4. Sistemas de conversión de texto a voz

UN SISTEMA DE CONVERSION DE TEXTO A VOZ PARA CASTELLANO.

Juan C. Pérez
Enrique Vidal.

Dpto. de Sistemas Informáticos y Computación (DSIC)
Universidad Politécnica de Valencia.

Resumen

Se presenta un sistema de conversión de texto a voz sin restricciones para el castellano. La síntesis de voz se logra por codificación de la onda sonora en PCM (Pulse Coded Modulation). La transcripción ortográfico-fonética se lleva a cabo mediante reglas.
Las novedades más importantes residen en el uso de muestras naturales de voz, el tratamiento de la prosodia (entonación) y el funcionamiento en un ordenador doméstico de bajo coste sin soporte físico adicional. A esto hay que añadir la producción de una voz de inteligibilidad y naturalidad superiores a las obtenidas hasta ahora en equipos de características similares y aún superiores.

Introducción

El origen de la síntesis del habla que hoy conocemos se produce con el desarrollo de los sistemas digitales que permiten utilizar computadores para la generación y el control de los sonidos vocales. En este sentido las primeras aproximaciones consistieron en el almacenamiento digital de fragmentos de voz. Este sistema presenta dificultades de diverso tipo, sobre todo respecto a la alteración de la frecuencia fundamental, necesaria para conferir entonación a las frases. Sucesivos refinamientos, sustentados en gran medida por las aplicaciones civiles y militares de procesamiento de la señal para su transmisión, parecen conducir a la utilización del dominio de la frecuencia en lugar del dominio del tiempo, así como a la parametrización de la señal y su almacenamiento en forma de cadenas de vectores de parámetros. En este sentido aparecen técnicas como los vocoders, los sintetizadores por formantes, la síntesis por predicción lineal (LPC) y la síntesis homomórfica.

Los sistemas de síntesis más evolucionados son aquellos que permiten leer automáticamente, de viva voz, un texto de entrada en lenguaje natural. La conversión de texto a voz es una aplicación para la que que no se han obtenido todavía resultados totalmente satisfactorios. Esto se debe a las enormes dificultades que se plantean al intentar descubrir, directamente a partir del texto, los rasgos prosódicos a nivel de oraciones o párrafos completos. En la mayoría de los casos, esta componente de acentuación no puede ser interpretada sin una comprensión del texto, la cual se halla, por el momento, fuera de las posibilidades de la inteligencia artificial.

Las técnicas de síntesis de voz pueden clasificarse, por una parte, de acuerdo con el tamaño de las unidades que, concatenadas, van a constituir la señal acústica. En función de los resultados a obtener se optará por unidades mayores o menores. Las unidades de mayor tamaño exigen, obviamente, almacenar un mayor número de ellas. La elección tradicional en un sistema no restringido como el que nos ocupa oscila entre las sílabas, como unidad de mayor tamaño, y

los fonemas, como unidad menor, recayendo típicamente en las semisílabas o en los difonemas, de duración intermedia entre los extremos señalados.

Por otro lado, las diferentes formas de almacenar y reproducir la señal dan lugar a una nueva clasificación por este concepto. Las técnicas de codificación de la onda son las que se basan en almacenar un patrón de la propia señal vocal y restituirlo luego a través de algún tipo de subsistema acústico que reintegra esta onda. Los diversos métodos y sus variantes sólo difieren entre sí en el modelo de patrón utilizado para la codificación de la señal. En las técnicas de síntesis paramétrica no es la onda sonora, propiamente, la que se codifica sino un conjunto de características (parámetros) de ésta.

En concreto, el objetivo alcanzado con nuestro sistema ha sido la conversión de texto a voz para un texto en castellano sin ninguna restricción, siendo la producción de una **voz de calidad**, en una máquina de **bajo coste**, la aportación más importante. Una detallada y completa descripción del trabajo que presentamos se puede consultar en [Pérez,89].

Aspectos acústico fonéticos

La construcción del prototipo sobre un equipo de modesta potencia computacional exige reducir al mínimo los cálculos de suavización de transiciones y simulación de las características coarticulatorias. Esto se ha conseguido a través de la elección de unidades en las que la cadena fónica se divide por sus partes más estables.

Pretendíamos, también, satisfacer un requerimiento más importante: controlar algunos parámetros (típicamente el tono y la intensidad) de la zona que determina la inflexión tonal de la voz (la vocal). Así pues, se ha refinado más aún la subdivisión, dejando la zona en cuestión aislada y fácilmente controlable (susceptible de modificación). Las unidades obtenidas de esta forma estarían a medio camino entre el difonema y la semisílaba (con particularidades adicionales derivadas del control requerido para la prosodia). Este control se podrá conseguir a partir de **varios bancos de muestras pronunciadas en diferentes tonos.**

Pasamos a describir con detalle las unidades que se han usado.

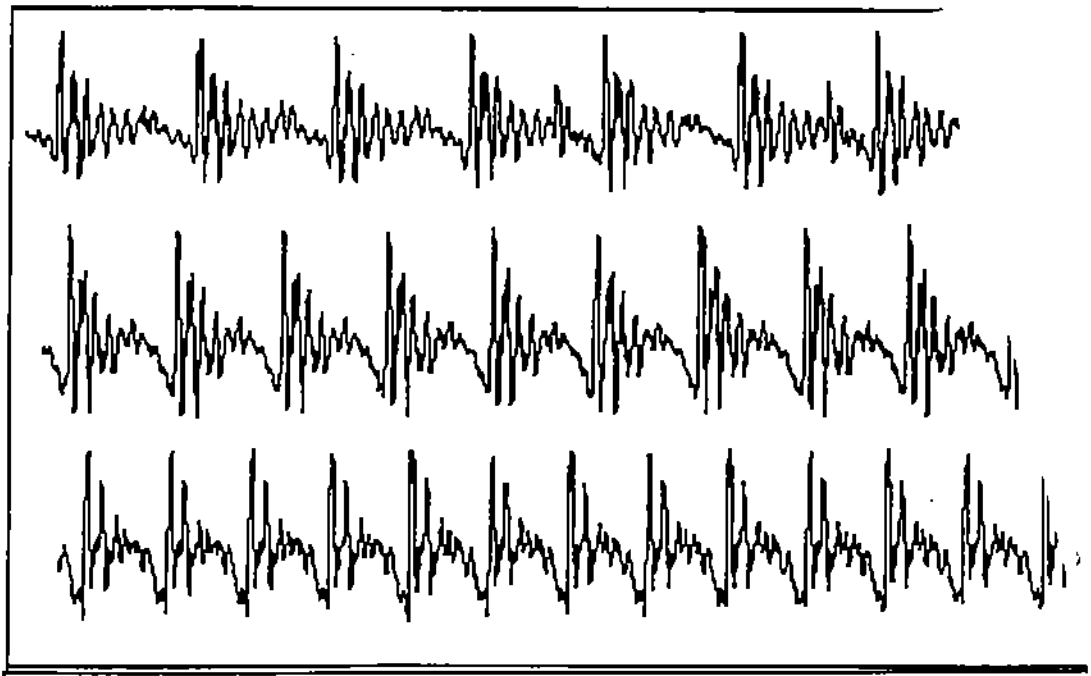


Figura 1.
Muestras correspondientes a la vocal "A" en tres tonos (bajo [1], medio[5] y alto[8])

