# Modelos de entonación analítico y fonético-fonológico aplicados a una base de datos del español de Buenos Aires.

**Jorge A. Gurlekian, Humberto Torres y Laura Colantoni**
Laboratorio de Investigaciones Sensoriales. CONICET. Instituto de Neurociencias Aplicadas. Hospital de Clínicas. UBA, Av. Córdoba 2351, 9 Piso, Sala 2. (1120) Buenos Aires. Argentina. *anagraf99@yahoo.com.ar*

**RESUMEN**

En este trabajo evaluamos un modelo analítico-cuantitativo y otro fonético-fonológico de las características entonativas obtenidas para una base de datos de 741 oraciones declarativas de foco amplio para el español de Buenos Aires. La descripción cuantitativa es la resultante de la aplicación del modelo de superposición de contornos de frecuencia fundamental propuesto por Fujisaki (2003) para diversas lenguas. La descripción fonética que utiliza la marcación de índices de juntura y tono ToBI (Beckman y Ayers, 1993) surge de la percepción de los grupos entonativos y de las prominencias y la aplicación de un método de etiquetado manual (Gurlekian y otros, 2001b), denominado ToBI ampliado (ToBI-A). Los resultados obtenidos, extienden la validez del modelo analítico para un gran número de oraciones declarativas de foco amplio del español. Asimismo se comprueba la validez del ToBI-A para producir un contorno de entonación comparable con el real a partir de las nuevas etiquetas propuestas. Para este fin, se realiza una prueba perceptual de comparación por pares y se calculan los errores respecto de la curva real que resultan ser del mismo orden para ambos modelos. También se verifica la validez del ToBI-A en las aplicaciones de síntesis de habla y para la definición fonológica de los acentos tonales de variedades no estudiadas del español.

**ABSTRACT**

We evaluate here the application of two intonational models –quantitative and phonetic- to the analysis of an Argentine Spanish database of 741 broad-focus declarative sentences. The analytic model is the superpositional model proposed by Fujisaki (2003) for several languages. The phonetic model is the result of the application of a labelling method (Gurlekian et al., 2001b) that incorporates psycho-acoustic measurements and a detailed description of the shape of the accent. Parameters generated by this labelling method were used to synthesize the intonational contours, which were then evaluated in a perception test. Results indicate the validity of Fujisaki´s model for describing a large database of Spanish broad-focus declaratives, and, thus, suggest the importance of Fujisaki's model for speech technology applications. The extended ToBI model (ToBI-A) is validated by the correlation coefficient and RMSE values as well as the results of the perception test. Ten native speakers of the variety under study judged the synthesized sentences as highly natural with only minor differences with the original contour. These results indicate that the ToBI-A (i) is adequate for a linguistically-meaningful description of the intonation of a new variety; (ii) can be adequately used for modelling intonation.

## 1 INTRODUCCION

El modelo analítico de entonación (Fujisaki, 2003) permite obtener un conjunto de parámetros reducido con los que se pueden representar contornos de entonación reales de una forma compacta y automática. La precisión de este modelado ha sido verificada para unas pocas frases del español de Buenos Aires (Fujisaki y otros, 1994) y en este trabajo el primer objetivo es comprobar su aplicación a una gran base de datos de oraciones declarativas de la misma variedad del español. La extensión de la validez del modelo analítico a una base de oraciones permitirá realizar asociaciones entre sus parámetros y aspectos lingüísticos únicamente dependientes del texto escrito para su aplicación en los sistemas de conversión de texto a habla de alta calidad.

Paralelamente, los desarrollos recientes en la teoría fonológica proponen una modelación de la entonación para las tecnologías del habla en dos pasos (Beckman y Pierrehumbert, 1986). El primero requiere la definición de los acentos tonales (Beckman y Ayers, 1993) y el segundo consiste en la predicción del contorno de F0 a partir de estos acentos. Para realizar el primer paso, se requiere de un inventario de acentos tonales. Aunque se han propuesto algunos para otras variedades, no se contaba con ninguno para el español de Buenos Aires. Por ello, se optó por una descripción fonética más que fonológica. Para obtener esta descripción llamada 'ToBI ampliado' (ToBI-A) se partió del ToBI tradicional (Beckman y Ayers, 1993) al que se le incorporó información sobre la codificación de los movimientos tonales alrededor del acento tonal (Gurlekian y otros, 2001b; 2003; Colantoni y Gurlekian, 2002, 2004). Se hipotetizó, además, que esta incorporación facilitaría la modelación del contorno de F0, dado que en intentos previos con el ToBI tradicional se requirieron técnicas sofisticadas de modelación (Ross, 1995; Black y Hunt, 1996). Así, el segundo objetivo de este trabajo es demostrar la validez del denominado ToBI-A en la generación directa de los contornos de entonación. Mediante la evaluación perceptual de un subconjunto de oraciones tomadas al azar realizada por oyentes nativos con entrenamiento musical, se verificará la adecuación de los contornos obtenidos.

Por último, las dos aproximaciones para la definición de los contornos de entonación, -analítica y fonético-fonológica serán comparadas con respecto de los contornos originales utilizando la raíz del error cuadrático medio (RMSE) y el coeficiente de correlación ($R^2$), que han demostrado ser los más efectivos para la evaluación cuantitativa (Escudero y otros, 2002).

El trabajo se estructura de la siguiente forma: primero se presenta en (2) la base de datos utilizada, en (3) se describe el modelo analítico y su aplicación a la base de datos, luego en (4) se presenta el método de etiquetado ToBI-A y su aplicación a la misma base, la evaluación perceptual del método ToBI-A y la comparación cuantitativa de ambas aproximaciones con los contornos reales se indica en (5). En la discusión, (6) se analizan las correspondencias de la evaluación perceptual y las medidas de similitud entre los contornos reales y estimados por el método ToBI-A.

## 2. BASE DE DATOS PROSODICA

La base de datos original (Gurlekian y otros, 2001b), de tipo relacional y basada en el lenguaje SQL, consiste en 741 oraciones declarativas que emplean el 97% de las sílabas del español en las dos condiciones de acento (sílaba acentuada y no acentuada) y en todas las variantes posicionales (inicial, media y final) dentro de la palabra. El 70 % de las oraciones fueron obtenidas de los periódicos que se publican en Buenos Aires. El resto fue creado por maestros de lengua quienes recibieron la instrucción de elaborar oraciones con palabras que contuvieran las sílabas menos frecuentes.

Las oraciones fueron grabadas por dos locutores hablantes nativos de Buenos Aires (masculino y femenino), pero en este trabajo, se analizan solamente las producidas por el hablante femenino. Cada una de las emisiones fue

etiquetada dos veces por cuatro fonoaudiólogas con entrenamiento musical, utilizando el programa de análisis, síntesis y etiquetamiento Anagraf (Gurlekian, 1997) (ver Figura 1). Cada onda tiene asociada once archivos de etiquetado: cuatro del etiquetado tonal, uno del fonético (Gurlekian y otros, 2001a), uno de datos acústicos (F0, energía y formantes), cuatro de categorías sintácticas y uno de clases de palabras. La base de datos, se pobló con los archivos de onda y los archivos de etiquetado asociados.

