Proposición de un marco adecuado para el estudio de contornos de F0 para síntesis de voz

Xavier Fernández Salgado y **Eduardo Rodríguez Banga**
Departamento de Tecnoloxías das Comunicacións
E.T.S.E. Telecomunicación, Universidade de Vigo
Campus Universitario. 36200 - Vigo
{xsalgado,erbanga}@tsc.uvigo.es
Telf: 986 812141 – 986 812676
Fax: 34 986 812116

Resumen La entonación humana, concretada en la curva de frecuencia fundamental (f0), es un complejo proceso en el que pueden observarse una serie de fenómenos (como la declinación o el acento de entonación) cuyo alcance se extiende sobre unidades de distinta longitud. Para un estudio adecuado de estos fenómenos en esta contribución se propone un modelo de descripción jerárquica que discretiza el contorno de f0 en una serie de puntos que se asignan a los distintos niveles implicados (nivel de grupo fónico, de grupo tónico y segmental). Esta descripción posibilita el alineamiento de los contornos, de forma que se pueda realizar fácilmente un estudio comparativo y se permita observar comportamientos y tendencias comunes susceptibles de ser aplicadas a un sintetizador de voz. Aunque la metodología aquí propuesta estaba inicialmente dirigida al estudio de la lengua gallega, resulta también adecuada para otras lenguas afines.

1 *Introducción*

Aunque la inteligibilidad de la cadena oral en los sistemas de síntesis de voz parece estar plenamente garantizada, el conseguir una mayor naturalidad es aún una tarea pendiente que involucra tanto la mejora de los métodos de generación de voz como la asignación de una adecuada prosodia.

En esta comunicación se realiza una aproximación al análisis de contornos de f0 que es la base de un nuevo método automático de generación de contornos para síntesis de voz [Salgado & Banga 00].

2 *Descripción del corpus*

Para la realización de este trabajo se dispone de un corpus prosódico que consta de 1000 frases en lengua gallega, y abarca las modalidades oracionales básicas: enunciativas, enunciativas con interpolación de otra sentencia, enumerativas, interrogativas totales y parciales, exclamativas e imperativas. La f0 ha sido extraída utilizando el estimador de f0 propuesto por [Tuerk]. Las frases del corpus han sido analizadas lingüísticamente de manera automática por la etapa de procesado de lenguaje natural de nuestro sintetizador de voz con el fin de asignar a cada alófono distintos atributos (posición en el grupo fónico, posición relativa dentro del grupo tónico, tonicidad, etc.). Como resultado de esto proceso se genera una base de datos que posibilita el tratamiento por medios automáticos de toda esta información

3 *Análisis jerárquico*

La entonación es un proceso cuya descripción pasa necesariamente por el estudio de fenómenos que se manifiestan a distintos niveles (segmental, frase, párrafo, etc.)[O'Shaughnessy 79]. En este trabajo consideramos en orden jerárquico descendente los niveles de grupo fónico, grupo tónico y segmental. En la literatura existen otras técnicas de análisis y modelado de la entonación que, implícita o explícitamente, utilizan una descripción jerárquica. Por ejemplo, el conocido modelo de Fujisaki [Fujisaki 97] separa el contorno de frecuencia fundamental en dos componentes denominadas "de frase" y de "acento". No obstante, el principal inconveniente de este modelo radica en la dificultad de asociar los movimientos del contorno de frecuencia fundamental con la información lingüística. Nuestro objetivo es, por tanto, el proporcionar una metodología que permita el descubrimiento de características prácticamente invariantes a nivel de grupo fónico, tónico y segmental, y que puedan ser fácilmente asociadas con la información lingüística disponible.

4 *Nivel de grupo fónico*

La representación de contornos a partir de una serie de puntos objetivo (*target points*) no es nueva. En [Hirst & Espesser 93] o en [Garrido 96], el contorno de frecuencia fundamental es discretizado en una serie de puntos H (altos, *High*), y L (bajos, *Low*). En [Garrido 96] a partir de estos puntos y mediante regresión lineal se generan *toplines* y *bottonlines*.