○ Las vocales A .. U (se denotarán con mayúsculas para distinguirlas de las vocales anejas a cada grupo consonántico, que no tienen el mismo valor que éstas) son las encargadas de controlar la inflexión. Son las únicas susceptibles de variar su frecuencia fundamental. Como se ha dicho, para evitar los problemas derivados del cambio de la frecuencia fundamental y de los armónicos, que desvirtúan el resultado fónico, *se han muestreado varios juegos de vocales en las diferentes frecuencias a utilizar en el discurso (fig. 1).*

○ Los diptongos se muestrean dejando *la vocal fuerte con sólo una pequeñísima parte de su duración (fig.2)* (por ejemplo en **"ie"**, la muestra es una **i** con inflexión final conducente a **e**, la **e** casi no llega a percibirse). De este modo para pronunciar el diptongo **"ie"** se emitirán las muestras **"ie" "E"**. En el diptongo **"iu"** se considera la **u** vocal fuerte y en el **"ui"** es la **i** la vocal fuerte. *Los triptongos se forman con dos diptongos y la vocal fuerte central intercalada.* P. ej. **"iau"** se realiza como "ia" "A" "au".

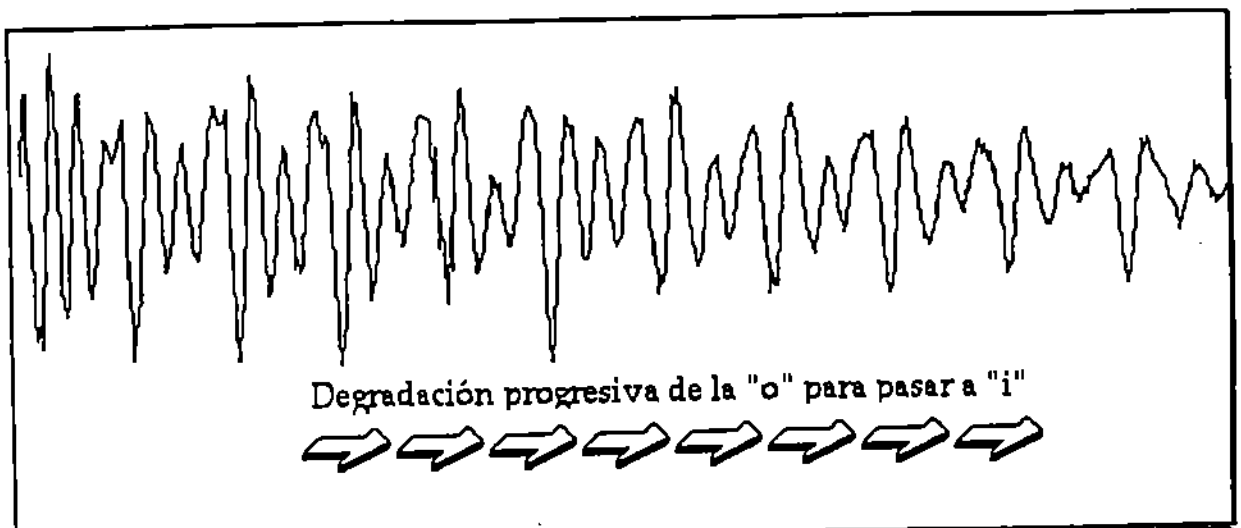


Figura 2.
Muestra completa correspondiente al diptongo "oi" .

○ *Las consonantes fricativas* (según el rasgo distintivo) *así como la vibrante "rr" y la "ch"*[5], se han muestreado de palabras que las contienen y *se han almacenado aisladas (fig.3).* Se concatenarán así con las unidades anteriores y posteriores sin más.

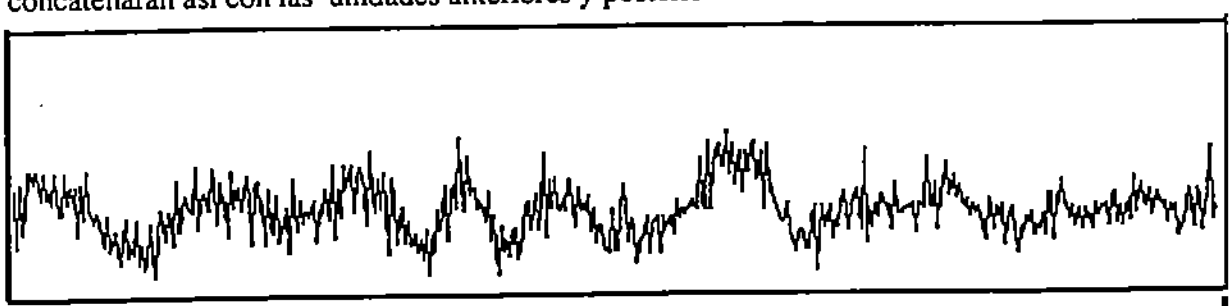


Figura 3.
Muestra completa correspondiente a la consonante aislada "z" .

○ *El resto de consonantes*, que irán en grupo con vocales, se muestrearán *antes y después de cada vocal (fig.4).*

5 Véase el punto "Variedades alofónicas" para obtener las equivalencias, en nomenclatura fonética normalizada, de los sonidos aquí representados.