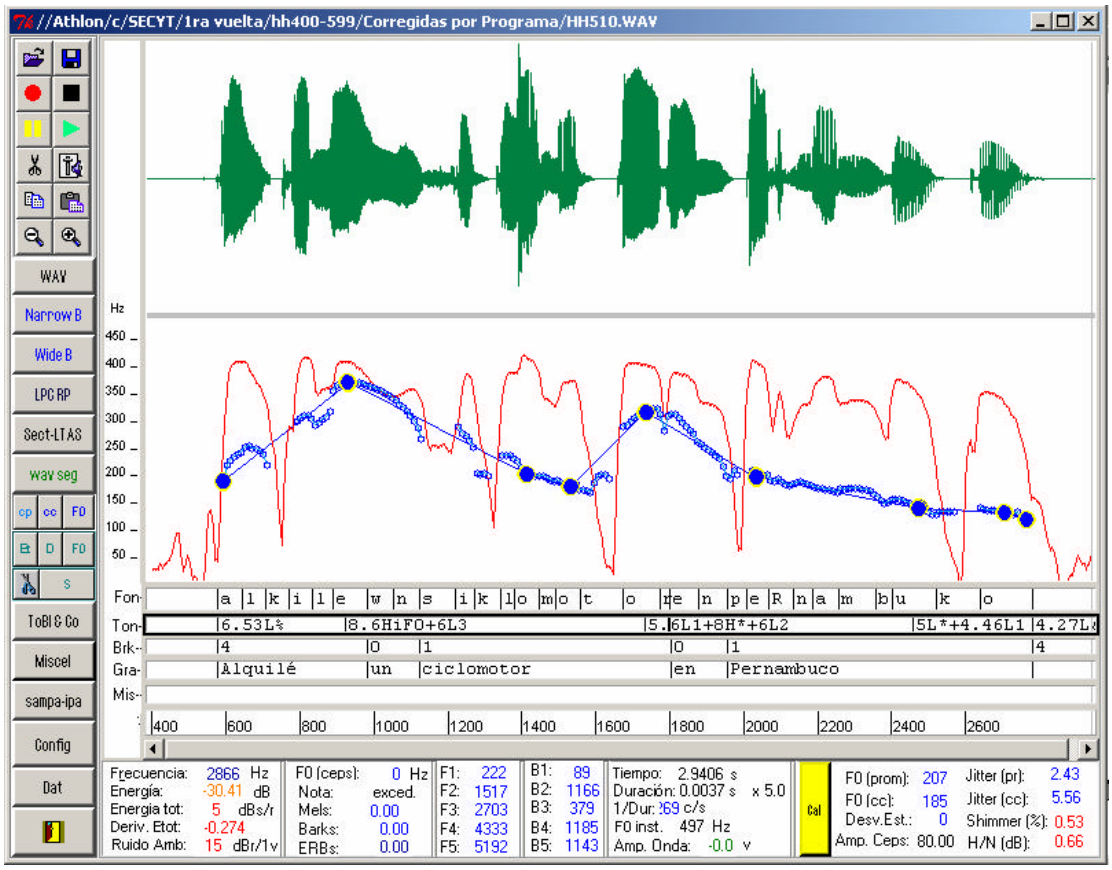



***Figura 1.*** *Se observa la forma de onda de la frase "Alquilé un ciclomotor en Pernambuco" (arriba), el contorno real de F0 (círculos pequeños), el contorno obtenido a partir del método de etiquetado ToBI Ampliado (círculos grandes unidos por líneas rectas continuas) y el contorno de Energía total (línea contínua). Más abajo se presentan los diferentes niveles de etiquetado (fonético, tonal, de junturas, grafémico y de misceláneas) obtenidos mediante el programa Anagraf (Gurlekian, 1997).*

## 3. MODELO ANALÍTICO

### 3.1 Descripción

La aproximación cuantitativa propuesta por Fujisaki (2003) se basa sobre la modelación del mecanismo de vibración de cuerdas vocales. Este modelo describe analíticamente el contorno de F0 en una escala logarítmica, como la superposición de dos componentes: los acentos tonales y acentos de frase. Los acentos de frase se generan

como respuesta a la excitación con una función delta llamada `comando de frase´, a un filtro lineal de segundo orden amortiguado críticamente. Los acentos tonales se obtienen como respuesta de otro filtro igual a la excitación de una función escalón llamada `comando de acento´. La ecuación en (1) expresa el modelo de entonación enunciado, donde las sumatorias representan los componentes de frase y acento respectivamente:

$$\ln(F0) = \ln(F_{min}) + \sum_{i=1}^{N_f} Af_i \cdot Gf_i(t - T0_i) + \sum_{j=1}^{N_a} Aa_j \cdot [Ga_j(t - T1_j) - Ga_j(t - T2_j)] \quad (1)$$

donde,

$$Gf_i(t) = \begin{cases} \alpha_i^2 \cdot t \cdot e^{-\alpha_i t} & \text{para } t \geq 0 \\ 0 & \text{para } t < 0 \end{cases} \quad (2)$$

y

$$Ga_j(t) = \begin{cases} \min[1 - (1 + \beta_j t)e^{-\beta_j t}, \gamma_j] & \text{para } t \geq 0 \\ 0 & \text{para } t < 0 \end{cases} \quad (3)$$

además,
**Fmín** es el nivel de base de F0
**Afi** es la Amplitud del comando de frase i
**Gfi** representa la respuesta al impulso del mecanismo de control de frase
**T0i** es el instante de tiempo donde ocurre el comando de frase i
**Aaj** es la Amplitud del comando de acento j
**Gaj** representa la respuesta al escalón del mecanismo de control de acento
**T1j** es el tiempo de inicio del comando de acento j
**T2j** es el tiempo de finalización del comando de acento j
$\alpha_i$ es el autovalor del mecanismo de control de frase, para el comando de frase i
$\beta_j$ es el autovalor del mecanismo de control de acento, para el comando de frase j
$\gamma_j$ es el valor máximo del componente de acento, para el comando de acento j

Los parámetros $\alpha$ = 2 y $\beta$ = 20 caracterizan las propiedades dinámicas de los mecanismos laríngeos de control de frase y acento y se consideran junto con $\gamma$ = 0.9 prácticamente constantes para todos los hablantes. `Fmín´ se obtiene de cada emisión. Finalmente, los parámetros que deben calcularse son la existencia o no de los comandos de frase, los valores de amplitud de los acentos de frase `Af´ y de los acentos tonales `Aa´ y los tiempos T0, T1 y T2.