En nuestra descripción el grupo fónico quedará esquematizado mediante un punto por cada grupo tónico. Por medio de una alineación de los eventos de acentuación, los contornos son asentados sobre una misma base de tiempos normalizada, permitiéndose su comparación directa y evitándose de esta forma los problemas de visualización que ocasionaría una simple alineación temporal.

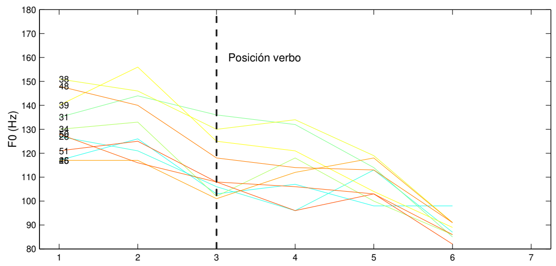
Como valores representativos para cada grupo tónico utilizaremos los valores del núcleo de la sílaba pretónica o los de la sílaba tónica. Estos puntos tienden a ser los valores más próximos y representativos de los puntos L y H en base al tipo de discretización utilizada a nivel de grupo tónico, que como veremos será de un punto por sílaba (sección 5). Esta aproximación se basa en la observación alineada de grupos tónicos coincidentes en la posición del acento (agudo, grave y esdrújulo en figura 2), en los cuales se puede observar que el punto más bajo era el de la vocal pretónica. Similarmente, al final de la vocal tónica, los acentos de entonación han alcanzado generalmente la altura máxima de f0 (tablas 1 y 2). Aunque varios autores señalan un desplazamiento del pico de f0 hacia la derecha de la vocal tónica [Llisterri et al 95], comportamiento corroborado por nuestras observaciones, la realidad es que es muy poca la subida posterior a la vocal tónica.

La unión de los puntos pertenecientes a la vocal tónica de cada grupo tónico permite generar una línea poligonal (supralínea), la cual será empleada para el estudio comparativo de estos contornos. De igual forma, uniendo los puntos de las vocales pretónicas obtenemos la línea inferior. No creemos oportuno generar líneas rectas imaginarias (creadas por ejemplo mediante regresión lineal), ya que estos contornos están lejos de poder aproximarse por una línea recta. La utilización de los puntos L para crear infralíneas no está exenta de ciertos problemas. Hemos podido observar como habitualmente dos acentos léxicos de palabras diferentes muy próximas no permiten una depresión de la f0, y por tanto no existe realmente un punto L. Este fenómeno también ha sido observado por [Prieto & Shih 95]. Por esta y otras razones (en posición inicial de frase no siempre es posible estimar un valor para la sílaba pretónica), creemos más oportuno utilizar las supralíneas en nuestro estudio, ya que tienden a ser más representativas, estables, y relacionables con una determinada estructura sintáctica.

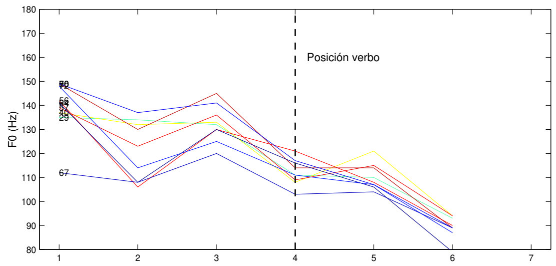
4.1 Aplicación a oraciones enunciativas

Siguiendo el marco descrito en la sección anterior, hemos realizado varios experimentos para distintas longitudes de enunciación, medida en número de grupos tónicos. La estructura sintáctica de las sentencias empleadas ha sido Su+Compl+Verbo+Compl, donde los complementos (uno o varios) son optativos. Es importante resaltar que estas frases no guardan ningún parecido semántico, ni el material fonético (en cuanto a sonoridad de los alófonos) es el mismo. La superposición de las correspondientes supralíneas (por ejemplo para 6 grupos acentuales) muestra distintos comportamientos en cuanto a la trayectoria global. Como hipótesis para explicar estos comportamientos se ha considerado la influencia de unidades menores que la sentencia (sujeto/predicado) en oraciones monoclausales. Además el punto de unión entre sujeto y predicado es tomado, a veces, como un punto para realizar el "phrasing" adecuado, como se indica en [Bachenko & Fitzpatrick 90]. Clasificando los contornos en función de la posición del verbo, obtenemos los resultados que pueden verse en la figura 1.

