

## MODELO DE SÍNTESIS DE HABLA CON DISFLUENCIAS BASADO EN MODIFICACIONES LOCALES SOBRE FRASES CONSTITUYENTES

Jordi Adell<sup>1</sup>, David Escudero-Mancebo<sup>2</sup> y Antonio Bonafonte<sup>1</sup>

<sup>1</sup>TALP Research Center, Universitat Politècnica de Catalunya

<sup>2</sup>ECA-SIMM Laboratory, Universidad de Valladolid

{antonio,jadell}@gps.tsc.upc.edu<sup>1</sup>, descuder@infor.uva.es<sup>2</sup>

### RESUMEN

La síntesis de habla con disfluencias será necesaria en aplicaciones futuras, como puede ser el doblaje automático o la traducción voz-voz. Esta comunicación presenta un modelo para la generación de habla sintética con disfluencias, que se basa en la inserción de disfluencias dentro de locuciones fluidas. Las frases con disfluencias se generan empleando los modelos prosódicos que se emplearían para generar las frases sin disfluencias, pero añadiendo modificaciones locales que afectan sólo a las unidades adyacentes a la posición que ocupa la disfluencia. En esta comunicación se explica el modelo propuesto y su aplicación para modelar las variaciones locales relativas a la duración segmental en las fronteras de las disfluencias.

### 1. INTRODUCCIÓN

Hoy en día, los sistemas de síntesis de voz han alcanzado un alto nivel de naturalidad [1], principalmente debido al uso de técnicas como la síntesis basada en selección de unidades u otras tecnologías [2] que se basan en el análisis de grandes corpus de voz. Por ahora, la principal aplicación de la síntesis de voz está centrada en el habla leída, dado que se trata de un estilo muy generalista cuya extensión a otras situaciones se considera realista. Pero hoy en día, e incluso más aún en el futuro, las aplicaciones de conversión texto-voz (CTV) como por ejemplo el doblaje automático de películas, la robótica, los sistemas de diálogo o los informativos multilingües; demandan de una riqueza en los estilos muy superior.

Para integrar la voz sintética en las tecnologías enumeradas en el párrafo anterior, los sistemas CTV deben simular la manera de hablar, en lugar de simular la manera de leer, de los humanos. Ambos estilos difieren significativamente debido a la inclusión de un buen número de factores prosódicos, uno de ellos la presencia de disfluencias. Las disfluencias se definen como una interrupción en el flujo del habla que no añade ningún contenido proposicional a la frase [3]. A pesar de ello, las disfluencias ofrecen indicaciones sobre lo que se está diciendo [4] y son tremendamente frecuentes en el habla espontánea [3]. Debido a esto, su inclusión en el habla sintética parece una necesidad clara para el futuro.

El estudio de las disfluencias ha sido realizado desde diferentes perspectivas, principalmente la fonética [5], la psicolingüística [6] y el reconocimiento del habla [7]. Estas diferentes perspectivas modelan las disfluencias teniendo en cuenta sus intereses específicos. El uso de las disfluencias en sistemas CTV conlleva consideraciones adicionales que fuerzan la propuesta de un modelo alternativo. El modelo que proponemos, a diferencia de

otras aproximaciones ya empleadas en sistemas CTV como [8] o [9], tiene en cuenta las frases fluidas asociadas a la frase disfluente que va a ser sintetizada, teniendo en cuenta las modificaciones locales que produce la inserción de dicha disfluencia. Estas modificaciones locales pueden afectar a la prosodia o a la calidad de la locución original. En esta comunicación mostramos la importancia de estas modificaciones locales, comprobando el impacto en la duración de las sílabas que rodean las disfluencias.

Primero hacemos una introducción del modelo de generación de disfluencias. Después, se presenta el procedimiento a seguir para aplicar el modelo, mostrando las alteraciones en la duración de las sílabas que rodean la disfluencia. Por último se plantea el trabajo futuro a realizar para completar este trabajo y las conclusiones.

