MODELADO Y ESTIMACIÓN DE LA PROSODIA MEDIANTE RAZONAMIENTO BASADO EN CASOS

Ignasi Iriondo , J. Claudi Socoró, Lluís Formiga, Xavier Gonzalvo, Francesc Alías y Pere Miralles *

Enginyeria i Arquitectura La Salle. Universitat Ramon Llull.
Departamento de Comunicaciones y Teoría de la Senyal
Ps. Bonanova 8, 08022 Barcelona

{iriondo; jclaudi; gonzalvo; llformiga; falias; is08388}@salle.url.edu

RESUMEN

En este artículo se presenta la utilización del aprendizaje analógico, en particular el razonamiento basado en casos, como herramienta de generación automática de la prosodia a partir de texto, el cual ha sido etiquetado de forma automática con atributos prosódicos. Se trata de un método basado en corpus para el modelado cuantitativo de la prosodia y su estimación en un sistema de conversión del texto en habla. El principal objetivo es conseguir un método común para predecir los 3 rasgos prosódicos principales: la curva de frecuencia fundamental (F0), la duración segmental y la intensidad. Se ha llevado a cabo una evaluación objetiva y subjetiva para considerar su uso en el ámbito de la síntesis del habla expresiva.

1. INTRODUCCIÓN

La conversión texto en habla (CTH) tiene como finalidad la transformación automática de cualquier texto escrito en la correspondiente realización sonora. La investigación en este campo se centró inicialmente en conseguir el mayor grado de inteligibilidad posible. Posteriormente el objetivo ha sido conseguir una mayor naturalidad, es decir en la capacidad de emular la riqueza del habla humana que es intrínsecamente expresiva, ya que posee la capacidad de complementar la información verbal con una intención, actitud o estado emocional determinados. En este contexto, la mejora de la expresividad de los sistemas CTH se ha debido a avances en el modelado de la prosodia y la generación de la señal de voz de una alta calidad.

El estudio de la correlación entre el habla y la expresividad es complejo y se ha abordado desde diferentes enfoques (ver estudio comparativo en [1]). Algunos trabajos se centran en estudiar las variaciones en el timbre de voz, mientras que otros se basan en un uso para-lingüístico de la prosodia. Los estudios orientados a la síntesis del habla expresiva suelen abordar los dos enfoques, como es el caso del *Affect Editor* [2] o el trabajo de Montero et al. [3]. Además hay que tener en cuenta que en la producción de sonido humana se mezclan aspectos verbales y no verbales fruto de la evolución. La incorporación de este tipo de sonidos (*affect bursts*) se ha demostrado clave en la mejora de la expresividad del habla sintética [4].

La síntesis del habla expresiva a partir de texto requiere la generación automática de una prosodia adecuada al estilo/emoción deseado y un módulo de síntesis de voz capaz de generar habla de alta calidad con altas variaciones de los rasgos prosódicos y de los parámetros de cualidad de voz relacionados con el timbre. En la actualidad, la técnica predominante es la síntesis del habla basada en corpus o selección de unidades [5].

Este artículo presenta la utilización del Razonamiento basado en casos (CBR) aplicado a la generación automática de parámetros prosódicos (contorno de la frecuencia fundamental (F0), duración e intensidad de los fonemas) de habla expresiva en español. Por lo tanto, este trabajo está enfocado principalmente al modelado y estimación de la prosodia del habla expresiva. Con este propósito, se ha desarrollado un corpus de habla expresiva, el cual se utiliza también en el proceso de síntesis concatenativa del habla. Las medidas objetivas utilizadas para evaluar la precisión del modelado prosódico son el error cuadrático medio (RMSE) y el coeficiente de correlación (ρ). La evaluación subjetiva se ha llevado a cabo mediante un test perceptual utilizando la escala MOS. Las locuciones del test se han generado mediante la resíntesis de frases del corpus que no han formado parte del entrenamiento, a las que se les han modificado los parámetros prosódicos según la predicción obtenida automáticamente a partir de su texto.