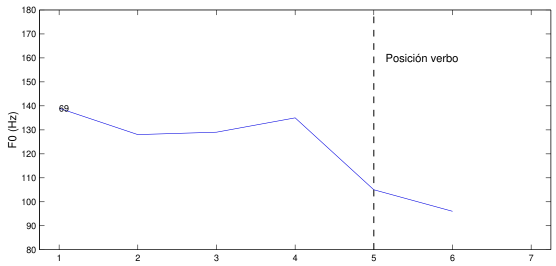
Se puede observar como la estructura sintáctica parece afectar la forma del contorno. Varios son los fenómenos comprobables. El primero es la dependencia del contorno de la posición de la frontera sujeto-predicado en cuanto a que el grupo tónico anterior al verbo marca el inicio del progresivo descenso de la f0 (fácilmente observable en la figura 1(b)). El segundo es la notable alternancia entre puntos altos y bajos a modo de "diente de sierra", lo que puede ser un indicativo de distintos grados de tonicidad. Otro efecto a señalar es el descenso de la frecuencia fundamental entre la penúltima y la última sílaba



(a) Frontera entre el 2^o y 3^{er} g. t.



(b) Frontera entre el 3^{er} y 4^o g. t.



(c) Frontera entre el 4^o y 5^o g. t.

Figura 1: Supralínea para contornos de 6 grupos tónicos, en función de la posición de la frontera sujeto-predicado

tónicas (*final lowering*) .

Concluimos, por tanto, que la la descripción en términos de la supralínea parece adecuada para manifestar fenómenos que ocurren a nivel de grupo fónico.

5 *Nivel de grupo tónico*

El grupo tónico (*syntactic stress group* según [Thorsen 83]) es la unidad intermedia escogida en nuestra descripción jerárquica . Considerando el paralelismo entre palabras función y átonas (y palabras léxicas y tónicas), el grupo tónico coincide en la mayoría de ocasiones con la unidad denominada sintagma. Ésta es una de las razones para escoger la unidad de grupo tónico en lugar de otras unidades relacionables también con el acento como puede ser el grupo acentual. En [Marín et al 96] se justifica la utilización de esta unidad a nivel prosódico y sintáctico.

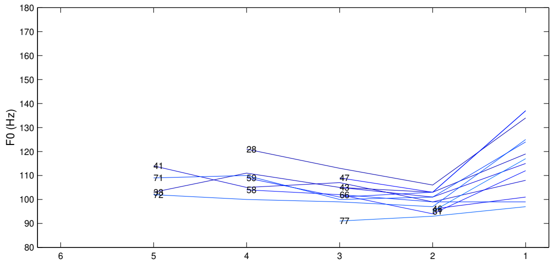
El contorno de grupo tónico es esquematizado mediante un punto en el núcleo silábico de cada sílaba, siendo, por tanto, un método homogéneo en cuanto a que realiza la discretización empleando siempre el entorno sonoro de cada sílaba. De esta manera se intentan visualizar los movimientos asociados a la sílaba acentuada, o los movimientos implicados en la realización de tonos finales o frontera. Clasificaremos los grupos tónicos de acuerdo a su posición en el grupo fónico como intermedios y finales.

5.1 Grupos tónicos intermedios en grupos fónicos enunciativos

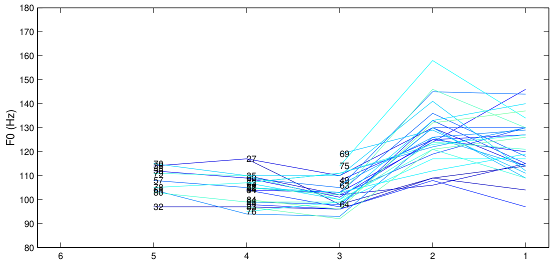
Utilizando la comparación visual de contornos alineados de acuerdo a la discretización indicada anteriormente, considerando un mismo tipo de acento (agudo, grave o esdrújulo), y una misma posición dentro del grupo fónico, los resultados son los que pueden observarse en la figura 2.