### 2. INSERCIÓN DE DISFLUENCIAS EN SISTEMAS DE SÍNTESIS POR SELECCIÓN DE UNIDADES

La síntesis de disfluencias en el marco de la síntesis por selección de unidades presenta una serie de dificultades a tener en cuenta. Primero, la mayoría de los sistemas de selección de unidades que existen tienen un inventario cerrado de unidades que no contiene en absoluto disfluencias. Esto hace que los métodos de aprendizaje automático que se aplican para analizar la prosodia, no sean capaces de modelar automáticamente este fenómeno a partir de los datos. Además, no sólo los modelos prosódicos sino también los modelos de análisis del texto, como por ejemplo el etiquetado de *Part-of-Speech*, dependen mucho de que las frases de entrada tengan una estructura que se corresponda con una sintaxis correcta y un orden, en términos de acentos y de grupos de entonación, también correcto; lo cual no sucede en una frase con disfluencias, ya que cuando la fluidez de una frase se rompe, su estructura también se rompe. Por último, el habla sintética con disfluencias precisa del uso de nuevas unidades segmentales que no están definidas en las bases de datos convencionales, como pueden ser los fillers o los fonemas interrumpidos.

Nuestro modelo distingue tres elementos de cara a generar una frase dada que incluya disfluencias (*Disfluent Sentence* (DS)). Primero, la frase original que iba a ser pronunciada antes de que apareciera la disfluencia (*Original Sentence* (OS)). Después la frase objetivo (*Target Sentence* (TS)) que hubiera sido dicha si no hubiera habido ningún motivo que provocara la disfluencia. Tercero, el *Editing Term* (ET), que de acuerdo a la terminología defendida en [10] es la clave o indicador de la disfluencia (por ejemplo, el relleno de la pausa). Podemos ilustrar estos términos con un ejemplo tomado de [11]: *Go from left to mmm from pink again to blue* donde los elementos de la disfluencia se identifican como:

*Go* RM[from left to]  $\downarrow$  <sup>IP</sup> ET[mmm], RR[from pink again to] *blue*.

Trabajo parcialmente financiado por los proyectos de investigación AVIVO del Gobierno de España (TEC2006-13694-C03) y por el proyecto ACME de la Junta de Castilla y León (VA077A08)

siendo *RM* (*Reparandum*), *RR* (*Repair*), *ET* (*Editing Term*) e *IP* (*Interruption Point*) los términos empleados en [10]. La frase del ejemplo (DS) está relacionada con las frases: *Go from left to right* y *Go from pink again to blue* que son las frases OS y TS respectivamente; y el ET es *mmm*. Estas relaciones se modelan como sigue:

$$\begin{aligned}
 DS &= PrevRM|RM|ET|RR|PostRR \\
 OS &= PrevRM|RM|PostRM \\
 TS &= PrevRM|RR|PostRR
 \end{aligned}$$

donde *PrevRM* es la parte de la frase que precede el *RM*; *PostRM* es la parte que sigue al *RM* y *PostRR* es la parte que sigue al *RR*. El *PostRM* existe sólo en OS, porque no se trata de una parte real de ninguna locución, sino que en su lugar, en la DS, se pronuncia *RR*. En el ejemplo presentado arriba, *PostRM* sería *right*. Dado una DS, un sistema CTV de selección de unidades puede generar correctamente las correspondientes OS y TS. Si el inventario de unidades del CTV incluye un cierto número de frases con disfluencias, entonces el sistema CTV podrá elegir entre un conjunto de ETs. Además, existen evidencias de que la inclusión de disfluencias provoca modificaciones locales en las propiedades acústicas de los términos *RM*, *PostRM* y *PrevRM* [12, 9].

Nuestra propuesta de generación de disfluencias opera en tres etapas (Figura 1). Primero, utiliza OS para obtener los parámetros prosódicos relativos a *PrevRM* y *RM*. También genera TS para obtener los parámetros prosódicos de *RR* y *PostRR*. Estos parámetros prosódicos se utilizan para guiar la búsqueda en el inventario de síntesis. En una segunda etapa, se obtiene el término ET desde el inventario. Finalmente, se aplican modificaciones locales a las sílabas adyacentes al término *ET*. Estas modificaciones se corresponden con las desviaciones locales que pueden aparecer en las fronteras de los elementos descritos en esta sección (*PrevRM*, *RM*, ...).