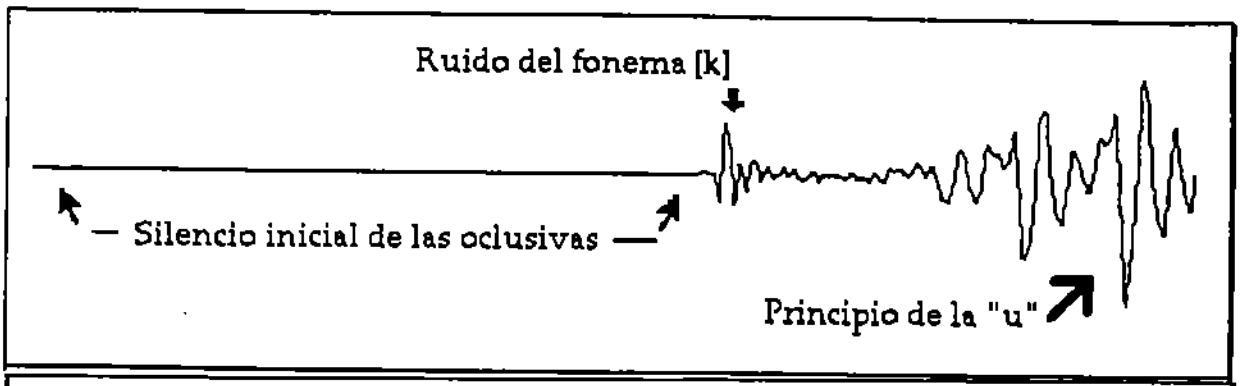


Figura 4.
Muestra completa correspondiente a la unidad consonántica "ku" .

Cada uno de estos grupos (o unidades difonemáticas: ba, be, etc.), hay que reiterar que se ha extraído de una voz real, cada uno por separado. Con todo esto tenemos un total de 6 + 130 + 14 = 150 unidades más las vocales.

Hay que señalar que, hasta el presente, no ha sido propuesta en la literatura una técnica que contemple el aspecto prosódico (y otros complementarios que se mencionan a continuación) de este mismo modo. Sin duda debido a esto, se ha podido conseguir una calidad de voz superior a la de otros trabajos realizados para funcionar en equipos de las mismas y, en algunos casos, superiores prestaciones.

Variedades alofónicas.

Hasta ahora hemos hablado de consonantes y vocales como si fueran éstas las que tomaran un carácter de realización fónica a la hora de hablar. En realidad, la pronunciación de una frase se suele representar como una sucesión de fonemas que da lugar al sonido que finalmente emitimos o escuchamos. Pero la definición lingüística del fonema no es sencilla [Alarcos,50].

Los fonemas no se corresponden, de ninguna manera, con los sonidos que se producen al pronunciar las diferentes combinaciones de letras que encontramos escritas. La determinación de los fonemas de una lengua se basa en una serie de cotejos entre palabras, tales que la distinción entre ellas se deba a los sonidos diferentes que las caracterizan. Esta oposición entre sonidos, o independencia fonológica, indica cuándo éstos son fonemas realmente distintos o meras **realizaciones diversas del mismo fonema** (alófonos). Los fonemas del castellano son cinco vocálicos y diecinueve consonánticos. Vamos a dar una lista de ellos en la que incluimos también las unidades que los representan así como las principales variedades alofónicas que se distinguen de cada uno:

Fonema	Unidades asociadas	Variantes alofónicas[6]
/a/	A y diptongos "ai","au","ia" y "ua".	[a] , [a,˙] , [a,˛]
/e/	E y diptongos "ei","eu","ie" y "ue".	[e] , [e,˙] , [e,˛]
/i/	I y diptongos "ai".."ui" y "ia".."iu".	[i] , [i,˛] , [j]
/o/	O y diptongos "oi","ou","io" y "uo".	[o] , [o,˙] , [o,˛]
/u/	U y diptongos "au".."ou" y "ua".."uo".	[u] , [u,˛] , [w]
/p/	pa...pu , ap...up	[p]
/b/	ba...bu , ab...ub	[b] , [b,¯]
/t/	ta...tu , at...ut	[t] , [t,˙]
/d/	da...du , ad...ud	[d] , [d,¯]
/k/	ka...ku , aK...uk	[k]
/g/	ga...gu , ag...ug	[g] , [g,–]
/c,^/	CH	[c,^]
/y/	Ya...Yu , aY...uY	[y] , [y,^]
/f/	F	[f]
/q/	Z	[q] , [z,˙] , [z,¸] , [z]
/s/	S	[s] , [s,¸]
/x/	J	[x]
/r/	Ra...Ru , aR...uR	[r] , [ɹ]
/r,¯/	RR	[r,¯]
/l/	La...Lu , aL...uL	[l] , [l,˙] , [l,¸]
/l,√/	LLa...LLu , aLL...uLL	[l,˛]
/m/	Ma...Mu , aM...uM	[m] , [m,^]
/n/	Na...Nu , aN...uN	[n] , [n,˙] , [n,¸] , [n,,]
/n,√/	Ña..Ñu , aÑ..uÑ	[n,˛]

Existen, además, numerosas variantes alofónicas que aparecen con la combinación de dos fonemas al unirse. Este es el caso de los alófonos correspondientes a los diptongos, y otros que no citaremos.

No es el concepto fonológico de fonema el que debe hacerse corresponder con las unidades de síntesis, sino el de alófono, dado que las características físicas del sonido vienen determinadas por éste último y no por el primero. Sin embargo la caracterización del fonema es también interesante dado que permite distinguir, por definición, entre cualesquiera palabras diferentes. Por este motivo, si- como en nuestro caso- no se ha considerado conveniente incluir todas las variedades alofónicas en el banco de muestras, por la extensión que esto supondría, el mínimo subconjunto de éstas, capaz de representar distintamente la totalidad de las locuciones del idioma, es el de los fonemas.

Esta es la aproximación seguida en este caso. Se ha escogido la variedad alofónica más representativa y general para cada fonema (típicamente la que se representa mediante el mismo símbolo que el fonema en la notación común) de modo que sólo sea necesario almacenar las

6 Según Alarcos[50] Véanse allí los detalles sobre los fonemas, sus nombres y características, así como los de sus realizaciones fonéticas o sonidos.

muestras asociadas a éstos. La reducción en cuanto a consumo de almacenamiento es muy considerable pudiendo cifrarse en más de un 50% respecto al requerido por el uso de la totalidad de los alófonos diferenciables en el castellano.[1]

Control de la frecuencia fundamental

Como se ha comentado, la técnica utilizada para conferir entonación a la voz se basa en seleccionar cada unidad de uno entre varios juegos pronunciados por el locutor en diferentes tonos. Esta característica es original y permite evitar el problema derivado de la alteración artificial del tono.