Se observa que, claramente, el movimiento de subida se localiza entre la vocal de la sílaba pretónica y la tónica, alcanzando aquí el máximo o la mayor parte de la altura final del movimiento. Este hecho justifica la aproximación empleada en la generación de supralíneas en la sección 4 donde se asignaba el punto H a la vocal tónica. En nuestro marco de descripción no es posible observar el tipo de movimiento después de la sílaba tónica en el caso de grupos tónicos agudos, pero sí en graves y esdrújulos. En el caso de graves el contorno puede mantenerse o descender. Un cuidadoso examen manual de cada uno de los contornos para tratar de relacionar estos comportamientos de la f0 con factores tales como mantenimiento de la frecuencia fundamental debido a la continuidad de sonoridad en la cadena de alófonos inmediata, no reveló una posible relación en cuanto a distribución de casos. En el caso de esdrújulos, al menos para los casos observados, es mayoritario el mantenimiento o incluso subida de la f0 en la sílaba posterior a la tónica.

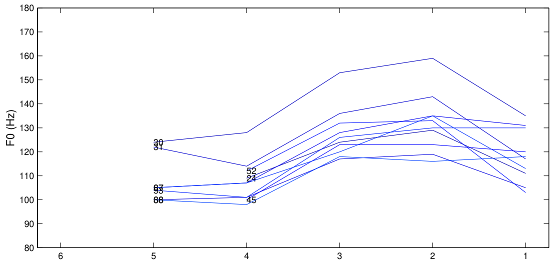
Se debe tener en cuenta que la excursión de f0 es mayor en los grupos tónicos iniciales y que se va reduciendo paulatinamente hacia el final, hasta no ser posible la observación de



(a) Agudos



(b) Graves



(c) Esdrújulos

Figura 2: Grupos tónicos (posición 3^a) de frases enunciativas largas

este movimiento en los grupos tónicos finales enunciativos.

5.2 Grupos tónicos finales

De todos las modalidades básicas de sentencias recogidas en nuestro corpus, sólo presentaremos 2 casos significativos de grupos tónicos finales, que son el final de las interrogativas totales (*yes-no questions*) y el de enunciativas simples. Si para los grupos acentuales intermedios era pertinente la consideración del tipo de acento, las pruebas hechas para grupos acentuales finales no muestran dependencia de este factor. En éstos no se observa ningún movimiento especial, evolucionando el contorno de f0 de forma más o menos plana hacia el valor final de f0. Este contorno puede ser observado en la figura 3. Recordaremos que no siempre es posible estimar un valor de f0 para la última vocal de grupo fónico. En estos casos se considera que tiene el mismo valor que la anterior.

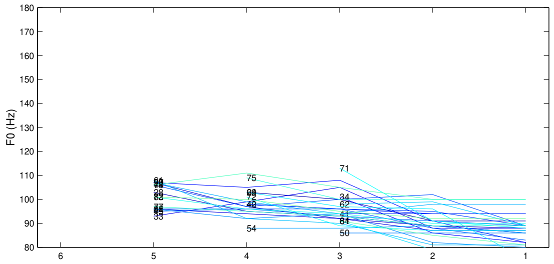


Figura 3: Grupos tónicos finales graves de enunciativas

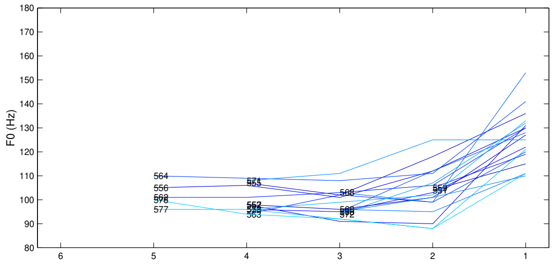


Figura 4: Grupo tónico final de interrogativas totales graves

Para el caso de interrogativas totales se consideró, en principio, la posibilidad de la dependencia de la longitud en número de grupos tónicos, ya que para cadenas fónicas suficientemente grandes y debido al propio fenómeno natural de declinación, era posible que no se realizase el movimiento de anticadencia señalado por [Quilis 64]. Sin embargo las gráficas realizadas para grupos tónicos finales de grupos fónicos cortos (4 grupos tónicos o menos) y largos (más de 4 grupos tónicos) resultaron semejantes. Los resultados para el caso de largos se muestran en la figura 4. Se comprueba como el movimiento de subida más importante se sitúa entre los núcleos silábicos penúltimo y último, aunque a veces ocurre con anterioridad una ligera subida.