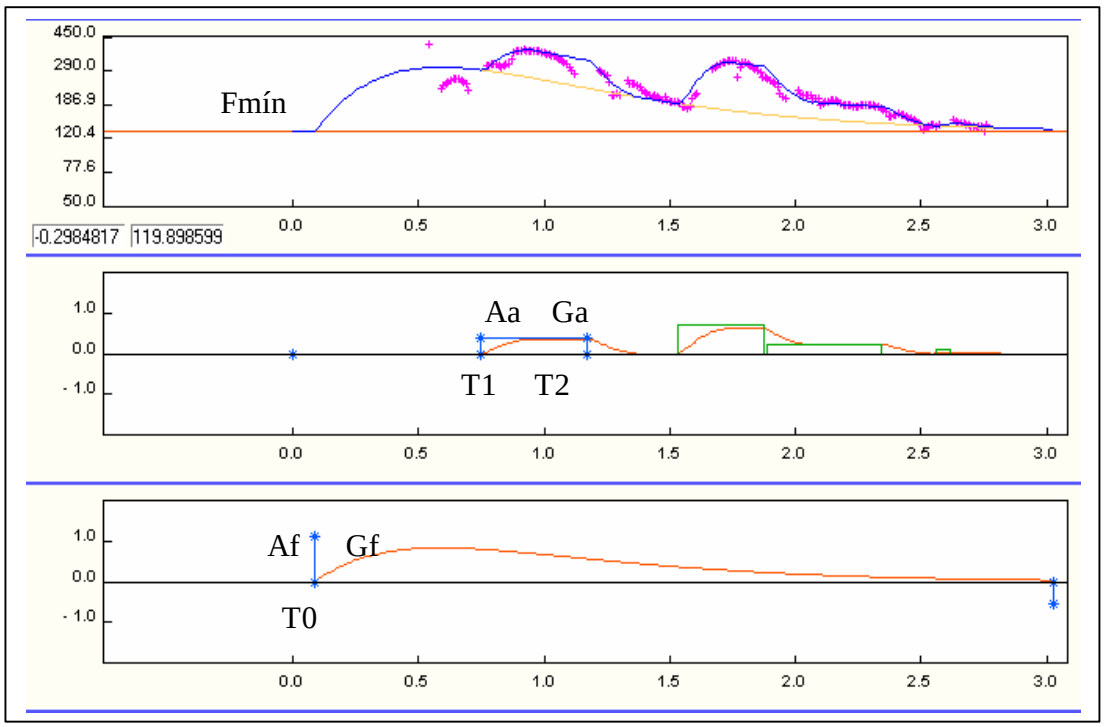



***Figura 2.*** En la parte superior se observa el contorno real de F0 (línea de cruces) y el obtenido mediante la aplicación del modelo de Fujisaki (línea continua) para la frase: "Alquilé un ciclomotor en Pernambuco". En la parte media se observan los comandos de acento tonal y la respuesta obtenida a cada escalón. En la parte inferior se ve el impulso al inicio que corresponde al comando de frase y la respuesta continua obtenida mediante el programa AutoFujiPos, (Mixdorff, 2000).

### 3.2 Aplicación

La aplicación del modelo analítico requirió del cálculo de los parámetros del modelo para cada una de las 741 emisiones. Para ello se utilizó el programa AutoFujiPos (Mixdorff, 2000) que recibe como entrada el contorno real de F0 en formato ASCII y genera a la salida los parámetros del modelo. Este programa estima los parámetros del modelo partiendo del contorno de F0 calculado por el método de programación dinámica ESPS/RAPT (Talkin,1995) y continúa con una serie de pasos que contemplan una estilización cuadrática "spline", la separación de los componentes de frase y acento mediante filtrado, y una inicialización de comandos.

Luego, se optimiza la configuración inicial de parámetros con tres pasadas del método "hill-climb". En la parte final del procedimiento los parámetros de los componentes de frase y acento se optimizan separadamente y luego juntos, usando el contorno estilizado como contorno final. Finalmente, se realiza un ajuste fino considerando una versión ajustada del contorno extraído como contorno final.

La verificación de la efectividad de los parámetros calculados se comprobó al generar contornos de F0 para las 741 oraciones a partir de los parámetros obtenidos, que se compararon con los contornos de F0 originales (ver Anexo 1).

En síntesis: la aplicación del modelo requirió de los siguientes pasos:

1. Cálculo del contorno de F0 original. (ANAGRAF, método ESPS/RAPT )
2. Transformación del contorno a formato ASCII
3. Cálculo de los parámetros de Fujisaki (Autofujipos)
4. Cálculo del contorno de F0 mediante los parámetros de Fujisaki para las 741 oraciones (ver Anexo 1)
5. Comparación de los contornos original y calculado. Medición de la raíz del error cuadrático medio (RMSE), y Coeficiente de Correlación ($R^2$).

La ecuación utilizada para el cálculo del RMSE es:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (ERB_{orig}(i) - ERB_{est}(i))^2}{N}} \qquad (4)$$

donde,

$$ERB = 16.7 \log_{10}\left(1 + \frac{F0}{165.4}\right) \qquad (5)$$

y,

F0 valor en Hz
$ERB_{orig}(t)$ valor en ERB para cada ventana temporal de análisis espectral del contorno original
$ERB_{est}(t)$ valor en ERB para cada ventana temporal de análisis espectral del contorno estimado
N total de ventanas temporales de análisis espectral que contienen F0 en la oración original.

## 4. TOBI AMPLIADO

### 4.1 Descripción

Para el desarrollo de nuestro modelo de etiquetado, hemos adoptado como referencia el modelo métrico-autosegmental (Pierrehumbert, 1980; Beckman y Pierrehumbert, 1986). Este marco teórico tiene la doble ventaja de haber sido aplicado al español (Sosa 1991, Sosa 1999) y de haber resultado en un método de transcripción relativamente estandarizado, conocido como ToBI – *Tonal and Break Indices* - (Beckman y Ayers, 1993). Originalmente propuesto para el inglés, el ToBI ha sido adaptado a varias lenguas incluido el español (Beckman y otros, 2002). El SP-ToBI (ToBI para el español) propone un inventario reducido de acentos (ver Tabla 1), obtenido a partir del estudio de algunas variedades latinoamericanas y peninsulares, entre las que no se encuentra el español de la Argentina. Por ese motivo, cuando iniciamos nuestra investigación sobre esta variedad, decidimos no restringirnos al inventario propuesto, sino optar por un modelo que, en principio, permitiera todas[1] las combinaciones lógicas de ambos tonos, como se muestra en la Tabla 2.
Nuestro modelo de etiquetado, al que denominamos ToBI-A (ToBI ampliado) se complementa con otras dos modificaciones: (i) una descripción detallada de la forma del acento; (ii) el uso de una escala psico-acústica para la descripción del F0. La primera modificación incorpora información sobre la dirección de la pendiente de F0 y la cantidad de sílabas en las que se extiende el movimiento tonal. Así, para etiquetar el acento tonal, se identifican

[1] Dos unitonales, ocho bitonales y ocho tritonales derivados.

primero las sílabas prominentes[2]. Luego, el etiquetador decide si esa prominencia se alcanza por una subida, una bajada de F0 o una combinación de ambas. Una vez tomada la decisión, se marca la cabeza del acento (H* o L*) y se insertan las marcas asociadas en el punto más alto o más bajo del F0, respectivamente. Es importante destacar aquí que el signifcado de las marcas tonales difiere ligeramente con el propuesto por Pierrehumbert (1980) o Beckman y Pierrehumbert (1986), donde se supone que el tono con el asterisco ocurre en las proximidades del acento, pero no coincide necesariamente con su intervalo temporal. Aquí el asterisco indica que el pico o el valle del F0 está alineado con la sílaba acentuada. Terminada la marcación de la cabeza, se agrega información (de ser necesaria) acerca de la entrada y la salida del acento. Para ello se tienen en cuenta el número de sílabas (indicadas con 'm'[3]en la Tabla 2) en las que se alcanza un pico desde un valle (o viceversa) y el número de sílabas en el que se sale de un pico o un valle.

