundamental-Frequency Tracings of Danish", *Phonetica*, 36, pp. 57-78.

(Toledo & Gurlekian, 90).- TOLEDO, G. - GURLEKIAN, J. (1990).- "Entonación del español: ¿existe la preplanificación?", *Estudios de Fonética Experimental*, IV, pp. 27-49.

('t Hart, 74).- 'T HART, J. (1974).- "Discriminabiliy of the size of pitch movements in speech", *IPO Annual Progress Report*, 9, pp.56-63.

('t Hart & Collier, 75).- ' T HART, J. - COLLIER, R. (1975).- "Integrating different levels of intonation analysis", *Journal of Phonetics*, 3, pp. 235-255.

('t Hart *et al.*, 90).- 'T HART, J. - COLLIER, R. - COHEN, A. (1990).- *A Perceptual Study of Intonation. An Experimental - Phonetic Approach to Intonation*, Cambridge: Cambridge University Press.