6 Nivel segmental

El nivel más bajo de esta descripción es el segmental, el cual será descrito mediante un punto por cada alófono. Esta etapa supone un refinamiento sucesivo al análisis de grupos tónicos en el entorno próximo de los eventos que más influyen en la curva de entonación, que son el acento léxico y las fronteras. Aunque el movimiento de subida en torno al acento era observable con claridad en el nivel de grupo tónico, esta descripción resulta muy poco detallada, ya que no localiza si el acento se realiza antes de la sílaba tónica, al inicio de ésta, o justo en la vocal tónica. A continuación se analiza qué ocurre en el entorno próximo de la vocal tónica en función del tipo de alófonos que la rodean. Este estudio estará limitado a acentos iniciales de frases enunciativas que constan de una sola cláusula. Al igual que en las jerarquías superiores antes expuestas, el camino de análisis de la entonación se basa en el estudio del contorno de f0. No obstante existen otras aproximaciones válidas, a través, por ejemplo, del estudio perceptual de esta realización, sometiendo a distintos sujetos a estímulos (sintéticos o naturales) con distintos patrones de curvas [Rietveld & Gussenhoven 95].

6.1 Descripción del movimiento de f0 asociado al acento

Aunque gran parte de la literatura existente está dedicada a la situación de los máximos de f0 (por ejemplo [Llisterri et al 95] [Prieto & Shih 95], para el castellano), menos abundante es una descripción más precisa de la trayectoria de subida de f0.

Para este estudio se utiliza un valor de f0 por alófono (estimado como valor central del alófono o valor promedio), considerando este punto como perteneciente a una curva continua de f0. Es decir, también se asignará un valor de frecuencia fundamental en aquellos casos en los que no ha sido posible estimarla o en tramos sordos. Al mismo tiempo se calcula para cada alófono la diferencia de f0 respecto al valor anterior, denominando cf0 a este parámetro, y el incremento de f0 dentro de cada alófono (f0dif). Como resultado, todos los alófonos tienen asignado un valor cf0, mientras que f0dif sólo se atribuye a alófonos sonoros.

6.2 Desarrollo del acento en sílaba sonora

En este caso se exige que la vocal tónica esté rodeada de alófonos sonoros, siendo el inicio silábico una consonante. Se consideran 2 posibilidades a priori, dependiendo de que el alófono posterior a la vocal tónica pertenezca o no a la misma sílaba.

		Inicio sil.	Núcleo	Sig. sil.
Posición	1	11.15	7.08	3.75
g.t.	2	9.93	6.21	2.27

Tabla 1: Valor medio del incremento de f0 en hz. dentro de cada alófono (f0dif) en sílaba sonora de 2 alófonos seguida de alófono sonoro (frases 24-84, 1^o y 2 ^{o} g.t.)

		Ini. sil.	Núcleo	Núc./coda
Posición	1	11.67	7.75	4.08
g.t.	2	9.8	4.3	0.4

Tabla 2: Valor medio del incremento de f0 en hz. dentro de cada alófono (f0dif) en sílaba sonora de 3 ó más alófonos sonoros (frases 24-84, 1^o y 2^o g.t.)

Comparando los resultados obtenidos en ambos casos (tablas 1 y 2), podemos ver que los incrementos son similares independientemente de que el alófono posterior pertenezca a la misma sílaba o no. Además puede observarse como el mayor incremento en el aumento de f0 se produce en la consonante sonora, siendo notable, aunque menos importante en la vocal tónica, y afectando de una forma residual al siguiente alófono. Esto parece confirmar el desplazamiento del pico de f0 hacia posteriores alófonos, aunque la mayor parte de la subida se produce en el inicio silábico. Estos datos concuerdan con [Arvaniti et al 88], donde se cita que para el griego actual los acentos prenucleares empiezan con una fuerte subida en el comienzo silábico de la sílaba acentuada, pudiendo alcanzar el máximo en la siguiente. Finalmente recordamos que los valores que aparecen en las tablas son valores medios. Sin embargo existen diferentes grados de acentuación o prominencia, que se reflejan directamente en las excursiones de f0. Además pueden estar influyendo fenómenos microprosódicos, como la f0 intrínseca de las vocales [Whalen & Levitt 95].