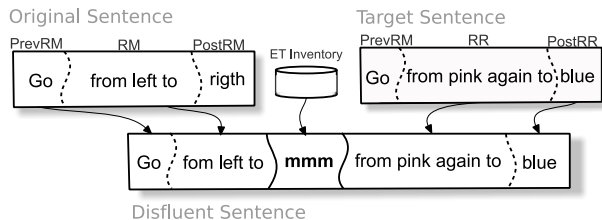


Figura 1: Proceso de generación de habla con disfluencias aplicado a una frase de ejemplo

El ritmo es una de las variables prosódicas que puede ser desviada con respecto a los valores predichos para las locuciones sin disfluencias. En la siguiente sección, vamos a mostrar como el ritmo de las frases con disfluencias sigue una tendencia general que es similar a la seguida por las frases sin disfluencias y que sufre una desviación significativa en las unidades próximas al *ET*. En este trabajo se consideran tres tipos de disfluencias: alargamientos, repeticiones y pausas rellenas. La tabla 1 describe los elementos de modelado para estos tres casos de estudio.

Tabla 1: Elementos del modelo para diferentes tipos de disfluencias. 1ª y 2ª indican cada una de las realizaciones en una repetición.

type	RM	ET	RR
alargamientos	∅	∅	∅
pausas rellenas	∅	filler (e.g. mmm)	∅
repeticiones	1ª	∅	2ª

### 3. MODIFICACIONES LOCALES DEL RITMO

En la literatura se encuentran dos categorías principales a la hora de describir el ritmo de las distintas lenguas: *Accent-timed* y *syllable-timed*. Ambas categorías están relacionadas con el principio de isocronía por el cual, las unidades del lenguaje tienden a estar equiespaciadas en el tiempo [13]. El español, al igual que otras lenguas procedentes del latín, se considera *syllable-timed* [14, 15]. Dado que en este trabajo vamos a centrarnos en el caso del español, parece apropiado medir el ritmo de las unidades suprasegmentales (como por ejemplo las frases) como la duración media de las sílabas en dichas unidades. Sirva la tabla 2 como referencia de los valores de duración media de las sílabas medidas para todo el corpus.

Tabla 2: Duración media de las sílabas y límites de los intervalos de confianza al 99%

	mean	lower bound	upper bound
no-acentuadas	105ms	102ms	108ms
acentuadas	136ms	132ms	140ms
pre-Pausal	222ms	210ms	236ms
todas	123ms	120ms	125ms

El corpus empleado en este trabajo es una selección de frases del corpus desarrollado para el proyecto europeo LCSTAR. Se grabó en un laboratorio y recoge diálogos de dos personas a las que se pidió que completaran una determinada tarea por teléfono. La comunicación fue semi-duplex de manera que la base de datos está grabada en base a los turnos de intervención [16]. Aunque es habla de laboratorio, es espontánea porque los locutores no tenían ninguna guía en sus intervenciones. Los hablantes pronuncian las disfluencias de manera natural y además las disfluencias son muy frecuentes porque necesitaban planificar los turnos de intervención a la vez que realizaban sus tareas. Se han utilizado 100 frases de cuatro hablantes diferentes (3 hombres y 1 mujer). En total las disfluencias que se han analizado son: 133 pausas rellenas, 71 repeticiones y 65 alargamientos. La segmentación fonética se realizó automáticamente y fue corregida manualmente.

Se computan una serie de medidas para cada elemento descrito en la sección 2 (Tabla 3).  $D_{syl}^{-1}$  es la última sílaba de *PrevRM* y  $D_{syl}^1$  es la primera sílaba de *PostRR*. Esperamos encontrar que el ritmo permanece constante a lo largo de las frases y que los cambios se producen en las fronteras de *ET* de acuerdo al modelo propuesto.