La segunda modificación propuesta por el ToBI-A consiste en el en el uso de una escala psicoacústica para la descripción del F0: la escala de proporciones ERB "Equivalent Rectangular Bandwidth" (ver 3.2, ecuación 5), (Patterson 1976; Glasberg y Moore 1990). Existen evidencias en favor del uso de una escala de ERBs en la percepción de prominencia. De acuerdo con Hermes y Van Gestel (1991), las prominencias de movimientos tonales en diferentes registros son iguales cuando sus excursiones son iguales en la escala de ERB. El nivel en ERB se aplica a todos los tonos que constituyen el acento tonal (indicado con 'n'[4] en la Tabla 2). Por ejemplo, la etiqueta [7H* + 5L1] representa un tono alto H (cabeza del acento tonal) con un nivel de 7 ERBs medido en el instante donde F0 es máximo; la curva cae a 5 ERBs (medido en el instante en el que se alcanza el valle) una sílaba más tarde. La información relativa al instante temporal del tono complementario no es relevante (ver 4.2, paso 2).

***Tabla 1.*** *Inventario Fonológico de Acentos Tonales para el SP-ToBI (Beckman y otros, 2002). (*L tono bajo, H tono alto, * cabeza de acento, % tono de juntura, ¡! tonos escalonados).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **H*[5]**<br>***[5]** | **-** | **L*+H**<br>**L*+!H** | **-** | **-** | **L+H***<br>**L+(¡)H*** | **-** | **H+L*** | **-** | **-** |

***Tabla 2.*** *Inventario Fonético del ToBI-Ampliado. (*L tono bajo, H tono alto, * cabeza de acento, n valor de F0 en ERBs, y m número de sílabas).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **nH*** | **nL*** | **nL*+ nHm** | **nL*+ nLm** | **nLm+ nL*** | **nLm+ nH*** | **nH*+ nLm** | **nHm+ nL*** | **nH*+ nHm** | **nHm+ nH*** |

La combinación de los acentos bitonales descriptos en la Tabla 2 dan lugar a la definición de ocho acentos tritonales. Estos se emplean cuando debe indicarse el movimiento del contorno a ambos lados de la cabeza de acento. (Ver ecuación 6 en 4.2).

El nivel tonal se completa con la descripción de los eventos en el final (y en el comienzo) del grupo entonativo, conocidos como acentos de frases y tonos de juntura. Los acentos de frase H- y L- se han mantenido en el ToBI-A,

[2] La prominencia está determinada por los siguientes factores (Ladd, 1996): (a) Relaciones abstractas entre sílabas, palabras o entre frases determinadas por una estructura jerárquica (factor fonológico métrico (b) B. Relaciones concretas (psico)acústicas de acento entre esas unidades debidas cambios en la sonoridad duración, F0 y timbre (factor fonético). En el castellano, los acentos tonales se perciben y se asocian con sílabas métricamente fuertes.

[3] La variable 'm' puede tomar cualquier valor de 1 a 7, pero normalmente no incluye valores superiores a 3.

[4] La variable 'n' puede tomar valores de 1 a 12.

[5] En el SP-ToBI estos acentos se utilizan cuando no pueden aplicarse algunos de los tres tipos de acentos tonales definidos. En el ToBI Ampliado, H* y L* se emplean en los casos en que los tonos adyacentes ya están definidos por otro acento o el tono de juntura.

como se propone en el ToBI original, aun cuando los acentos bitonales puedan reemplazarlos (Sosa, 1991, 1999). Esta decisión se ha tomado, una vez más, con el fin de lograr una descripción más exhaustiva del contorno entonativo. Los tonos de juntura contemplados en el SP-ToBI son H%, L% y M%, en los casos donde se alcanza un nivel medio después de acentos H*. En el modelo ToBI-A, se utilizan sólo los tonos, L% y H% con la indicación del nivel "n" en todos los casos. Se entiende que los niveles indicados en ERBs agregan información más detallada que la reflejada por M%.

Además del nivel tonal, el ToBI tradicional utiliza cuatro niveles de trascripción: ortográfico, de junturas, y misceláneas. El nivel de juntura incluye marcas que indican el grado de separación entre las palabras y grupos entonativos. En el modelo propuesto se utilizan cinco índices de juntura como se presentan en el ToBI original (Beckman y Ayers, 1993). La ventaja de mantener estas categorías reside en que la fusión de sílabas tiene grados que son representados por los índices 0, 1 y 2. Los límites de palabras que separan grupos entonativos con pausa o cambios tonales se representan con los índices 3 y 4.

Es importante tener en cuenta que este método ha sido diseñado para la modelación (y no sólo para la descripción) de la entonación del español de Buenos Aires, de ahí el nivel de detalle del etiquetado. Este nivel de detalle sería sumamente difícil de manejar si no fuera por la utilización de una base de datos prosódica (ver § 2), que permite hacer recuentos estadísticos de los tipos de acentos y, así, colapsar categorías. Por ejemplo, trabajos preliminares (Colantoni y Gurlekian 2002, 2004) indican que el número teórico original de dieciocho acentos puede ser reducido a seis (H*+L, L+H*, H*+H, L*+L, L*+H, H+L*), dado el bajo nivel de ocurrencia (o la no ocurrencia) de algunas variantes. Además, la aplicación de este modelo ha mostrado (Colantoni y Gurlekian, 2004) que la realización más frecuente de los acentos prenucleares en las declarativas de foco amplio (un pico alineado dentro de la sílaba acentuada y un valle en la postónica) no se corresponde directamente con ninguna de las etiquetas propuestas por el SP-ToBI y difiere de las realizaciones descriptas para otras variedades del español (Garrido y otros, 1993; Prieto y otros, 1995; Hualde, 2002; Face 2001, entre otros). Esta alineación temprana, sin embargo, ha sido descripta en un contexto pragmático diferente, i.e. la realización del foco de contraste (De la Mota, 1997; Face 2001). Se espera así que la extensión en la utilización del ToBI-A nos permita llegar a una definición de acentos tonales fonológicos del español de Buenos Aires y de otras variedades de la Argentina.

### 4.2 Aplicación

La aplicación del ToBI-A consiste en la reconstrucción del contorno de F0 a partir del etiquetado de la base de datos prosódica. Como se mencionó, las etiquetas utilizadas pueden tener distinta extensión y complejidad. Para la modelación de los contornos se considera en forma genérica que el acento tonal más complejo es un tritonal compuesto por tres términos del tipo:

$$[n(L;H)m] + n(L;H)^* + [n(L;H)m] \qquad (6)$$

En la ecuación (6) los corchetes indican estructuras opcionales. Los paréntesis indican opciones alternativas L ó H, con su correspondiente valor de n y m. El segundo término está siempre presente y asociado con la marca temporal que realizó el etiquetador al marcar la cabeza del acento tonal en la vocal de la sílaba prominente y por lo tanto no requiere el valor de m (número de sílabas a la derecha o izquierda).

La Tabla 3 muestra la totalidad de símbolos utilizados en el ToBI-A para describir los acentos tonales y los contornos de entonación.

***Tabla 3.*** *Parámetros del ToBI Ampliado*

| **Marca** | **Tono** | **ERB n** | **Sílaba m** | **Tiempo** |
|---|---|---|---|---|
| **1** | {L, H, -1} | {-1; R+: 1: 12} | {-1, I :0:7} | |

| | | | | |
|---|---|---|---|---|
| **2** | {L*; L%; L-, H*; H%; H-, -1} | {-1; R+: 1: 12} | | {-1, (R +)} |
| **3** | {L; H; -1} | {-1; R+: 1: 12} | {-1, I: 0:7} | |

Donde, -1: indica ausencia de marca , L: tono bajo, H: tono alto, *: cabeza de acento, %: tono de juntura., - acento de frase, R+: real positivo, I: entero.