Al modificar la frecuencia fundamental de un sonido como el de la voz, es preciso mantener la estructura de formantes que caracteriza a este sonido, con independencia del tono, al ser emitido naturalmente. Para técnicas paramétricas de almacenamiento, esto, aunque puede conseguirse, dista mucho de resultar sencillo y, aun así, el resultado difícilmente llega a sonar del todo natural. La simplicidad del método usado aquí, sin embargo, nos permite obviar este proceso y obtener algoritmos sencillos y de coste computacional extremadamente reducido que se ejecutan de forma eficiente en sistemas con recursos limitados.

La desventaja evidente de nuestra aproximación reside en la ocupación de memoria. En nuestro prototipo, por ejemplo, el tamaño del almacenamiento de muestras es de 400 Kilobytes. Este tamaño viene marcado por la frecuencia de muestreo utilizada que ha sido de 15 KHz. Mediante el uso de una frecuencia menor y una optimización de las técnicas de almacenamiento, esto puede reducirse a 150 Kbytes. Utilizando técnicas de compresión y/o parametrización de la señal, se calcula que la ocupación total puede ser inferior a los 50 Kbytes. De todas formas, incluso la cifra inicial puede considerarse más que aceptable si tenemos en cuenta la continua reducción de costes de las unidades de almacenamiento (memoria).

Aspectos prosódicos. Síntesis de la curva melódica

Como se ha dicho, la conversión de texto a voz, al contrario de otros tipos de síntesis del habla menos sofisticados, debe llevar a cabo una simulación de los rasgos prosódicos básicos a fin de permitir la inteligibilidad y correcta interpretación de la voz resultante. En castellano, como en la mayor parte de los idiomas de todo el mundo, la entonación juega un papel determinante a la hora de dar sentido completo a un enunciado. Así lo expresa Navarro Tomás en su, ya clásico, "Manual de pronunciación española" [Navarro,16] : " (...) *El conocimiento de la entonación es, pues, de la mayor importancia, tanto para la recta inteligencia de lo que se oye como para la expresión justa de lo que se quiere decir.* (...) "

La forma de conferir la entonación adecuada a las frases se basa, como se ha comentado, en utilizar vocales adquiridas en diferentes tonos para trazar de este modo la curva melódica deseada. El problema real que se presenta estriba en la elección de una curva adecuada para cada caso. Este es el papel que debe cumplir el análisis (léxico, sintáctico y prosódico) del texto.

Durante el desarrollo del sistema se han estudiado en bastante profundidad muy diversas técnicas de análisis del texto como las gramáticas formales y sus extensiones, redes de transiciones recursivas y aumentadas, sistemas de producción, redes semánticas, tramas e incluso redes neuronales. Muchos de estos estudios han revelado interesantes y prometedoras vías de investigación. Las conclusiones extraídas han sido, sin embargo, indicadoras de una excesiva complejidad (tanto conceptual como computacional) en los algoritmos necesarios para

[1] En [Iglesias, 82] se utilizaron 44 alófonos que supondrían prácticamente el doble de espacio que los 24 usados aquí.

llevar a la práctica, en un sistema de bajo coste, la mayoría de estas aproximaciones. Dada la necesidad de algoritmos eficientes y sencillos, se ha desarrollado un método pragmático de análisis basado fundamentalmente en la puntuación ortográfica que, pese a su simplicidad, ha proporcionado resultados más que aceptables. (Para obtener detalles puede consultarse [Pérez,89])

Esquema lógico, lenguaje de codificación y comentario general sobre el prototipo

Describiremos el prototipo construido, capaz de admitir como entrada un texto normal y llevar a cabo su correcta lectura, manteniendo a su vez una adecuada interacción con el usuario.

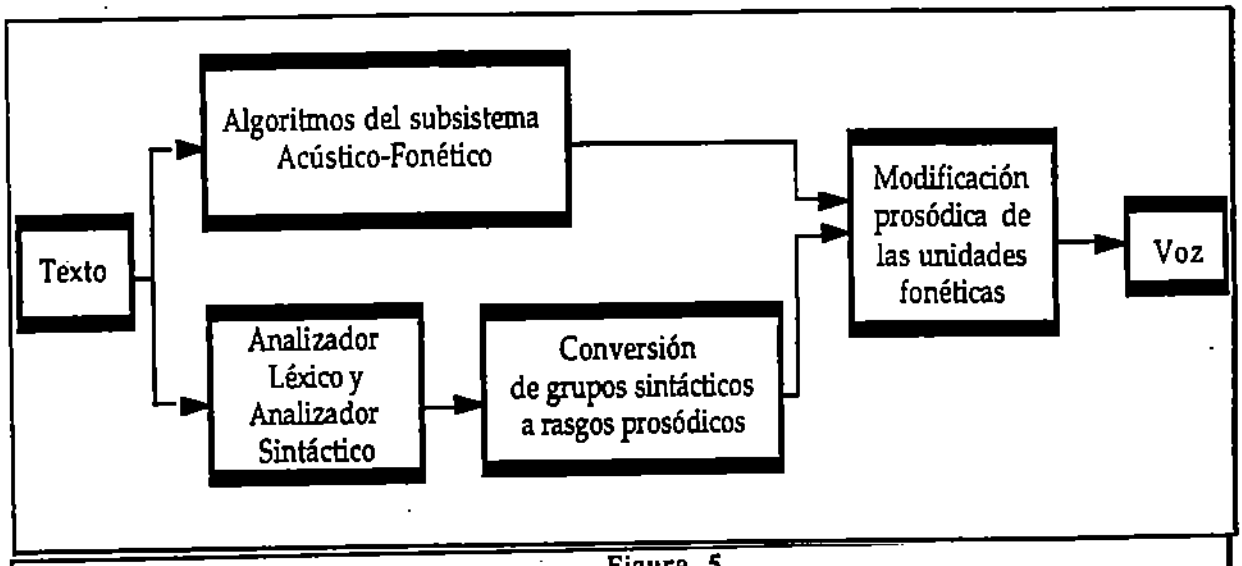


Figura 5.
Flujo de los datos en el sistema propuesto. El texto es procesado tanto por los algoritmos del sistema acústico-fonético como por los analizadores léxico y sintáctico. Con los datos procedentes de ambas ramas se confecciona la cadena de unidades que forman la voz.