Tabla 3: Lista de variables relacionadas con el ritmo.

Variable	Definición
$\bar{R}_s$	Duración media de las sílabas de una frase.
$R_w^N$	Duración media de las sílabas en la N-ésima palabra desde la disfluencia.
$R_{DF}$	Duración media de las sílabas en la disfluencia, o una parte de ella: $R_{hes}$ , $R_{RM}$ , $R_{RR}$ , $R_f$
$D_{syl}^N$	Duración de la N-ésima sílaba desde la disfluencia.

El estudio se basa en diagramas tipo *boxplot* [17] que mostrarán las desviaciones en las fronteras de los *ET*. Para facilitar la interpretación, las variables se trazan en el orden en el que aparecen en las frases. Hemos utilizado el test estadístico conocido como *Least Significant Difference* (LSD) o *t-test* múltiple para comparar las medias de las distribuciones y así establecer diferencias [18]. Todos los niveles de significatividad presentados en esta comunicación son al 99% de confianza.

Los alargamientos (en inglés *hesitations*) consisten en una extensión no prevista de una sílaba. En la figura 2 se observa que

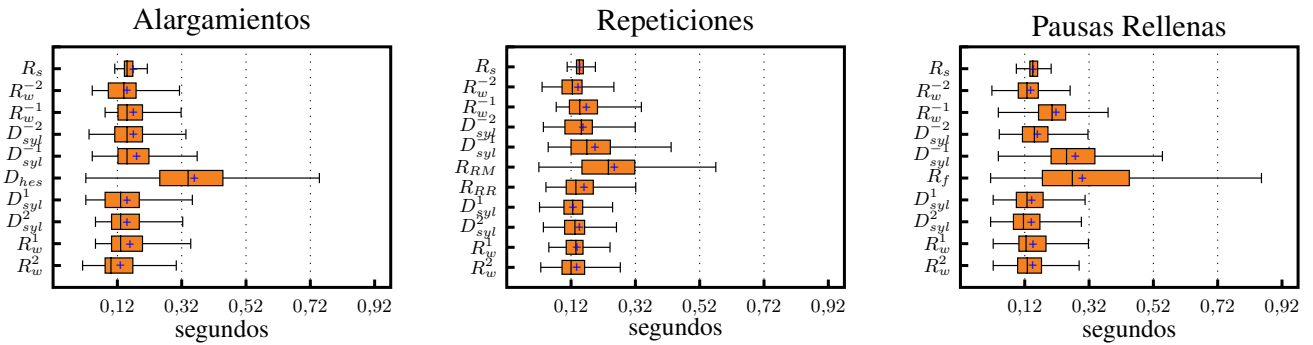


Figura 2: Boxplot de variables relacionadas con el ritmo para cada tipo de disfluencia

no hay una diferencia significativa entre el ritmo en las palabras que preceden al alargamiento ( $R_w^{-1}$  y  $R_w^{-2}$ ) frente a  $R_s$ . Tampoco lo hay entre las palabras que siguen al alargamiento ( $R_w^1$  y  $R_w^2$ ). Sin embargo, la sílaba alargada (i.e.  $D_{hes}$ ) tiene mayor rango y mayor valor medio. El test *LSD* muestra que la única diferencia significativa existente es la que existe entre  $D_{hes}$  y el resto de variables. Los intervalos del test *LSD* indican que el alargamiento de la sílaba está entre 141ms y 239ms con respecto a la duración media de las sílabas en la frase. Podemos resumir las modificaciones locales en el caso de los alargamientos como el alargamiento de la sílaba extendida, mientras que el resto de variables permanecen aparentemente inalteradas.