El procedimiento para obtener los contornos de entonación a partir del etiquetado manual con ToBI-A y su comparación con el contorno real es el siguiente:

1. Para cada una de las marcas temporales asociadas a la cabeza de acento tonal, tonos de juntura y acentos de frase, de una frase entonativa se definen los valores de F0 a partir de los valores en ERBs extraídos del segundo término.
2. Se analizan los términos a la izquierda y derecha del núcleo del acento. Teniendo en cuenta el número de sílabas indicado, se coloca una marca temporal en el centro[6] de la vocal de la sílaba correspondiente. Se asigna el valor de F0 indicado por el valor en ERBs en ese instante.
3. Se calculan por interpolación lineal los valores de F0 entre los puntos definidos.
4. Se compara la similitud física entre ambos contornos mediante el cálculo del error RMSE (calculada como en (4) y el coeficiente de correlación $R^2$ entre los valores del contorno real y el etiquetado para cada instante.

## 5. EVALUACIÓN DEL TOBI AMPLIADO y COMPARACIÓN CON EL MODELO ANALÍTICO

### 5.1 Evaluación perceptual y cuantitativa del ToBI Ampliado

La similitud perceptual de los contornos de entonación fue evaluada mediante el juicio de diez oyentes nativos con entrenamiento musical. La tarea de los sujetos consistió en asignar una valoración numérica de 0 a 5 al grado de semejanza entonativa de la emisión resintetizada[7] con el contorno estimado mediante el etiquetado con ToBI-A respecto de la original. La estimación numérica se realizó de acuerdo con la Tabla 4. Del total de oraciones se han seleccionado 19 al azar para realizar esta prueba de percepción auditiva, empleando el método de comparación por pares.
La emisión original tambien fue resintetizada con su contorno de F0 original intacto, para que los efectos secundarios producidos por el método de síntesis afectaran de manera idéntica a ambas señales.

***Tabla 4.*** *Definición de las categorías utilizadas en la prueba perceptual.*

| Escala Subjetiva | Niveles de similitud |
|---|---|
| **5** | Idéntico, indistinguible del original |
| **4** | Muy similar, con una alteración |
| **3** | Similar, con dos alteraciones |
| **2** | Poco similar, con tres alteraciones o más |
| **1** | Se percibe algo artificial |
| **0** | Muy artificial |

[6] La colocación de esta marca en el inicio y final de la vocal no produjo cambios significativos en el valor del error.
[7] Se empleo el método PSOLA (Moulines y Laroche, 1995) en el programa Anagraf.

Se presentan a continuación (ver Tabla 5) los resultados de la evaluación perceptual más los valores de RMSE de cada frase con respecto del contorno original obtenido mediante el método de ToBI-A.

***Tabla 5.*** *Resultado de la Evaluación Perceptual y comparación con las medidas cuantitativas de RMSE en semitonos (S.T.), ERBs y correlación $R^2$, obtenidos para la descripción de ToBI ampliado.*

| | **Evaluación Perceptual ToBI Ampliado** | | | **Evaluación Cuantitativa ToBI Ampliado** | | |
|---|---|---|---|---|---|---|
| **Nro. de Frase** | **Valor Medio** | **Desviación Estándar** | | **RMSE S.T.** | **RMSE ERBS** | **Coef. $R^2$ Correlación** |
| 1 | 3.83 | 0.73 | | 2.11 | 0.503 | 0.934 |
| 2 | 3.60 | 0.82 | | 1.41 | 0.327 | 0.941 |
| 3 | 4.57 | 0.51 | | 1.57 | 0.368 | 0.938 |
| **4** | **2.98** | **0.87** | | **2.34** | **0.552** | **0.808** |
| 5 | 3.20 | 1.17 | | 1.96 | 0.467 | 0.841 |
| 6 | 4.15 | 0.63 | | 2.24 | 0.524 | 0.854 |
| 7 | 4.78 | 0.42 | | 1.81 | 0.426 | 0.925 |
| 8 | 4.44 | 0.70 | | 1.57 | 0.363 | 0.902 |
| 9 | 4.67 | 0.48 | | 1.17 | 0.282 | 0.971 |
| ***10*** | ***5*** | ***0*** | | **1.01** | **0.235** | **0.982** |
| 11 | 4.37 | 0.51 | | 1.49 | 0.340 | 0.959 |
| 12 | 4.06 | 0.56 | | 2.26 | 0.534 | 0.821 |
| 13 | 4.03 | 0.73 | | 2.19 | 0.519 | 0.919 |
| 14 | 5 | 0 | | 1.17 | 0.262 | 0.946 |
| 15 | 4.47 | 0.52 | | 1.51 | 0.357 | 0.933 |
| 16 | 3.68 | 0.91 | | 2.56 | 0.609 | 0.856 |
| **17** | **3.87** | **0.94** | | **3.01** | **0.718** | **0.827** |
| 18 | 4.57 | 0.51 | | 1.33 | 0.319 | 0.954 |
| 19 | 4.03 | 0.73 | | 2.19 | 0.510 | 0.877 |
| **Totales** | ***3.96*** | ***0.59*** | | ***1.83*** | ***0.43*** | ***0.904*** |

La evaluación perceptual de la similitud de los contornos entonativos es el marco de referencia para evaluar la adecuación de la aproximación mediante el método de ToBI-A. La prueba de evaluación auditiva en la que se comparaban los contornos obtenidos a partir del método de etiquetado ToBI-A con los reales refleja una respuesta promedio cercana a 4, en una escala que va de 0 a 5; es decir, se estima que los contornos obtenidos son muy similares con una alteración. (Ver Anexo 1).

La evaluación cuantitativa con el RMSE medio de estas 19 frases es de 0.43 ERBs o 1.83 ST, valor que está levemente por encima del umbral de detección de cambios tonales, estipulada en 1.5 ST (Toledo, 2000; Bertrán y otros, 2002).

En general, la evaluación objetiva está de acuerdo con la evaluación subjetiva. A menor RMSE y menor distancia en ST, las evaluaciones perceptuales se acercan a la máxima similitud (ver estímulos 10,15, 7, 8 y 18). Sin embargo debe notarse que al tratarse de valores promedio algunas oraciones pueden presentar el error concentrado en un segmento produciendo una diferencia tonal anómala perceptible, o tener el error distribuido a lo largo de la oración sin producir un cambio perceptual. Esta afirmación puede corresponder a las evaluaciones del los estímulos 4 y 17. Por ejemplo, el estímulo 17 presenta un RMSE promedio de 0.718 ERBs que representa 3 ST. Para este estímulo los sujetos respondieron con una evaluación de 4. El estímulo 4 produjo un RMSE de .0552 que representa 2.34 semitonos, es decir, mejor aproximación de los contornos, pero fue evaluado 3, es decir dos diferencias tonales.

Otra medida habitualmente utilizada que complementa el RMSE es el valor del Coeficiente de Correlación que mide la desviación del F0 medio a intervalos temporales, indicando con el valor 1, si ambos contornos tienen exactamente la misma tendencia a subir o bajar. En este caso la medida del Coeficiente de Correlación se

corresponde con la evaluación perceptual de los estímulos 4 y 17. (ver Tabla 5). El Coeficiente de Correlación promedio entre los contornos obtenidos por ToBI-A y los originales para toda la base es de 0.810.