Este artículo se organiza como sigue. La sección 2 explica el modelo prosódico, detallándose los atributos prosódicos extraídos del texto de entrada y los parámetros prosódicos a predecir. La sección 3 explica la adaptación de la técnica CBR al modelado prosódico, entrando en detalle de nuestra propuesta en la sección 4. En la sección 5 se muestran los resultados de los experimentos realizados. Y finalmente, la sección 6 proporciona las conclusiones y el trabajo futuro.

* Este trabajo ha sido parcialmente financiado por la Comisión Europea, proyecto SALERO FP6 IST-4-027122-IP.

2. MODELADO Y ESTIMACIÓN DE LA PROSODIA

Los parámetros que determinan la prosodia de un texto hablado son esencialmente la duración e intensidad segmental, el posicionamiento y duración de las pausas y el contorno de F0 [6]. En el ámbito de los sistemas CTH, la literatura en modelado prosódico es muy extensa. La curva de entonación es el parámetro prosódico con una mayor presencia, distinguiéndose entre métodos cuantitativos (como TILT [7], Fujisaki [8] o Bezier [9]) y métodos cualitativos (ToBI [10] o Intsint [11]). Para el modelado de la duración segmental se han utilizado métodos basados en reglas [12], o métodos estadísticos tales como redes neuronales [13] o árboles de clasificación y regresión (CART) [14]. El modelado de la intensidad es el menos presente en la literatura aunque hay algunos trabajos específicos en esta dirección tales como [15]. En nuestro trabajo previo [16], se confirmó la importancia de este parámetro en el modelado del habla para ciertas emociones.

En el habla natural, la duración de los sonidos depende del contexto en que se encuentran. La mayoría de estudios (por ejemplo [17] [18] [19]), utilizan el fonema como unidad básica para la duración, aunque haya aproximaciones basadas en difonemas o sílabas. Según estos estudios, los factores que influyen en la duración de los sonidos se deben básicamente a: *i)* la identidad del fonema y los de su contexto (habitualmente, el anterior y el posterior), *ii)* información sobre el acento, *iii)* información sobre la posición del fonema en la frase y en la sílaba. Cada estudio presenta su manera particular de codificar esta información.

La predicción de la curva de intensidad se suele llevar a cabo generalmente a nivel de fonema. Aunque muchos sistemas CTH no consideran este rasgo, los factores a tener en cuenta [6] están también relacionados con la identidad del fonema, el acento y la posición.

En [9] se realiza un estudio de diversos trabajos a cerca de las unidades de entonación y los factores que caracterizan cada una de estas unidades. Existen diferentes tipos de unidades utilizadas para modelar el contorno de entonación: las unidades inferiores a la sílaba y la sílaba (microentonación), el grupo acentual (GA) —relacionado con el ritmo del habla—, el grupo de entonación (GE) y otras unidades superiores (planificación del discurso). Dicho autor propone el uso del GA como unidad básica para el modelado de la entonación. Además, se concluye que algunos de los factores a considerar a nivel de grupo acentual para modelar la entonación estan relacionados con: *i)* el tipo de GE al que pertenece el GA, *ii)* la posición del GA dentro del GE, *iii)* la posición del acento, *iv)* la posición del GE dentro de la frase, y *v)* el número de sílabas del GA y del GE. La curva de entonación de cada GA se puede modelar con diferentes funciones (polinomios, Bezier [9], logarítmicas).

3. CBR APLICADO AL MODELADO DE LA PROSODIA

El aprendizaje artificial (ML) comprende un conjunto de técnicas que permiten reconocer una situación problemática y reaccionar utilizando la estrategia aprendida para un nuevo problema. La utilización de ML puede ser interesante en aquellos dominios en que la experiencia es escasa y la codificación del conocimiento que la describe es limitada o fragmentaria y por lo tanto incompleta. La predicción de los rasgos prosódicos a partir del texto es una tarea compleja en la cual intervienen muchos elementos (lingüísticos, fonéticos, pragmáticos). La utilización de técnicas de ML para dicha tarea puede deparar resultados válidos dentro del ámbito de los sistemas CTH. El proceso de ML se puede ver como la suma de dos fases: selección más adaptación. En una primera fase el sistema escoje (selecciona) las características más relevantes de un objeto, las compara con otras de conocidas a través de algun sistema de comparación, y cuando las diferencias son significativas, adapta el modelo de aquel objeto según el resultado de la comparación. Según sean estas dos fases los sistemas de ML se clasifican en Aprendizaje Analógico, Inductivo, Evolutivo y Conexionista. Las técnicas más utilizadas en el modelado prosódico pertenecen al aprendizaje inductivo, tales como árboles de decisión [14], y al aprendizaje conexionista, tales como Redes Neuronales [13].