Las repeticiones se componen de dos elementos principales: las locuciones repetidas que corresponden a  $RM$  y  $RR$ . La figura 2 muestra que mientras  $R_{RR}$  es parecido a  $R_s$ ,  $R_{RM}$  y  $D_{syl}^{-1}$  son más largos que  $R_s$ . Estas observaciones coinciden con las realizadas en trabajos previos [12, 19]. El test *LSD* muestra que existe una diferencia significativa entre los valores medios de  $R_{RM}$  y de  $D_{syl}^{-1}$  con respecto al resto de propiedades. Podemos concluir que para el caso de las repeticiones,  $RM$  parece pronunciarse más despacio que el resto de la frase y que la última sílaba de  $PrevRM$  ( $D_{syl}^{-1}$ ) es alargada sistemáticamente. *LSD* también indica que el valor medio de  $D_{syl}^{-1}$  es entre 10ms y 88ms más largo que el de  $R_s$ . También  $R_{RM}$  es entre 70ms y 148ms

más largo que la duración media de las sílabas en la frase. Sólo el 17% de las muestras contienen un silencio en ese punto. Si quitamos estos ejemplos de los datos, la diferencia sigue siendo significativa.

Las pausas rellenas (*filled pauses*) no pueden ser consideradas estrictamente como sílabas. Sin embargo, la duración del relleno ( $R_f$ ) puede ser comparado con la duración de una sílaba. Puede verse que todas las medidas, excepto  $R_f$ ,  $D_{syl}^{-1}$  y  $R_w^{-1}$ , siguen una distribución similar a la que sigue  $R_s$  (figura 2). Este hecho también se apoya en la observación de que el test *LSD* indica que no hay diferencia significativa con respecto a los valores medios del resto de variables. Según el test *LSD*, la sílaba pre-filler ( $D_{syl}^{-1}$ ) –a diferencia de la sílaba pre-pausa– es entre 100ms y 161ms más larga que  $R_s$ . La duración de la pausa rellena tiene un valor medio de 300ms y una desviación estándar de 185ms lo que implica que la longitud de la pausa rellena supera claramente el rango de variación de las sílabas pre-pausales (ver Table 2). De nuevo, el ritmo de la frase global no se modifica y sólo se observan alteraciones locales en forma de alargamientos previos a la sílaba previa al relleno.

Para cada tipo de disfluencia analizada, se han identificado los elementos clave que las definen, observando una serie de modificaciones locales a la disfluencia que deberán ser reproducidas en la síntesis. Estas modificaciones deberán ser aplicadas una vez que las partes correspondientes de OS y TS han sido sintetizadas y ensambladas.

En experiencias previas, hemos construido para ello sistemas de reglas [20] o modelos de regresión [21]. En la figura 3 mostramos los modelos de regresión para el caso de los alargamientos. Estudios previos han mostrado que la duración percibida de un alargamiento es la compuesta por la duración de la sílaba más la duración del silencio posterior ( $D_{ala} = D_{sil} + D_{syl}$ ) [9]. A partir de las observaciones del corpus, podemos inferir que la sílaba alargada solo puede serlo hasta un máximo y que sólo a partir de ese momento aparece un silencio. Para alargamientos superiores a este valor podemos modelarlos mediante reglas de regresión. Finalmente, el modelo propuesto puede expresarse como:

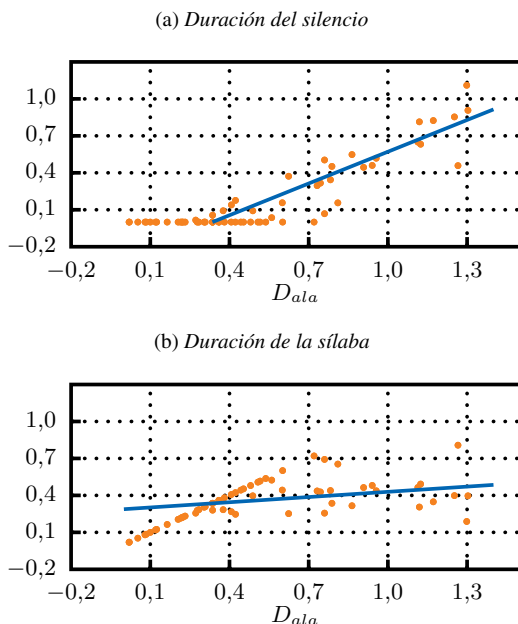


Figura 3: Líneas de regresión de la duración de la sílaba y el silencio posterior en un alargamiento

$$D_{sil} = \begin{cases} 0 & \text{if } D_{ala} \leq 0,337 \\ 0,819 * D_{ala} - 0,174 & \text{if } D_{ala} > 0,337 \end{cases} \quad (1)$$

$$D_{syl} = \begin{cases} D_{ala} & \text{if } D_{ala} \leq 0,337 \\ 0,181 * D_{ala} + 0,174 & \text{if } D_{ala} > 0,337 \end{cases} \quad (2)$$