### 5.2 Comparación

La Tabla 6 muestra el error RMSE en Hz, ERBs y en semitonos (ST) y el coeficiente de correlación $R^2$ obtenidos al comparar el contorno real y el estimado por los dos modelos para las 741 oraciones.

***Tabla 6.*** *RMSE y $R^2$ obtenidos para las 741 oraciones*

| Modelo \ Medida | **RMSE-Hz**[8] | **RMSE-ERBS** | **RMSE-ST** | **$R^2$** |
|---|---|---|---|---|
| **Fujisaki** | 16 | 0.311 | 1.40 | 0.93 |
| **ToBI Ampliado** | 29 | 0.565 | 2.39 | 0.81 |

La comparación sugiere que la aproximación con el ToBI-A presenta índices de error mayores a los obtenidos con el modelo analítico. Hay que tener en cuenta sin embargo que estamos comparando un método manual que atiende a las características lingüísticas del texto con un modelo matemático iterativo que las ignora. Aunque los errores son mayores, la comparación del ToBI-A con otras presentaciones similares refleja un desempeño equivalente. El RMSE promedio obtenido con el ToBI-A de 29 Hz es similar al obtenido con el ToBI estándar para el inglés. Sin embargo, para obtener este nivel de error, los modelos que utilizan el ToBI estándar deben complementarse con técnicas de predicción de parámetros complejas. Por ejemplo, Ross (1995) presenta un RMSE de 34 Hz utilizando etiquetamiento manual de marcas ToBI y un sistema de modelación dinámica. Black y Hunt (1996), quienes emplean una combinación de ToBI y técnicas de Regresión Lineal, reportan 34.8 Hz y una correlación de 0.62. Estos resultados muestran que la utilización directa de los parámetros del ToBI-A, aún con una simple interpolación lineal, pueden compararse exitosamente con las técnicas de modelación que convierten las marcas ToBI tradicionales en contornos de F0.

El RMSE obtenido para el total de las oraciones cuando se emplea el modelo analítico de Fujisaki es de 0.311 ERBs ó 16 Hz. y la Correlación es de 0.93. Estos valores son comparables con los datos presentados por Escudero (2002), quien obtuvo un RMSE 10.4 Hz mediante la parametrización con funciones de Bézer, y una correlación de 0.89 para una base de datos en español equivalente en número de oraciones. Para el inglés, Dusterhoff (2000) presenta un RMSE de 9.1Hz y un Coeficiente de Correlación de 0.74 estimando los parámetros del método TILT con árboles de regresión (CARTs).

## 7. CONCLUSIONES

Se confirma en este trabajo la aplicación del modelo analítico para una base de datos con un gran número de oraciones. Esta extensión, que ya ha sido verificada en córpora de pocas frases (Fujisaki y otros, 1994), permitirá la aplicación del modelo en los sistemas de conversión de texto a habla. Es necesario aun establecer una relación adecuada y automática entre los parámetros del modelo y la información lingüística presente en el texto.

El método ToBI-A también describe satisfactoriamente los contornos entonativos obtenidos mediante el etiquetado manual. Este modelo presenta una doble ventaja: (i) permite asociar más naturalmente las marcas prosódicas con las características del texto para su aplicación en sistemas TTS; (ii) facilita una caracterización fonética y fonológica del español de Buenos Aires y permitirá postular un inventario de acentos para esta variedad.

[8] Para obtener el valor en Hz de los semitonos se ha considerado el F0 promedio del hablante de esta base de datos (200Hz).

## 8. TRABAJOS FUTUROS

La utilización de una modelación más elaborada que la interpolación lineal debería mejorar aun más la aproximación obtenida con el ToBI-A, como puede deducirse de los ejemplos indicados en el Anexo 1. Se intentará asociar los parámetros de los modelos con las características sintácticas y de las estructuras silábica y de palabras de las oraciones mediante el uso de árboles de regresión. Se investigará la relación entre los modelos de Fujisaki y ToBI-A (Mixdorff y Fujisaki, 2000), para potenciar las ventajas de cada uno en una propuesta híbrida. Se espera definir el inventario de los acentos tonales típicos en esta variante del español mediante un análisis estadístico de los tipos fonéticos de acentos tonales y una posterior reducción a los acentos tonales fonológicos.

**Agradecimientos**
Este trabajo fue posible gracias a un subsidio PIP-CONICET, Nro 02489. Buenos Aires. Argentina

## REFERENCIAS BIBLIOGRÁFICAS

Beckman, M. E. y Pierrehumbert, J. (1986): «Intonational structure in Japanese and English», *Phonology Yearbook*, 3, pp. 255-309.

Beckman, M. E. y Ayers, G. M. (1994): «Guidelines for ToBI labelling», en: http://ling.ohio-state.edu/phonetics/E_ToBI. The Ohio State University Research Foundation.

Beckman, M. E., Díaz-Campos, M., McGory, J. T. y Morgan, T.A. (2002): «Intonation across Spanish, in the Tones and Break Indices Framework», *Probus* 14, pp. 9-36.

Bertrán, A. P., Planas, A. M., Celdrán, E. M., Escandell, A. O. y Céspedes, M. C. A. (2002): «Umbrales tonales en español peninsular», *II Congreso Nacional de Fonética Experimental*, Sevilla, Barrio, Marina, y otros. (eds.).

Black, A. y Hunt, A. (1996): «Generating F0 contours from ToBI labels using linear regression», *ICSLP96*, Philadelphia, PA, vol. 3, pp. 1385-1388.

Clark, R. A. J. y Dusterhoff, K. E. (1998): «Objetive methods for evaluating synthetic intonation», *ICSLP 98***.**

Colantoni, L. y Gurlekian, J. A. (2002): «Modeling intonation for synthesis: pitch accents and contour patterns in Argentine Spanish», *Laboratory approaches to Spanish phonology*, University of Minnesota.

Colantoni, L. y Gurlekian, J. A. (2004): «Convergence and intonation: historical evidence from Buenos Aires Spanish», *Bilingualism: Language and Cognition*, 7 (2), pp. 107-119.

De la Mota, Carme. (1997): «Prosody of sentences with contrastive new information in Spanish», en A. Botinis, G. Kouroupetrogl, N. Fakotakis, y E. Dermatas (eds.), *Intonation: theory, models and applications. An ESCA workshop*, Atenas, pp. 75-78.

Dusterhoff, K. (2000): «Synthesizing Fundamental Frequency Using Models Automatically Trained from Data», Tesis doctoral, University of Edinburgh.

Escudero Mancebo, D. y Cardeñoso Payo, V. (2001): «Modelo cuantitativo de entonación del español», *Revista de la SEPLN*, pp. 233-240.

Escudero Mancebo, D., González Ferreras, C. y Cardeñoso Payo, V. (2002): «Evaluación objetiva y subjetiva de entonación sintética», *Actas de las Jornadas de Tecnologías del Habla*., Departamento de Electrónica y Tecnología de Computadores de la Universidad de Granada. Sevilla, España.

Face, Timothy. (2001): «Focus and early peak alignment in Spanish intonation», *Probus*, 13, 223-46.

Fujisaki, H. (2003): «Prosody, Information and Modelling with emphasis on Tonal Features of Speech», *Proceedings Workshop on SLP*, Mumbai, India.

Fujisaki, H., Ohno, S., Nakamura, K, Guirao, M. y Gurlekian, J. A. (1994): «Analysis of accent and intonation in Spanish based on a quantitative Model», *ICSLP 94*, Yokohama, pp. 355-358.