En primer lugar, se lleva a cabo un proceso de transcripción de las abreviaturas, los signos especiales y los números. Una vez tenemos el texto en forma de caracteres directamente legibles hay que localizar las palabras que no se deben acentuar prosódicamente, como por ejemplo "de", "hasta", etc. que se almacenan en número de 84 y figuran en una tabla de rápido acceso. Por otra parte, es necesario acentuar prosódicamente los monosílabos que no se inscriban en la categoría anterior. A continuación debemos determinar los grupos fónicos, etiquetar la cadena con la información prosódica, dividir las palabras en sílabas y transcribir estas sílabas en unidades fonéticas. Después hay que asignar las cualidades prosódicas de bajo nivel (acentos prosódicos, fundamentalmente) a las unidades fonéticas. Para ello hay que tener en cuenta los diptongos, sinalefas, etc.

Seguidamente habrá que sumar la curva melódica ya construida (la que ha resultado de la prosodia a nivel más bajo) con la que determinen las etiquetas correspondientes a la prosodia de los grupos fónicos, obtenidas tres etapas antes. De esta forma la curva melódica total interpreta ambos niveles prosódicos. Finalmente se generan las salidas necesarias para que la señal alcance al subsistema acústico que la convierta en voz. Debemos tener en cuenta asimismo los tiempos de pausa entre grupos fónicos dependiendo del carácter de éstos y de la puntuación del texto.

Características funcionales del prototipo

El modelo construido se puede definir como **un sintetizador por codificación de la onda sonora en PCM para conversión de texto a voz mediante reglas.**

Detallamos a continuación las características fundamentales del prototipo.

- El funcionamiento en tiempo real de todo el proceso, incluyendo el tratamiento de números, abreviaturas y **entonación** natural de acuerdo con los signos de puntuación del castellano. Es decir que se trata de un sistema capaz de tomar **un texto cualquiera** y leerlo de viva voz como lo haría un lector humano.

- El control de la prosodia a partir de muestras naturales de voz en frecuencias diversas. Esta particularidad no parece haber sido utilizada en sistemas anteriores a este y va íntimamente ligada a la subdivisión en unidades que hemos comentado. Ambos conceptos se han desarrollado conjuntamente y dan lugar a una filosofía original en el tema de la prosodia.

- Utilización exclusiva de material procedente de muestras de voz humana sin tratar. Esto se puede conseguir, una vez más, gracias a la subdivisión utilizada que suprime la necesidad de suavizaciones u otros tratamientos en la transición entre unidades.

- Un énfasis proyectado especialmente hacia técnicas capaces de analizar el texto fuente y extraer la mayor cantidad posible de información prosódica, en orden a explotar al máximo las buenas posibilidades de control tonal y de duración que ofrece el sistema. El control preciso de la *cantidad vocálica* (duración de las vocales en función de su contexto silábico) y la localización de las palabras inacentuadas mediante consulta en una tabla son dos ejemplos de técnicas de análisis que se hallan incorporadas en el sistema.

- El análisis del texto a nivel ortográfico deletreando los numerales y sustituyendo las abreviaturas más comunes por sus equivalencias. Se lleva a cabo una pronunciación castellana aproximada de los términos foráneos tal y como recomienda la Real Academia de la Lengua. Las siglas no recogidas en la tabla de abreviaturas se deletrean para su fácil identificación por el oyente. Se controlan todos los signos de puntuación, incluso los menos comunes como el guión y los puntos suspensivos. Este completo tratamiento ortográfico no se suele encontrar en sistemas de este tamaño.

- La independencia del método respecto al sistema físico utilizado, siendo muy fácil de adaptar a uno u otro computador y requiriendo de éste una potencia muy pequeña.

Ejemplos de comportamiento.

Para dar una idea adecuada de las múltiples dificultades a las que se enfrenta un sistema de conversión de texto a voz, vamos a citar algunos ejemplos de frases conflictivas.

Frase de entrada:

Ha salido de entre los arbustos.

Este es un caso típico de abundancia de palabras inacentuadas. No se debe pronunciar **"Há-salído-dé-éntre-lós-arbústos"** sino **"Há-salído-de-entre-los-arbústos"**. Para ello el sistema comprueba que las palabras "de", "entre" y "los" se encuentran en la tabla de palabras inacentuadas y "ha" no está en esa tabla.

Frase de entrada:

Lavar, planchar, coser, etc. son labores ingratas.

Vemos aquí la aparición de un punto que puede llevar al sistema a decidir el final de la frase. Esto daría lugar a una entonación que la haría difícilmente comprensible. En nuestro caso, sin embargo, se detecta la aparición de la abreviatura "etc." en la correspondiente tabla. Se traduce, lógicamente, a "etcétera". Adicionalmente, El análisis sintáctico detecta una posible

enumeración (no por la aparición de "etc." sino por la estructura de la frase) y asigna los tonemas (modelos de entonación) correspondientes.

Frase de entrada:

¡ Pero bueno ! ¿ Qué pasa aquí ?

Los signos de admiración e interrogación exigen una pronunciación específica. El sistema lo reconoce y da lugar a la entonación adecuada. En cuanto al tipo de interrogación, se asigna el tonema correspondiente a interrogación absoluta a toda aquella que contenga un pronombre interrogativo. Serán, por tanto, interrogaciones relativas el resto.

Conclusión

Se ha descrito un sistema de conversión texto-voz que presenta cierto número de aspectos innovadores y un prototipo de aplicación de este sistema. El prototipo funciona en tiempo real sobre un sistema físico de bajo coste y la calidad de voz obtenida es muy satisfactoria.

Los usos y posibles aplicaciones de un sistema como el presentado aquí son numerosos. Además de los obvios pero inestimables servicios que, en combinación con un OCR (reconocedor óptico de caracteres), puede prestar al usuario invidente, la conversión de texto a voz puede ser una ayuda importante en tareas de corrección de textos y otras.

En concreto, un sistema como el que se presenta, capaz de permitir la identificación de errores en la pronunciación de cualquier palabra, se puede usar con gran conveniencia en la corrección inteligente- por parte del mismo escritor, secretaria u operario- de fallos en la acentuación ortográfica, defectos de puntuación y otros errores de introducción de texto. La diferencia claramente apreciable entre las frases "**No se como se hace**" y "**No sé cómo se hace**" (la primera tiene dos faltas de acentuación) nos llama la atención inmediatamente al escucharlas. Errores tipográficos como "**acalchofa**" o "**ampilficador**", que son difícilmente detectables en la lectura, no se escapan cuando los oímos pronunciados.