El modelo propuesto ha sido implementado en nuestro sistema de CTV [22]. Se han realizado tests perceptuales informales que nos permiten afirmar que las modificaciones locales de las duraciones segmentales hacen que la inclusión de disfluencias no provoque una degradación significativa de la naturalidad general del sistema. También se ha observado que si no se aplican estas variaciones el resultado es menos natural.

#### 4. CONCLUSIONES Y TRABAJO FUTURO

En esta comunicación se ha presentado un modelo para la generación de disfluencias en sistemas CTV que se apoya en la consideración de tres frases constitutivas: La frase con disfluencias (DS), la frase original (OS) y la frase objetivo (TS). OS y TS proporcionan el contexto donde el *Reparandum* y el *Repair* respectivamente son pronunciados correctamente. El término *ET* se inserta en la frase final, utilizando la base de datos de síntesis y aplicando una serie de modificaciones locales a las sílabas adyacentes. La aportación principal es que este modelo puede ser aplicado en síntesis de voz utilizando los modelos prosódicos previamente entrenados con voz sin disfluencias.

Para estudiar las modificaciones locales en el ritmo, hemos considerado tres tipos de disfluencias: alargamientos, repeticiones y pausas rellenas. Para estos tres tipos de disfluencias, hemos comprobado que existen variaciones significativas con respecto al ritmo general de la frase, pero que dichas variaciones afectan exclusivamente al propio *ET* y a la parte final de *Reparandum*. Estos resultados apoyan la oportunidad del modelo propuesto.

Actualmente se están estudiando las variaciones locales provocadas en los contornos de *F0*. En el futuro no se descarta estudiar otros aspectos como la calidad de la voz o la energía. Nuestro propósito es el estudio de las variaciones de un conjunto de parámetros prosódicos que nos permita sintetizar las disfluencias, aprovechando los modelos previamente entrenados para voz sin disfluencias, aplicando a posteriori modificaciones locales que afecten sólo al entorno de la disfluencia.

Por otro lado, otro problema que debe ser tratado, se refiere a la predicción de la posición que deben ocupar las disfluencias. En este artículo sólo hemos abordado la cuestión de cómo sintetizar las disfluencias, y no hemos considerado donde incluir dichas disfluencias. Ésta es una tarea compleja que está fuera del ámbito de nuestra investigación por el momento. Otro aspecto importante a tener en cuenta es la elección de un *PostRM* adecuado para las repeticiones. A pesar de la importancia de estos dos aspectos, hay que tener en cuenta que en aplicaciones de generación de voz en sistemas de diálogo, esta información puede ser aportada al sistema CTV en base a una serie de reglas. También hay otras aplicaciones como la traducción voz-voz donde la posición de las disfluencias puede venir dada de acuerdo a la locución que debe ser traducida.