Garrido, Juan, Llisterri, Joaquim, De la Mota, Carme, y Ríos, Antonio. (1993): «Prosodic differences in reading style: Isolated vs. Contextualized sentences», *EUROSPEECH '93*, 573-576.

Glasberg, Brian y Moore, Brian. (1990): «Derivation of auditory filter shapes from notched-noise data», *Hearing Research*, 47, 103-38.

Gurlekian, J. A. (1997): «El laboratorio de audición y habla del LIS», en M. Guirao (ed.): *Procesos Sensoriales y cognitivos.* Buenos Aires, Dunken., pp.55-81.

Gurlekian, J. A, Colantoni, L, y Torres, H (2001a): «El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados», *Revista Fonoaudiológica*, 47, 3, pp 58-70.

Gurlekian, J. A., Rodríguez, H. Colantoni, L. y Torres, H. (2001b): «Development of a Prosodic Database for an Argentine Spanish Text to Speech System», *IRCS Workshop on Linguistic Databases*, Philadelphia, pp. 99-104.

Gurlekian, J. A., Colantoni, L. y Torres, H. (2003): «Modelo de etiquetamiento prosódico para las tecnologías de habla», en *XV Congreso de la Sociedad Chilena de Lingüística*, Octubre, 2003, Santiago, Chile.

Hermes, Dik y Van Gestel, Joost (1991): «The frequency scale of speech intonation», *Journal of the Acoustical Society of America*, 90, 97-102.

Hualde, José I. (2002): «Intonation in Spanish and other Ibero-Romance languages: Overview and status questions», en Caroline Wiltshire and Josquin Camps (eds.), *Romance Phonology and Variation: Selected Papers from LSRL 30*, Amsterdam, Benjamins, pp. 101-115.

Ladd, D.R. (1996): *Intonational Phonology*, Cambridge, University Press.

Mixdorff, H. (2000): «A novel approach to the fully automatic extraction of Fujisaki model parameters», *ICASSP 2000*, Istanbul, 3, pp.1281-1284.

Mixdorff, H. y Fujisaki, H.(2000): «A quantitative descrption of German prosody offering symbolic labels as a by product», en ICSLP2000, Pekin, China, vol 2, pp.90-101.

Moulines, E. y Laroche, J. (1995): «Non parametric techniques for pitch scale and time domain scale modification of speech», *Speech Communication* 16, pp. 175-205.

Patterson, R.D. (1976): «Auditory filter shapes derived with noise stimuli», *Journal of the Acoustical Society of America*, 59, pp. 640-654.

Pierrehumbert, Janet. (1980): «The phonology and phonetics of English intonation», *Tesis doctoral. MIT.*

Prieto, Pilar, Van Santen, Jan, and Hirschberg, Julia. (1995): «Tonal alignment patterns in Spanish», *Journal of Phonetics*, 23, pp. 429-51.

Ross, K. N. (1995): «Modelling of Intonation for Speech Synthesis», *Tesis doctoral.* Boston University. School of Engineering.

Sosa, J. M. (1991): «Fonética y fonología de la entonación del español hispanoamericano», *Tesis doctoral*. Univ. de Massachussets..

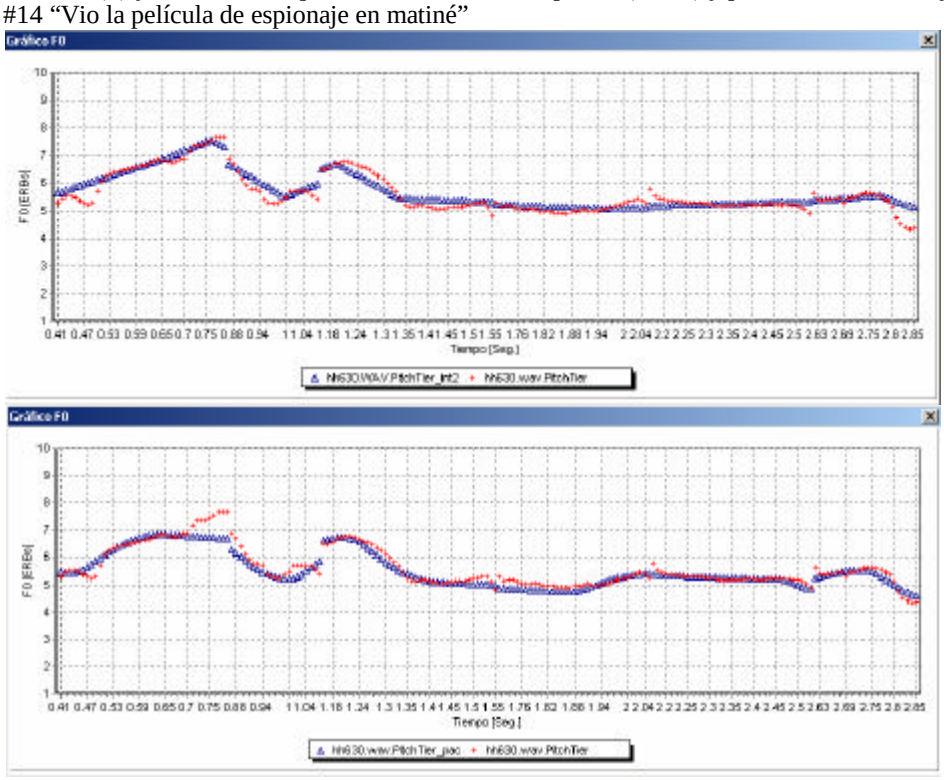
Sosa, Juan Manuel. (1999): *La entonación del español: su estructura fónica, variabilidad y dialectología*, Madrid, Cátedra.

Talkin, D. (1995): «A Robust Algorithm for Pitch Tracking (RAPT)», en Kleijn, W. B. y Paliwal, K. K. (eds.): *Speech Coding and Synthesis.* New York, Elsevier.

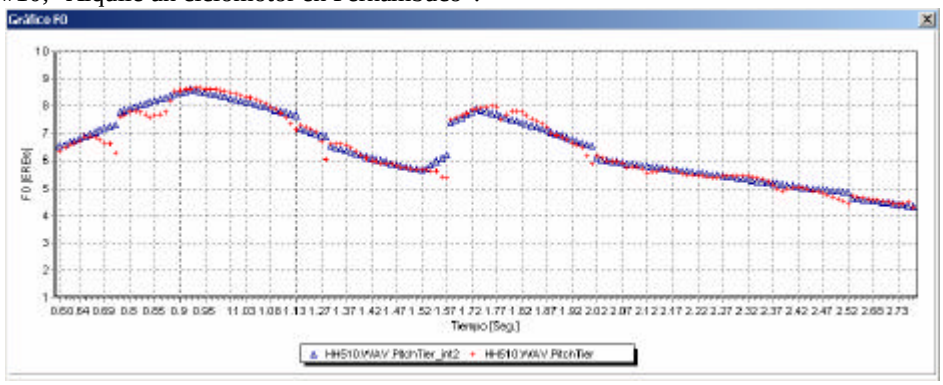
Toledo, G.A.(2000): «Taxonomía Tonal en español», *Language design*, 3.