En este trabajo, presentamos una aplicación del Razonamiento basado en casos (CBR) al modelado de la prosodia, ya que esta técnica de aprendizaje analógico permite un tratamiento sencillo de atributos de diferente naturaleza. El ciclo principal del CBR puede descomponerse en cuatro tareas (Ciclo 4R): recuperar los casos mas similares (*retrieve*), reutilizarlos para resolver el problema (*reuse*), revisar la solución propuesta (*revise*) y aprender de la experiencia (*retain*) [20] (figura 1). A continuación, se detalla la adaptación de estos pasos al modelado prosódico basado en corpus.

La inicialización del sistema no es propiamente una tarea del ciclo 4R del CBR pero será imprescindible para conseguir una memoria de casos que sea un equilibrio entre representatividad y compactación. En primer lugar hay que identificar los atributos que definen los casos para cada uno de las tres tareas del sistema: predicción de la duración y de la energía de los fonemas y de la curva de entonación de los GAs. Estos atributos y su naturaleza son los que se muestran en el apartado 4. Posteriormente, se genera el conjunto de entrenamiento mediante la unión de los rasgos prosódicos del corpus con los atributos prosódicos extraídos del análisis lingüístico del texto. La reducción de casos se consigue mediante el agrupamiento de casos representados por los mismos atributos. En el caso de reducir a un único caso cada grupo, el parámetro prosódico será la media de todos los casos del grupo. En cambio, si un conjunto de casos con los mismos atributos se divide en k subconjuntos, se hace necesario un proce-

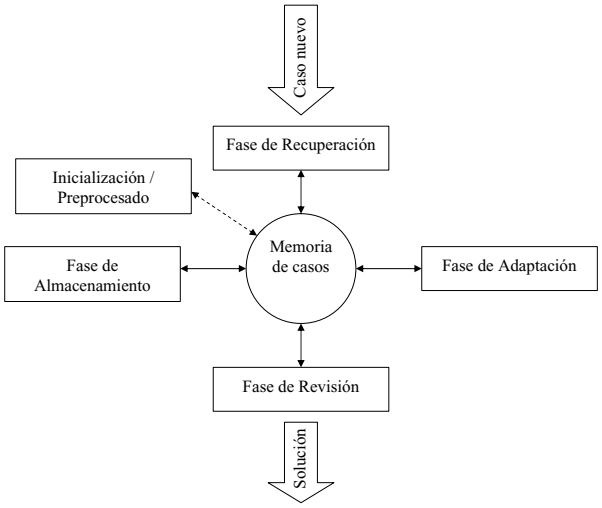

Figura 1. *Ciclo 4R del CBR.*

dimiento para seleccionar el mejor candidato en función del contexto.

El objetivo de la tarea de recuperación es mapear la solución desde la memoria de casos al nuevo problema. Se recupera el caso (o los *k* casos) más similar utilizando una métrica adecuada a los atributos que lo representan.

La tarea de reutilización trata de solventar un nuevo caso a partir de la información almacenada en la memoria de casos. En primer lugar se predice la duración de los fonemas, ya que la recuperación de la curva de F0 se hace sobre un eje temporal normalizado (ver figura 2). De esta forma, una vez conocida la duración de cada fonema se expande el eje temporal y se asocia el valor de F0 de cada fonema según el polinomio recuperado. En el caso de recuperar más de una muestra ($k > 1$) hay que realizar la adaptación para obtener únicamente una solución. Para el caso de la entonación, proponemos solventar este proceso decisional mediante la búsqueda de un camino óptimo que maximice la continuidad de F0 entre GAs.