6.3 Realización del acento en sílaba con alófonos sordos rodeando la vocal

El entorno estudiado es el de consonante sorda en el comienzo silábico más otra consonante sorda inmediatamente después de la vocal

	F0dif	$Cf0_1$	$Cf0_2$	$Cf0_1 + Cf0_2$
2^o g.a.	-15	7.2	18.5	25.7
3^o g.a.	-25	8.7	20.8	29.5

Tabla 3: Parámetros asociados a la realización del acento en entorno sordo

tónica. Existe por tanto una discontinuidad en la curva de f0. En este entorno el estudio será realizado a partir de los valores de cf0, parámetro que como se ha dicho anteriormente mide el incremento respecto al valor anterior de f0, y que es estimado incluso para alófonos sordos a través de valores interpolados. También se utilizará el valor de f0dif de la vocal tónica, es decir, el incremento entre el valor inicial y el final de ésta. Los datos numéricos han sido extraídos en los grupos acentuales 2º y 3º. $Cf0_1$ se refiere al incremento entre la primera consonante sorda y el alófono anterior a ésta , y $cf0_2$ se refiere al incremento entre la vocal tónica y la sorda anterior.

Llama la atención el valor negativo de f0dif. Esto indica que en el tramo sonoro de la vocal tónica la f0 en vez de subir, baja, lo que parece confirmar el comportamiento señalado por [Allen et al 87] para el inglés, en el que la f0 en vocal precedida de comienzo silábico sordo tiende a decrecer. Este comportamiento se debe a efectos microprosódicos debido a la interrupción de la sonoridad.

También es notable el valor positivo de $cf0_1$ más $cf0_2$, el cual refleja el incremento total de f0 entre el alófono anterior a la consonante sorda y la vocal tónica, es decir, el incremento relacionado con el acento. Por tanto, en entorno sordo el acento de entonación no se realiza mediante un aumento de f0 en el tramo sonoro, sino que se produce un salto de f0 en el periodo de duración de la consonante sorda.

7 *Aplicación a la generación de prosodia en un sintetizador de voz*

La descripción presentada en las secciones anteriores es la base de un modelo jerárquico para la generación de f0 en nuestro sintetizador de voz. El contorno de f0 es generado en tres etapas sucesivas en las que se escoge el representante adecuado (en una base de datos de contornos naturales de f0) para los niveles de grupo fónico, tónico y alofónico. A nivel de grupo fónico el contorno (semejante a los que aparecen en la figura 1) será seleccionado en función de la estructura sintáctica y modalidad oracional. Éste constituirá el esqueleto del contorno de f0 final. Cada uno de los puntos asignados (valor de f0 en vocal tónica) asignados en este nivel constituyen los puntos de anclaje sobre el que "colgaran" los patrones a nivel de grupo tónico. Para ello el nivel medio de f0 de estos será aumentado (o disminuido) de manera que el valor de f0 de la vocal tónica del patrón de grupo tónico coincida con el correspondiente del grupo fónico seleccionado en la etapa anterior. La selección del grupo tónico se realiza en función de la posición de éste en el grupo fónico y del tipo de acento. Finalmente a nivel alofónico se asignarán valores de f0 a cada alófono perteneciente al entorno de los eventos prosódicos más importantes (acentos y fronteras) siempre que pueda encontrarse un patrón local que coincida en la secuencia de alófonos sonoros y sordos en las sílabas acentuada y final. En otro caso los valores de f0 se obtienen a partir de interpolación de los valores asignados en los niveles superiores.

Este modelo jerárquico permite una concatenación coherente de los distintos patrones de grupos tónicos, preservando de esta manera la declinación natural de f0. Otras ventajas son la escalabilidad (insertando representantes en la base de datos) y la rápida adaptación del modelo a nuevos locutores por medio de la creación de nuevas bases de datos prosódicas.