La automatización de ciertos cometidos relacionados con la información telefónica (en empresas de consultoría, seguridad, publicidad, etc.) puede suponer una importante mejora de algunos servicios. La comodidad de introducir una serie de textos escritos en lengua natural y que puedan ser pronunciados en voz alta en diferentes circunstancias puede aumentar la productividad y agilidad de empresas y profesionales.

También en la comprobación de datos introducidos en un sistema informático puede jugar un papel importante la síntesis de voz. Si hemos introducido, por ejemplo, gran número nombres con sus direcciones y números de teléfono y queremos comprobar la exactitud de todo ello (en una aplicación importante como nóminas, fichaje policial, etc), la tarea puede ser ardua y poco fiable sin la ayuda de alguien que lea los datos introducidos mientras otra persona los comprueba en la lista impresa original. El uso del sistema permite la lectura automática facilitando la rápida comprobación por un sólo operario. Elimina incluso la posibilidad de error en la lectura oral, siendo la fiabilidad, en muchos contextos, cercana al cien por cien tras un breve período de adaptación a la voz sintética.

Finalmente, junto con un sistema de reconocimiento de voz multilocutor, puede constituir un conveniente canal de comunicación hombre-máquina accesible a usuarios con o sin formación técnica, directamente o a través de la línea telefónica.

Referencias

Alarcos, T. (1950) "Fonología española", Biblioteca románica hispánica, Gredos, Madrid.

Allen, J. (1976) "Synthesis of Speech from Unrestricted Text", Proceedings of the IEEE, Vol.64, No.4, April 1976.

El-Imam Y.A. (1987) "Speech Synthesis by Concatenating Sub-Sylabic Sound Units", Proc. of IEEE Intern. Conf. ICASSP, 1987, pp.2416-2417.

Fernández, L.C. ; Martinez, J.A. ; Alvarez, J. (1987) "Sistemas digitales de almacenamiento y síntesis de señal vocal en telefonía". Mundo Electrónico núm 171. pp.91-96.

Fu, K.S. (1982) "Syntactic Pattern Recognition and Applications", Prentice Hall.

Grosjean, F. ; Gee, J.P. (1987) "Prosodic Structure and Spoken Word Recognition", Cognition, 25. pp.135-155.

Hamon, C. ; Moulines, E. ; Charpentier, F. (1989) "A Diphone Synthesis System based on Time-Domain Prosodic Modifications of Speech", Proc. of IEEE Intern. Conf. ICASSP 89. pp.238-241.

Hunnicut, M.S. (1987) "La Síntesis de Voz como Ayuda Técnica". Mundo Electrónico núm 170. pp.63-68.

Iglesias E. ; Meneses J. ; Muñoz E. ; Quilis A. (1982) "Sistema de conversión de texto a voz en español a partir de demisílabas" Quinto Congreso de Informática y Automática, Mayo 82. pp.805-809.

J.Martí en Vidal, E. ; Casacuberta, F. (1987) "Reconocimiento automático del habla", Apéndice "Sintesis del habla", Marcombo.

Lea, W.A. ; Medress, M.F. ; Skinner T-E. (1975) "A Prosodically Guided Speech Understanding Strategy", IEEE Transactions on Acoustics, Speech and Signal Processing. Vol ASSP-23, Nº 1, Feb.85. pp.30-38.

Lippmann, P. (1987) "An Introduction to Computing with Neural Nets", IEEE ASSP Magazine, April 87, pp.4-22.

Ljolje, A. ; Fallside, F. (1986) "Synthesis of Natural Sounding Pitch Contours in Isolated Utterances Using Hidden Markov Models", IEEE Transactions on Acoustics, Speech and Signal Processing. Vol ASSP-34, Nº 5, Oct.86. pp.1074-1080.

Llisterri, J. "La síntesis del habla: Estado de la cuestión" Boletín de la Sociedad Española para el procesamiento del Lenguaje Natural, núm 6. pp.19-41

Mannell, R. ; Clark, J.E. (1987) "Text-to-Speech Rule and Dictionary Development", Speech Communication 6. pp.317-324.

Mañas J.A. (1987) "Word division in spanish", Comm. of the ACM, July 1987, pp.612-616.

Nakamura, M. ; Shikano, K. (1989) "A Study of English Word Category Prediction Based on Neural Networks", Proc. of IEEE Intern. Conf. ICASSP 89.

Navarro, T. (1916) "Manual de pronunciación española", Madrid, CSIC Decimoctava edición, 1974.

Nilsson, N. (1987) "Principios de inteligencia artificial", Springer-Verlag.

O'Shaughnessy, D. (1984) "Design of a Real-Time French Text-to-Speech System", Speech Communication 3. pp.233-243.

O'Shaughnessy, D. (1987) "Specifying Intonation in a Text-to-Speech System using only a Small Dictionary", IEEE Transactions on ASSP 1987, pp.1430-1436.

Pérez, J.C. (1989) "Sistema de Conversión de Texto a Voz para Castellano" Proyecto Fin de Carrera. Facultad de Informática, Universidad Politécnica de Valencia.

Pierrehumbert, J. (1981) "Synthesizing intonation", J. Acoust. Soc. Am. 70(4), Oct. 81, pp.985-995.

Sandri, S. ; Vivalda, E. (1981) "A Formal Language for the Generation of Prosodic Rules", The Fourth F.A.S.E. Symposium, April 81, Venezia. pp.165-168.

Santos, A. ; Muñoz, P. ; Martinez, M. (1984) "Diseño y evaluación de reglas de duración en la conversión de texto a voz" Boletín de la Sociedad Española para el procesamiento del Lenguaje Natural, núm 6. pp.71-91.

Scordilis, M.S. ; Gowdy, J.N. (1989) "Neural Network based Generation of Fundamental Frequency Contours", Proc. of IEEE Intern. Conf. ICASSP 89. pp.219-222.

Torres B. ; Vidal E. (1982) "Sistema de síntesis automática del castellano hablado", Quinto congreso de informática y automática. Madrid, pp.405-409.

Umeda, N. (1976) "Linguistic Rules for Text to Speech Synthesis", Proc. of the IEEE, Vol. 64 Nº 4, April 76. pp.443-451.

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.