En esta implementación, no se requiere revisión de la solución propuesta. El almacenamiento se realiza únicamente en la fase de inicialización. Por lo tanto el sistema no tiene la posibilidad de añadir nuevos casos en la fase de explotación.

4. ENFOQUE

La extracción de atributos prosódicos a partir del texto se realiza de forma automática mediante nuestra herramienta de análisis lingüístico que proporciona la transcripción fonética del texto y lo marca en grupos de entonación (GE), grupos acentuales (GA), palabras y sílabas. El GE se define como una estructura coherente de entonación que no incluye ninguna ruptura prosódica importante. Las rupturas prosódicas se producen debido a las pausas o inflexiones significativas del contorno de F0. Por el momento, únicamente consideramos las rupturas definidas por los signos de puntuación. El GA se define como una palabra acentuada precedida, si es el caso, por una o más palabras no acentuadas.

Para su implementación práctica, finalmente y despues de provar diferentes configuraciones, se han utilizado los atributos mostrados en la tabla 1 para la duración, intensidad y F0 respectivamente. Esta tabla muestra la etiqueta utilizada, una breve descripción y el tipo de atributo[1].

Para el modelado de la duración segmental y de la intensidad, hemos escogido el fonema como unidad acústica básica. La duración de un fonema depende principalmente de su identidad y del contexto donde se encuentra. Para la intensidad se han utilizado atributos parecidos (ver tabla 1).

Tabla 1. *Características prosódicas para duración, energía y F0.*

Etiqueta	Atributo	Tipo
F0	Fonema anterior	D
F1	Fonema actual	D
F2	Fonema siguiente	D
ACENTUADO	Fonema acentuado	B
GA-en-GE	Posición de GA en GE	D
FON-en-GE	Posición de FON en GE	D
DURACION	Duración del fonema en ms	N
Etiqueta	Atributo	Tipo
F1	Fonema actual	D
ACENTUADO	Fonema acentuado	B
GA-en-GE	Posición de GA en GE	D
FON-en-GA	Posición de FON en GA	D
FON-en-GE	Posición de FON en GE	D
INTENSIDAD	Intensidad *rms*	N
Etiqueta	Atributo	Tipo
TIPO-GE:	Tipo of GE	D
GA-en-GE	Posición de GA en GE	D
ACENTO	Posición de la sílaba tónica	D
GA-en-FRA	Posición del GA en la frase	D
NUM-SIL	Número de sílabas del GA	N
$a_0, a_1, \ldots a_n$	Coeficientes del polinomio	A

Para el modelado de la curva de F0 hemos elegido el GA como unidad básica siguiendo la propuesta de Escudero en [9]. El GA incorpora la influencia de la sílaba (cada GA está compuesto de una sílaba tónica y el resto átonas) y la estructura a nivel de GE se consigue mediante la concatenación de GAs. Por contra, este modelo carece de variaciones debidas a microentonación. Por el momento, sólo diferenciamos entre GEs enunciativos, interrogativos o exclamativos, que son fácilmente detectables a partir de los signos de puntuación. El atributo ACENTO indica la posición de la la sílaba tónica en el GA y el número de sílabas está relacionado con la longitud del GA (ver Tabla 1).

[1](D) Discreto, (B) Binario, (N) Numérico, (A) Array numérico

Para la curva de entonación de cada GA se ha utilizado una representación cuantitativa mediante los coeficientes de un polinomio aproximador de grado n (ecuación 1). Para encontrar los coeficientes del polinomio se parte de una colección de puntos (x_i, y_i) que representan el valor de la F0 media de cada fonema. Este valor de F0 media está referenciado al instante central del fonema. Mediante el método de mínimos cuadrados se calcula el polinomio aproximador de grado n que minimiza el error dado en la fórmula 2.

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \qquad (1)$$

$$E = \sum_{i=0}^{m} \left(P(x_i) - y_i\right)^2 \qquad (2)$$

El eje temporal se normaliza entre 0 y 1 para todos los GAs, de forma que el instante 0 representa el inicio del primer fonema del GA y el instante 1 el final del último fonema del GA. Se ha estudiado también la posibilidad de incluir en el cálculo de los coeficientes del polinomio información contextual, es decir incluir los puntos correspondientes a los dos fonemas adyacentes de los GAs vecinos (ver figura 2).