8 *Conclusiones*

En esta comunicación se propone un entorno adecuado para la descripción de contornos de f0. La jerarquía empleada en este sistema permite la comparación, de forma síncrona, de diversos fenómenos con distinto alcance. A nivel de grupo fónico, en base a la discretización del contorno mediante un punto por cada grupo tónico, es observable el fenómeno de declinación y como la estructura sintáctica parece afectar a este contorno. Igualmente, a nivel de grupo tónico, mediante un punto por sílaba se manifiesta adecuadamente el movimiento asociado al acento de entonación, y al grupo tónico final de grupos fónicos. Finalmente, a nivel segmental y gracias a la utilización de parámetros como cf0 o f0dif, ha sido posible cuantificar movimientos de f0 en el entorno próximo de los eventos de entonación más influyentes perceptualmente. Actualmente, esta codificación o parametrización jerárquica del contorno de f0, es la base de experimentación de un nuevo modelo prosódico en nuestro sintetizador de voz.

Referencias

[Fujisaki 97] Hiroya Fujisaki. 1997. Analysis and modeling of fundamental frequency contours of greek utterances. Eurospeech 97. Vol. 1, pp. 465-468

[Salgado & Banga 00] Xavier Fernández Salgado, Eduardo Rodríguez Banga. 2000. A Hierarchical Model For Synthesizing F0 Contours. Enviado para evaluación a ICSLP 2000. China.

[Bachenko & Fitzpatrick 90] J. Bachenko, E. Fitzpatrick. 1990. A computational grammar of discourse-neutral prosodic phrasing in English. Computational Linguistics, 16(3), pp. 155-170

[Prieto 95] Pilar Prieto. 1995. Tonal alignment patterns in Spanish. Journal of Phonetics 23, pp. 429-451

[O'Shaughnessy 79] Douglas O'Shaughnessy. 1979. Linguistic features in fundamental frequency patterns. Journal of Phonetics 7, pp. 119-145

[Garrido 96] J. M. Garrido. 1996. Modelling Spanish Intonation for Text-to-Speech Applications, Tesis Doctoral, Universitat Autonoma de Barcelona

[Hirst & Espesser 93] Daniel Hirst, Robert Espesser. 1993. Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function. Travaux de l' Institut de Phonétique d' Aix, vol. 15, pp. 75-85

[Llisterri et al 95] J. Llisterri, R. Marín, C. De La Mota, A. Ríos. 1995. Factors affecting F0 peak displacement in Spanish. Eurospeech 1995, vol. 3 pp. 2061-2064

[Thorsen 83] N. Thorsen. 1983. Standard Danish intonation, phonetic data and their interpretation, Folia Linguistica, 17, pp. 187-220

[Quilis 64] A. Quilis, J. A. Fernández. 1964. Curso de fonética y fonología españolas, p. 149. C.S.I.C. Madrid

[Marín et al 96] Rafael Marín, Lourdes Aguilar, David Casacuberta. 1996. El grupo acentual categorizado como unidad de análisis sintáctico-prosódico. Actas del XII Congreso Lenguajes Naturales y Lenguajes Formales, Barcelona, PPU, pp. 487-494

[Rietveld & Gussenhoven 95] T. Rietveld , C. Gussenhoven. 1995. Aligning of pitch targets in speech synthesis: effects of syllable structure. Journal of Phonetics 23, 375-385

[Prieto & Shih 95] Pilar Prieto, Chilin Shih 1995. Effects of Tonal Clash on Downstepped H* Accents in Spanish. Eurospeech 1995, pp. 1307-1310

[Arvaniti et al 88] Amalia Arvaniti. D. Robert Ladd and Ineke Mennen.1998. Stability of tonal alignment: The case of Greek prenuclear accents. Journal of Phonetics 26, pp. 3-25

[Whalen & Levitt 95] D. H. Whalen, Andrea G. Levitt. 1995. The universality of intrinsic F0 of vowels. Journal of Phonetics 23, 349-366

[Allen et al 87] Jonathan Allen, M. Sharon Hunnicutt, Dennis Klatt. 1987. From text to speech: The MITalk system. Cambridge University Press

[Tuerk] Tuerk C. Dynamic pitch tracker. Cambridge University Engineering. ftp://svrftp.eng.cam.ac.uk/comp.speech/analysis/Pitch_Tracker-1.0.tar.Z