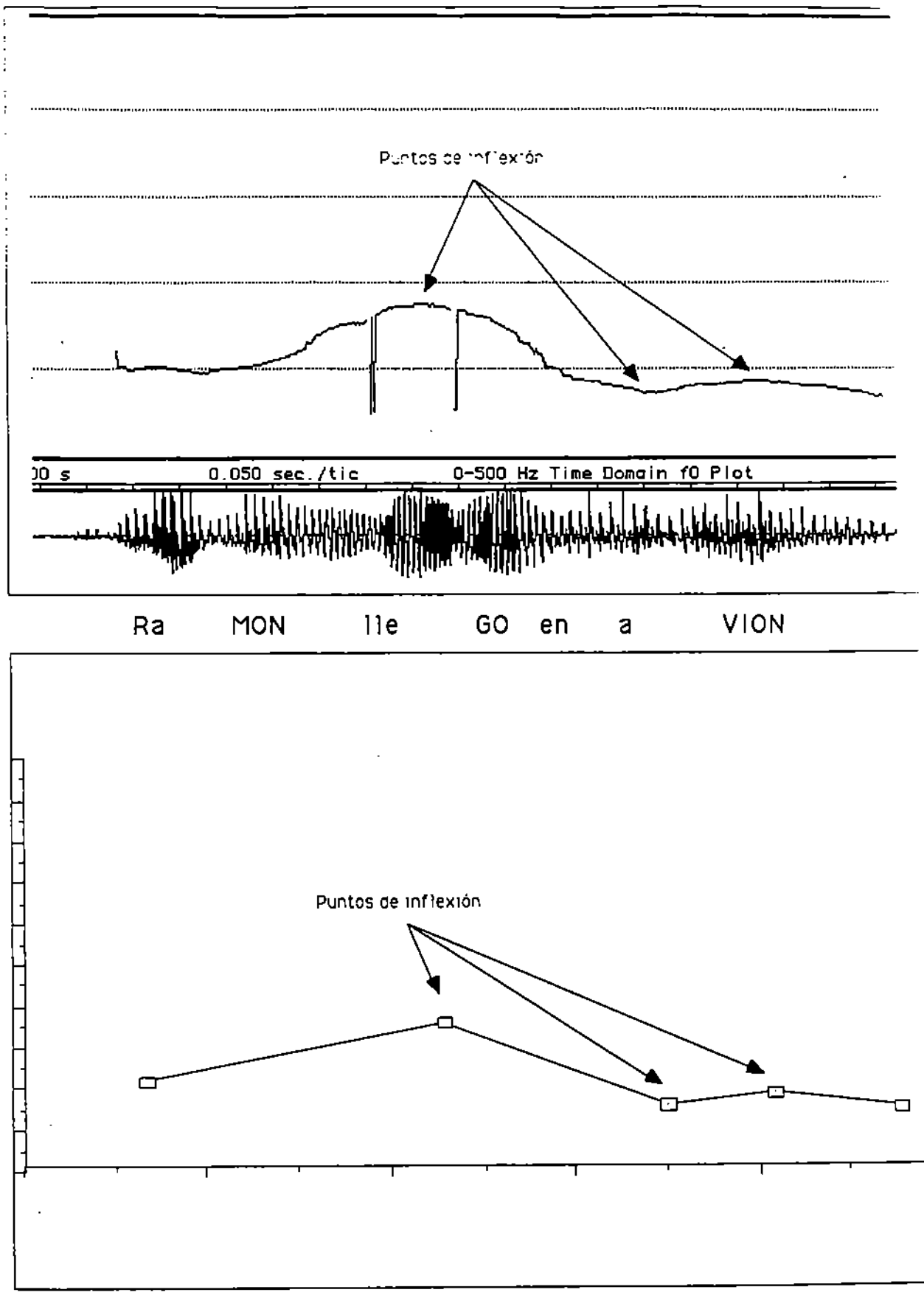


Figura 1: ejemplo de representación estilizada con su correspondiente curva original

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.