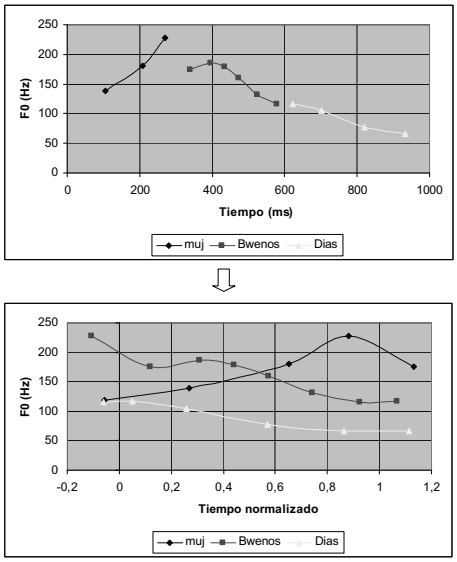

Figura 2. *Proceso de normalización del eje temporal de la curva de F0.*

5. EXPERIMENTOS Y RESULTADOS

5.1. Corpus

Para la realización de los experimentos hemos utilizado un corpus de habla desarrollado conjuntamente con el LAICOM del Departamento de Publicidad y Comunicación de la Universidad Autónoma de Barcelona. Un conjunto de textos extraídos de una base de datos de publicidad se han grabado por una locutora profesional en nuestro estudio de grabación. Los textos ya estaban clasificados por temáticas y se ha asociado un estilo expresivo a cada categoría. Los tres estilos iniciales son el neutro, el alegre y el sensual, que son los que se han realizado los experimentos. Recientemente, el corpus se ha ampliado con dos estilos más, agresivo y triste, que están en fase de segmentación y etiquetado.

5.2. Evaluación objetiva

El funcionamiento del sistema se ha evaluado mediante técnicas objetivas utilizando el error cuadrático medio (RMSE) y el coeficiente de correlación de Pearson (ρ). Los corpus de voz se han dividido en un 75 % para entrenar el sistema y un 25 % para evaluación. El RMSE mide la diferencia entre los valores observados y los predecidos en términos cuadráticos. El coeficiente de correlación mide el grado de dependencia lineal entre dos muestras de datos. Se calculan dichas medidas sobre las frases que forman el conjunto de test.

Los resultados para los tres estilos (neutro, alegre, sensual) se muestran en la Tabla 2. A nivel de entonación, el estilo sensual es el que presenta un mejor resultado debido a que es el estilo que presenta menos inflexiones. En cambio, presenta peores resultados en cuanto a la duración debido a que presenta segmentos de voz aspirada que tienden a alargarse. El estilo alegre es el que presenta mayores variaciones en la entonación, siendo el que obtiene un error mayor.

La comparación de resultados con otros trabajos se hace difícil al utilizar un corpus diferente. De todas formas, los resultados obtenidos son parecidos a los publicados por otros autores, la cual cosa nos permite considerar que es un buen punto de partida para el modelado de la prosodia. Existe un amplio margen para el ajuste y perfeccionamiento en ciertas tareas del modelado prosódico hecho hasta el momento, que seguramente permitirán mejorar los resultados finales.

5.3. Evaluación subjetiva

El experimento subjetivo se ha realizado mediante un test de percepción que pretende comparar la similitud entre la prosodia natural y la sintética. Los sujetos tienen que juzgar el grado de parecido (5= Muy alto, 4= Alto, 3= Cierto parecido, 2=Poco, 1=Nada) en la prosodia de dos versiones de la misma frase. Se indica a los usuarios que se fijen principalmente en la similitud de la prosodia. La frase con la prosodia natural se genera mediante resíntesis con los valores de F0 media, intensidad y duración etiquetados en el corpus. La otra frase se genera mediante resíntesis de la frase a partir de los valores de prosodia estimados. Para la modificación de la señal utilizamos nuestro sintetizador de voz basado en selección de unidades que realiza un ajuste de F0, duración y intensidad mediante TD-PSOLA [21]. Ambas versiones de ca-