#### 5. BIBLIOGRAFÍA

- [1] A. Aaron, E. Eide, y J.F. Pitrelli, "Conversational computers," *Scientific American*, vol. 292, no. 6, pp. 64–69, June 2005.
- [2] Mark Fraser y Simon King, "The blizzard challenge 2007," in *Proceedings of the Blizzard Challenge*, 2007.
- [3] Jea E. Fox Tree, "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech," *Journal of Memory and Language*, vol. 34, no. 6, pp. 709–738, December 1995.
- [4] Michiko Watanabe, Keikichi Hirose, Yasuharu Den, y Nobuaki Minematsu, "Filled pauses as cues to the complexity of following phrases," in *Proc. of Eurospeech*, September 2005, pp. 37–40, Lisbon, Portugal.
- [5] Shu-Chuan Tseng, *Grammar, Prosody and Speech Disfluencies in Spoken Dialogues*, Ph.D. thesis, Dpt. Linguistics and Literature, University of Bielefeld, April 1999.
- [6] Herbert H. Clark, "Speaking in time," *Speech Communication*, vol. 36, no. 1-2, pp. 5–13, January 2002.
- [7] Masataka Goto, Katunobu Itou, y Storu Hayamizu, "A real-time filled pauses detection system for spontaneous speech recognition," in *Proc. of EUROSPEECH*, Budapest, Hungary, 1999, pp. 227–230.
- [8] Shiva Sundaram y Shrikanth Narayanan, "An empirical text transformation method for spontaneous speech synthesizers," in *Proc. of EUROSPEECH*, Geneva, Switzerland, September 2003, pp. 1221–1224.
- [9] R. Carlson, K. Gustafson, y E. Strangert, "Cues for hesitation in speech synthesis," in *Proceedings of Interspeech 06*, Pittsburgh, USA, 2006.
- [10] Elizabeth Ellen Shriberg, *Preliminaries to a Theory of Speech Disfluencies*, Ph.D. thesis, Berkeley's University of California, 1994.
- [11] William Levelt y A. Cutler, "Prosodic marking in speech repair," *Journal of Semantics*, pp. 205–217, 1983.
- [12] Elizabeth E. Shriberg, "Acoustic properties of disfluent repetitions," in *Proc. of International Conference on Phonetic Sciences (ICPhS)*, 1995, vol. 4, pp. 384–387, Stockholm, Sweden.
- [13] Manuel Almeida, "Organización temporal del español: el principio de isocronía," *Revista de Filología Románica*, vol. 1, no. 14, pp. 29–40, 1997, Madrid.
- [14] Guillermo Andrés Toledo, *El ritmo en el español: estudio fonético con base computacional*, Number 361 in Biblioteca románica hispánica ; II. Estudios y ensayos. Gredos, 1988.
- [15] Mar Carrió y Antonio Ríos, "Compensatory shortening in spanish spontaneous speech," in *Proceedings of ESCA Workshop on Phonetic and Phonology of Speaking Styles*, September 1991, vol. 16, pp. 1–5, Barcelona, Spain.
- [16] David Conejero, Jesús Giménez, Victoria Arranz, Antonio Bonafonte, Neus Pascual, Núria Castell, y Asunción Moreno, "Lexica and corpora for speech-to-speech translation a trilingual approach," in *Proc. of Eurospeech*, September 2003.
- [17] Michael Frigge, David C. Hoaglin, y Boris Iglewicz, "Some implementations of the boxplot," *The American Statistician*, vol. 43, no. 1, pp. 50–54, February 1989.
- [18] D.J. Saville, "Multiple comparison procedures: The practical solution," *The American Statisticians*, vol. 44, no. 2, pp. 174–180, May 1990.
- [19] Jordi Adell, Antonio Bonafonte, y David Escudero, "Disfluent speech analysis and synthesis: a preliminary approach," in *Proc. of 3th International Conference on Speech Prosody*, May 2006, Dresden, Germany.
- [20] Jordi Adell, Antonio Bonafonte, y David Escudero, "Filled pauses in speech synthesis: towards conversational speech," *10th International Conference on Text, Speech and Dialogue (LNAI)*, vol. 1, pp. 358–365, September 2007, Springer Verlag.
- [21] Jordi Adell, Antonio Bonafonte, y David Escudero, "Statistical analysis of filled pauses' rhythm for disfluent speech synthesis," in *Proc. of the 6th IWSS*, Bonn, Germany, August 2007.
- [22] Antonio Bonafonte, Pablo Daniel Agüero, Jordi Adell, Javier Pérez, y Asunción Moreno, "Ogmios: The UPC Text-to-Speech synthesis system for spoken translation," in *TC-STAR Workshop on Speech-to-Speech Translation*, June 2006.