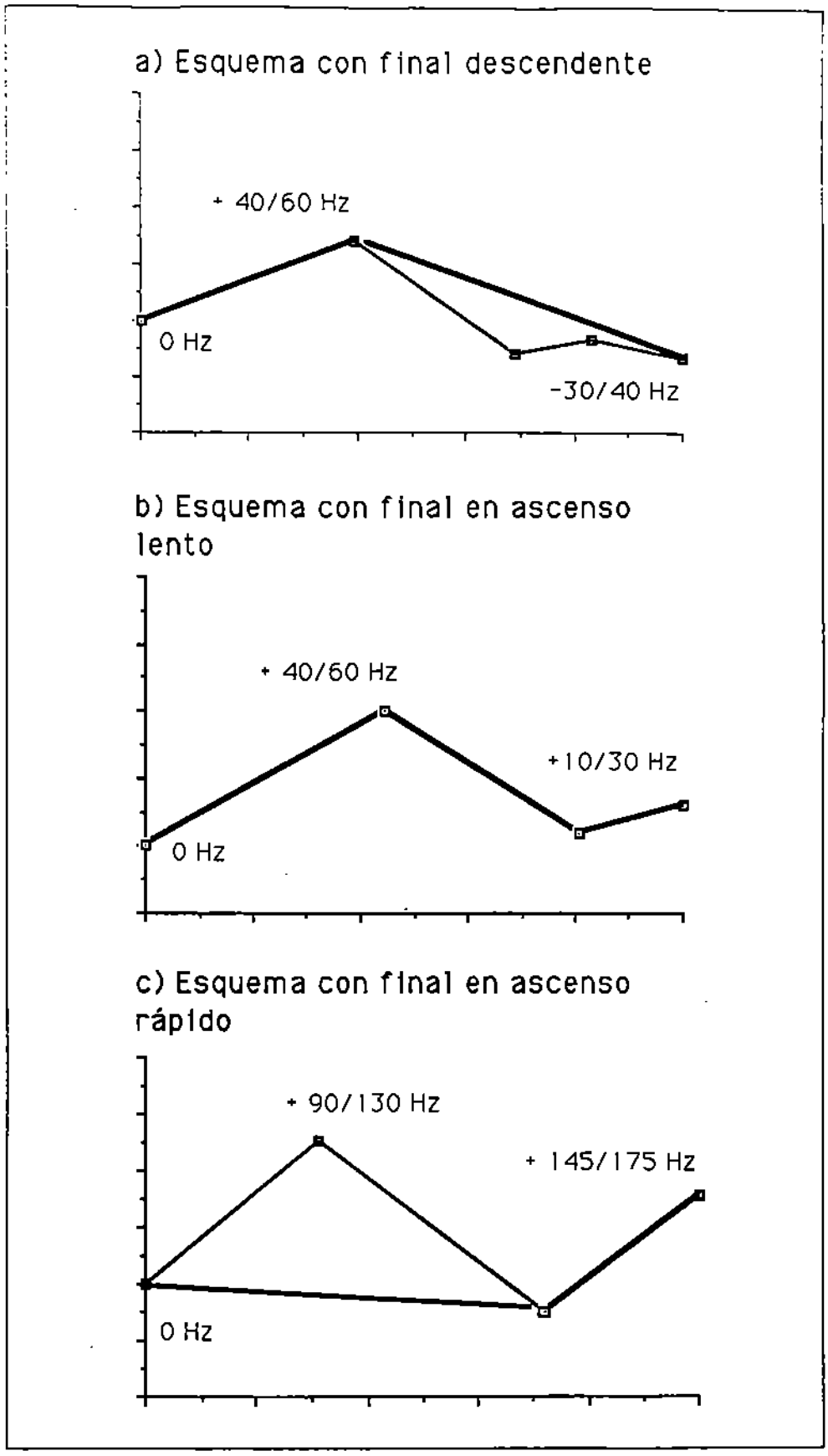


Figura 2: patrones melódicos básicos

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:

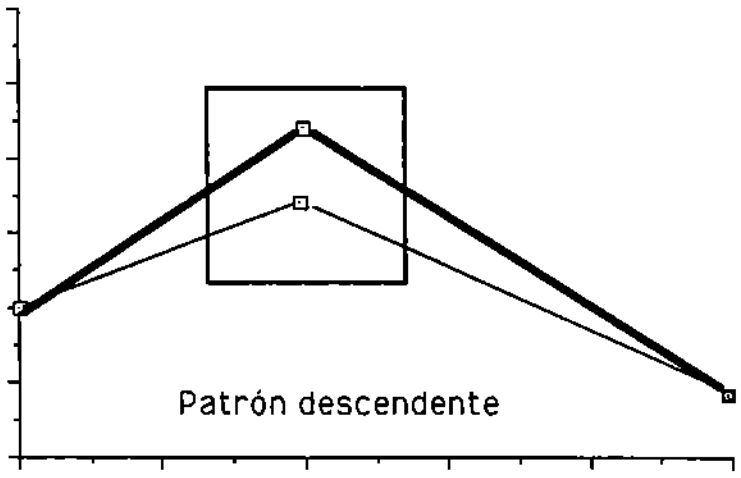
a) un análisis de la importancia de la información de niveles superiores al fonético (léxica y semántica, especialmente) en la identificación de los distintos tipos de frases: así, por ejemplo, habría que comprobar hasta qué punto influye la presencia de una partícula interrogativa en la identificación 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*.

1) Elevación de la altura tonal del primer pico



2) Esquema circunflejo

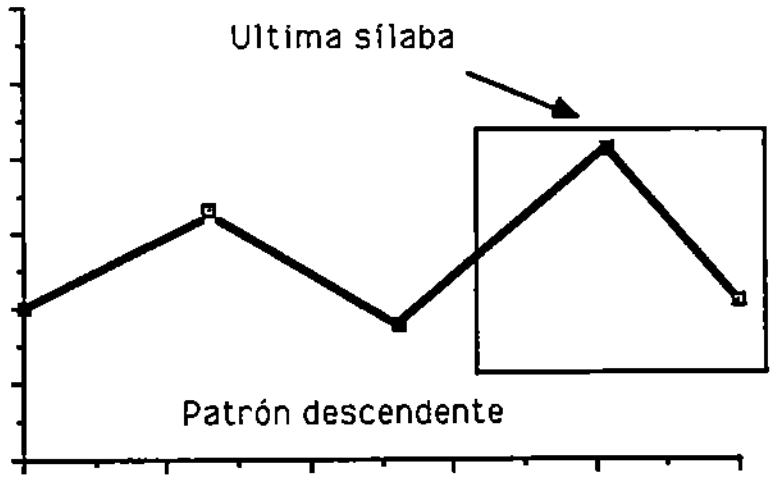
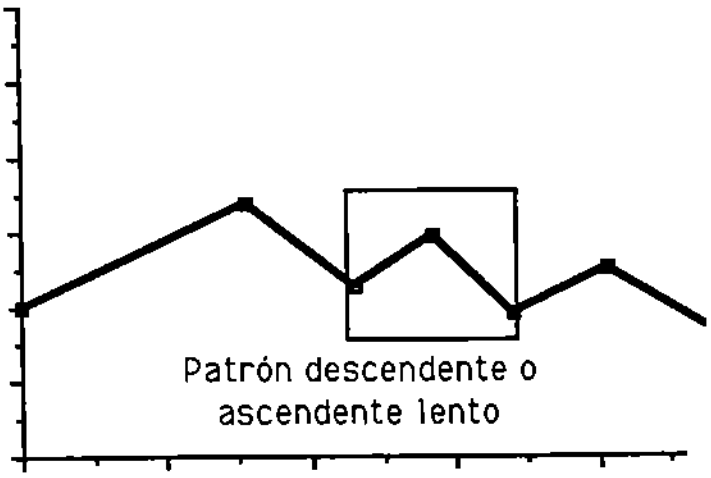


Figura 3a: esquemas superpuestos

3) Aumento del número de picos



4) Aumento del rango frecuencial

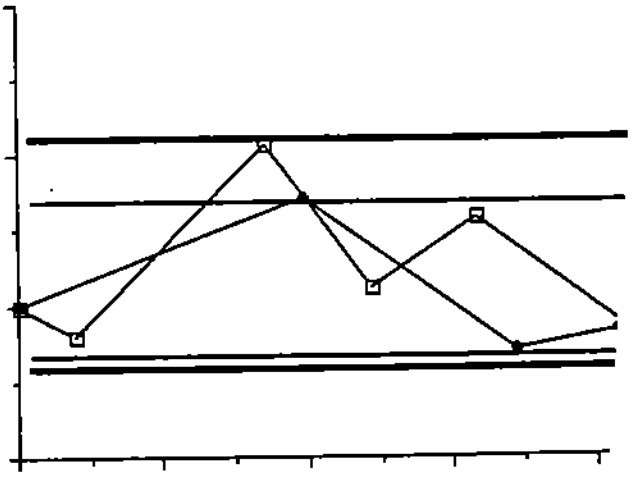


Figura 3b: esquemas superpuestos

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.