Tabla 2. *Resultados del experimento objetivo*

	Entonación (F0)		Duración		Intensidad			
Estilo	RMSE (Hz)	ρ	RMSE (msec)	ρ	RMSE (rms)	ρ	Frases de Test	GA por frase
Neutro	29,668	0,631	21,688	0,746	0,021	0,853	169	5,41
Alegre	73,157	0,520	23,698	0,635	0,024	0,810	124	5,33
Sensual	23,534	0,438	28,936	0,655	0,029	0,673	101	4,76

da frase se presentaron via interfaz web a 10 sujetos para su evaluación. Las frases se podían escuchar las veces que fuera necesario.

En la tabla 3 se muestra el porcentaje de similitud que han obtenido los 3 estilos. Destaca la correlación entre estos resultados y el RMSE obtenido para la entonación. En la figura 3 se muestra la puntuación por cada frase y cada estilo.

Tabla 3. *Porcentaje de similitud entre prosodia natural y sintética para los tres estilos*

	Neutro	Sensual	Alegre
Muy Alto	5,00 %	42,31 %	5,56 %
Alto	50,72 %	30,00 %	44,44 %
Cierto parecido	30,71 %	22,31 %	29,86 %
Poco	13,57 %	5,38 %	17,36 %
Nada	0,00 %	0,00 %	2,78 %

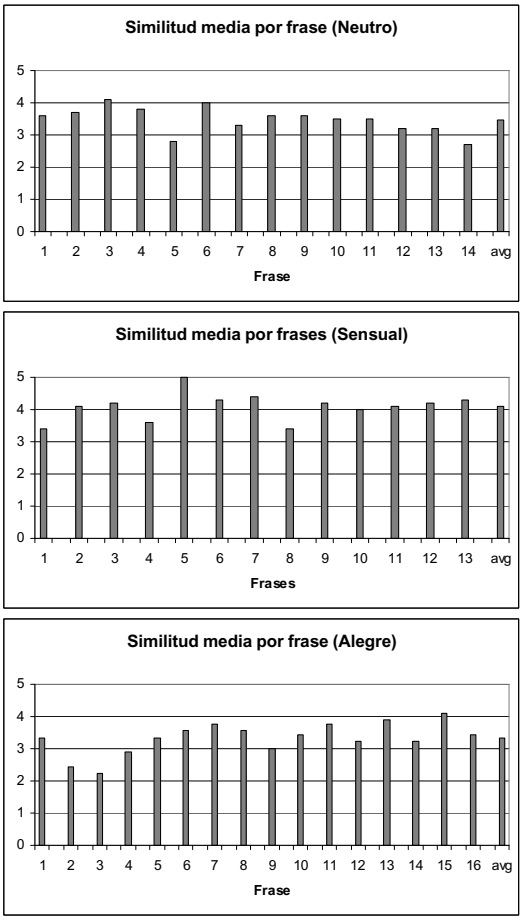

Figura 3. *Similitud media por frase y estilo (5=Muy Alta, 1=Nada).*

6. CONCLUSIÓN Y TRABAJO FUTURO

Este artículo presenta un nuevo enfoque para predecir la parámetros prosódicos desde texto en diferentes estilos expresivos. Se trata de la adaptación del CBR para el modelado cuantitativo de la prosodia (curva de frecuencia fundamental (F0), la duración segmental y la intensidad) para su uso en la conversión del texto en habla. Se ha llevado a cabo una evaluación objetiva y subjetiva para considerar su uso en el ámbito de la síntesis del habla expresiva. Con dicho fin se ha creado un corpus oral expresivo a partir de textos publicitarios en tres estilos (neutro, alegre y sensual) que se está ampliando los estilos agresivo y triste.

Como trabajo futuro, se pretende comparar esta técnica con otras técnicas de ML aplicadas sobre el mismo corpus, ampliar el estudio a los dos estilos nuevos y explorar el uso de nuevos atributos prosódicos y su relación con cada estilo.

7. BIBLIOGRAFÍA

[1] R.Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, y J. G. Taylor, "Emotion recognition in human computer interaction," *IEEE Signal Processing*, vol. 18, no. 1, pp. 33–80, January 2001.