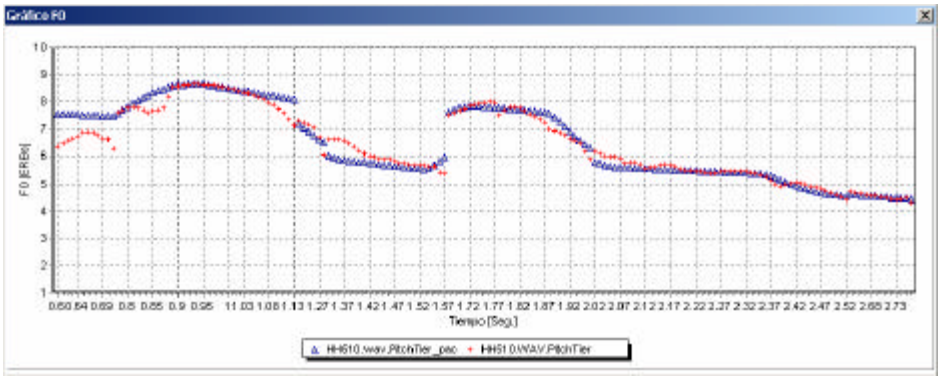
**ANEXO 1**

Contornos de F0 real (+) y estimados ( ) por el método ToBI Ampliado (arriba) y por el modelo de Fujisaki (abajo) para la frase #14 "Vio la película de espionaje en matiné"

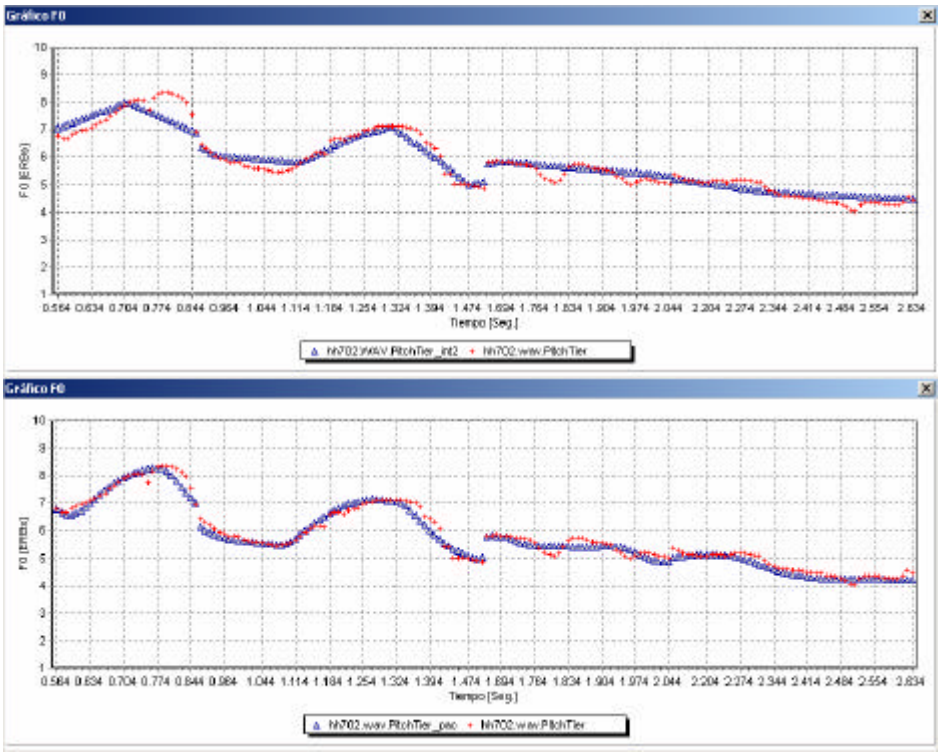



Contornos de F0 real (+) y estimados ( ) por el método ToBI Ampliado (arriba) y por el modelo de Fujisaki (abajo) para la frase #10, "Alquilé un ciclomotor en Pernambuco".






Contornos de F0 real (+) y estimados ( ) por el método ToBI Ampliado (arriba) y por el modelo de Fujisaki (abajo) para la frase #18 “Quiere comer la rosquilla desarmada”.



**Anexo 2.**
**Alineación temporal en el método de TOBI AMPLIADO**

Una vez percibido el acento tonal se ubica la sílaba asociada y se marca en el instante temporal correspondiente al valor más alto o más bajo de F0, luego se lee el valor en ERBs para ese instante y se define la cabeza del acento tonal H* o L*. A continuación se marca lo que ocurre con el contorno de F0 alrededor de esa marca temporal. Para ello, se observa cómo precede y continúa el contorno de F0, marcando las singularidades definidas por el valor de F0 en ERBs y el número de sílabas.

La indicación temporal de los tonos complementarios está dada por el número de sílabas a la derecha o izquierda de la cabeza de acento. En este trabajo se ha considerado que el tono complementario se ubica siempre en el centro de la vocal de la sílaba indicada en la etiqueta. Se ha demostrado que otras posiciones alternativas (inicio y final de la vocal) no reflejan cambios significativos en el error final.

Como ejemplo se analiza la frase "¿podemos encontrarnos a las ocho?". En la Figura 1 a la izquierda, luego de percibir el acento prenuclear alto H* (cabeza de acento) en la sílaba /De/ se debe decidir cómo completar la descripción del acento entre las siguientes alternativas: L+H*, H*+H ó H*. Las guías de etiquetado ToBI sugieren usar el acento L+H* cuando el tono bajo no puede ser predicho de un acento bajo precedente. En este ejemplo el tono precedente es un tono de juntura bajo L%, por lo que el acento queda mejor definido con el tono que continúa a la derecha: 7.11H*+7.76H1.
En la figura a la derecha, se percibe un acento nuclear alto H, al observar el acento precedente se observa una inflexión tonal no marcada previamente y en una pendiente de subida. En este caso se indica un acento 4.20H0+6.41H*. La terminación de este acento queda definida por el acento de juntura final.

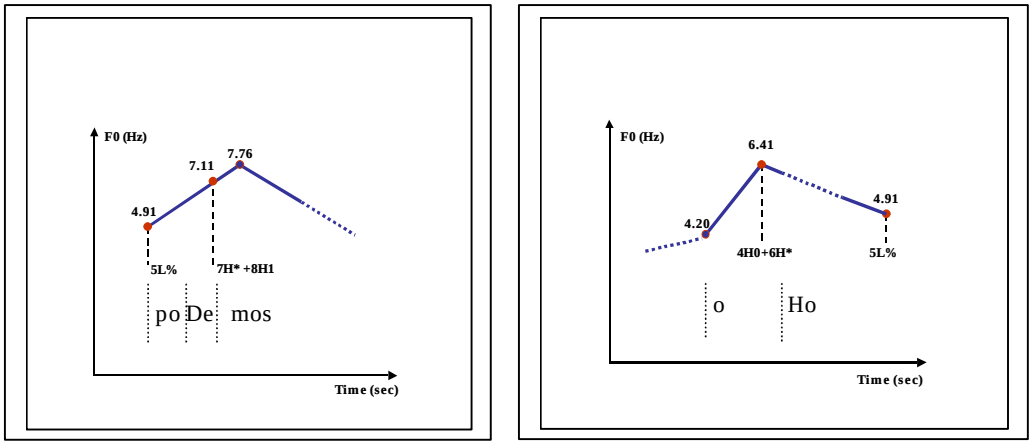



***Figura 1.*** *Ejemplos del etiquetado con ToBI Ampliado para la frase "podemos encontrarnos a las ocho?". Acento prenuclear al inicio del grupo entonativo (esquema a la izquierda) y acento nuclear al final del grupo entonativo (esquema a la derecha).*