[2] J. E. Cahn, "Generating expression in synthesized speech," M.S. thesis, Massachusetts Institute of Technology, 1989.

[3] Montero J.M., J. Gutiérrez Arriola, J. Colás, E. Enríquez, y J.M. Pardo, "Analysis and modelling of emotional speech in Spanish," in *Proceedings of 14th International Conference of Phonetic Sciences*, San Francisco, USA, 1999, pp. 957–960.

[4] M. Schröder, "Experimental study of affect bursts," *Speech Communication*, vol. 40, pp. 99–116, 2003.

[5] E. Eide, A. Aaron, y Cohen P. Donovan R. Hamza W. Mathes T. Picheny M. Polkosky M. Smith M. Viswanathan M Bakis, R., "Recent improvements

to the ibm trainable speech synthesis system," in *Proceedings of ICASSP'03*, Hong Kong, 2003.

[6] J. Llisterri, M. J. Machuca, C.de la Mota, M. Riera, y A. Ríos, *Entonación y tecnologías del habla*, Tecnologías del texto y del habla. Prieto, P. (Ed.) Teorías de la entonación. Ariel (Lingüística), Barcelona, 2003.

[7] Paul Taylor, "Analysis and Synthesis of Intonation using the Tilt Model," *Journal of Acoustical Society of America*, 2000.

[8] H. Fujisaki, S. Ohno, K.Ñakamura, M. Guirao, y J. Gurlekian, "Analysis of accent and intonation in Spanish based on a quantitative model," in *Proc. ICSLP*, 1994.

[9] D. Escudero, *Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversion Texto-Voz en Español*, Ph.D. thesis, Universidad de Valladolid, 2003.

[10] K. Silverman, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, y J. Hirschberg, "ToBI: A standard for labelling English prosody," in *Proceedings of ICSLP92*, 1992.

[11] D.J. Hirst, N. Ide, y Veronis J., "Coding fundamental frequency patterns for multilingual synthesis with INTSINT in the MULTEXT project," in *2nd ESCA/IEEE Workshop on Intonation*, 1994.

[12] D.H. Klatt, *Synthesis by rule of segmental durations in english sentences*, B. Lindblom and S. Öhman (Ed.), Frontiers of Speech Communication. New York: Academic, 1979.

[13] N.W. Campbell, "Analog I/O nets for syllable timing," *Speech Communication*, vol. 9, pp. 56–61, 1990.

[14] B. Möbius y J. van Santen, "Modelling segmental duration in German TTS synthesis," in *Proc. of ICSLP*, 1996.

[15] J. Trouvain, W. J. Barry, C.Ñielsen, y O. Andersen, "Implications of energy declinations for speech synthesis," in *3rd ESCA/ COCOSDA Workshop on Speech Synthesis, November*, Jenolan Caves, Australia, 1998, pp. 47–52.

[16] I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. Blanco, D. Bernadas, J. Oliver, D. Tena, y L. Longhi, "Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques," *Proc. of the I. W. on Speech and Emotion*, pp. 161–166, Sept. 2000.

[17] A. Febrer, J. Padrell, y A. Bonafonte, "Modeling Phone Duration: Application to Catalan TTS," in *3rd Int. Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

[18] Navas E., Hernáez I., y Sánchez J.M., "Modelo de duración para conversión texto a voz en euskera," *Procesamiento del Llenguaje Natural*, vol. 1, no. 3, 2002.

[19] J. P. Teixeira y D. Freitas, *Evaluation of a Segmental Durations Model for TTS*, Computational Processing of the Portuguese Language - 6th International Workshop, PROPOR 2003. N. Mamede, J. Baptista, I. Trancoso, M.G. Nunes (Eds), Edited by Springer, 2003.

[20] A. Aamodt y E. Plaza, "Case-based reasoning: foundational issues, methodological variations, and system approaches," *AI Commun.*, vol. 7, no. 1, pp. 39–59, 1994.

[21] E. Moulines y F. Charpentier, "Pitch-synchronous waveform processing techniques for TTS synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.