DISEÑO Y EVALUACIÓN DE UN MODELO DE DURACIÓN VOCÁLICA DEL ESPAÑOL PARA LA SÍNTESIS DE HABLA

Rafael Marín Gálvez

Departamento de Filología Española
Universidad Autónoma de Barcelona

1. Introducción[1]

En este estudio hemos diseñado un modelo de duración vocálica del español para su implementación en un conversor de texto a habla[2].

Algunos autores, como Van Santen (1992b), han señalado ya la necesidad de desarrollar modelizaciones como la que proponemos aquí en las que se incluyan aspectos tales como el orden de aplicación de las reglas de duración, la estructura que deben presentar los datos para poder obtener una implementación plausible, etc.

Uno de los motivos fundamentales de esta necesidad se debe a la enorme complejidad que presenta el fenómeno de la duración. En este sentido, Van Santen señala: "Generally, the number of parameters of a model increases linearly with the number of factors, while the number of cells increases exponentially"[3].

2. Diseño del modelo

2.1. Elección de las variables

Hemos elegido como factores que influyen en la duración vocálica aquellos que en Marín (en prensa) aparecen como estadísticamente significativos. De los

[1] En primer lugar, cabe decir que este trabajo es una continuación del estudio de Marín (en prensa). De dicho estudio hemos extraído gran parte de los datos que vamos a utilizar en el diseño de un modelo de duración vocálica.
Para obtener estos datos se entrevistó a dos locutores hablantes de español peninsular estándar, con edades comprendidas entre 35 y 45 años. El corpus de lectura se extrajo del editorial de un periódico. El número total de vocales analizadas es de 491.

[2] Los resultados que proporcionamos en este trabajo han sido incorporados ya en el conversor de texto a habla SINCAS de la Universidad Ramon Llull.

[3] Van Santen (1992b), p. 277.

resultados que aparecen en el trabajo citado se pueden extraer las siguientes reglas de duración:

-Una vocal en posición prepausal aumenta un 47% su duración con respecto a esa misma vocal en posición no prepausal.
-En posición no prepausal, una vocal acentuada aumenta un 20% su duración con respecto a esa misma vocal inacentuada.
-En posición prepausal, una vocal acentuada aumenta un 10% su duración con respecto esa misma vocal inacentuada.
-En posición prepausal, una vocal perteneciente a sílaba abierta aumenta un 20,4% su duración con respecto esa misma vocal en sílaba cerrada.
-La vocal [a] presenta un incremento de duración del 7,5% con respecto a las vocales [e], [o].
-Las vocales [e], [o] presentan un incremento de duración del 6,5% respecto a las vocales [i], [u].

Así pues, las variables que incluimos en el diseño del modelo son las siguientes: duración intrínseca, posición en la frase, acento y estructura silábica (esta última variable sólo aparece en el caso de las vocales en posición prepausal)[4].

2.2. Elección del modelo

Existen diversas posibilidades a la hora de elegir un modelo para la representación de nuestros datos[5]. Hemos considerado conveniente no elegir ninguno de estos sistemas a priori y confrontar nuestros datos con las predicciones de estos diferentes sistemas para saber cuál de ellos se ajusta más a nuestros resultados.

---

[4] Las variables que no hemos incorporado en la elaboración del modelo son: sonoridad y modo de articulación de la consonante posterior a la vocal.
[5] Vid. Van Santen y Olive (1990) para una detallada explicación de los diferentes modelos que han sido propuestos.

	- ac	+ ac	diferencia
a	61,43	77,50	16,07
e	58,82	70,83	12,01
i	57,21	64,04	6,83
o	58,95	69,35	10,40
u	53,61	68,50	14,89

Tabla 1. *Diferencias de duración de las vocales en función del acento. (Valores medios expresados en milisegundos).*

En la tabla 1 se puede observar que el incremento producido por el acento en la duración de las vocales en posición no prepausal es, excepto en el caso de [u], proporcional y no absoluto. Debido a ello, hemos optado por emplear un modelo multiplicativo, ya que los resultados de que partimos así lo exigen.

2.3. Elección del sistema de representación

Ante la posibilidad de representar nuestros datos mediante un sistema de reglas secuenciales o una tabla de consulta de las duraciones de cada vocal teniendo en cuenta los diversos parámetros que utilizamos, o la posibilidad de estructurar nuestros datos estableciendo reglas más generales y de un alcance más amplio mediante un diagrama arbóreo, hemos optado por esta segunda opción[6]. Con un sistema de representación de este tipo se consigue, además, una mayor simplicidad, lo cual permitirá una mayor facilidad en la implementación de los datos.

Los incrementos porcentuales que aparecen en el diagrama de la figura 1 corresponden a una estructuración de las reglas de duración vocálica que hemos proporcionado anteriormente y a una estilización de los resultados que de éstas se derivan[7].

---

[6] Van Santen (1990) distingue cuatro sistemas de representación: sistemas de reglas secuencias, tablas de consulta, sistemas de ecuaciones y diagramas arbóreos.

[7] A lo largo del presente estudio hemos modelado los resultados de forma que pudieran integrarse en el sistema que hemos adoptado, pero solo hemos mantenido aquellas estilizaciones que no alejaran las predicciones del modelo de los datos que poseíamos.

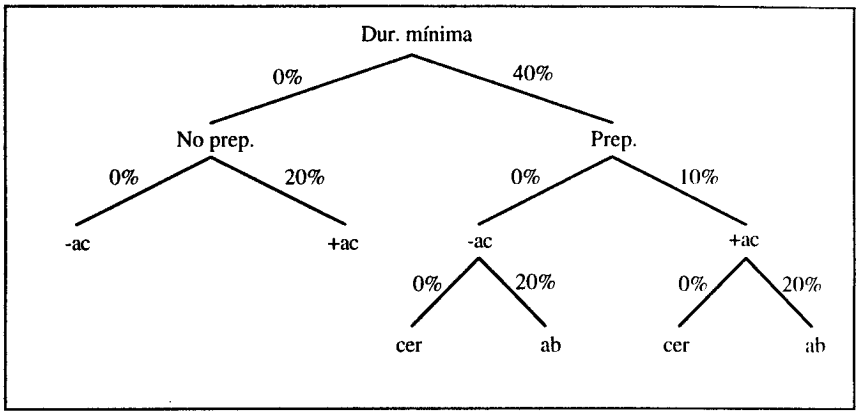

Figura 1. *Modelo de duración vocálica. 1ª aproximación.*

Este árbol se aplica de forma general a cada una de las cinco vocales. Para ello debemos establecer una duración base o de partida, que en nuestro caso coincide con una hipotética duración mínima que, de acuerdo con los datos de que partimos, queda establecida de la siguiente forma: [a]: 63 ms., [e]: 58 ms., [i]: 54 ms., [o]: 58 ms., [u]: 54 ms.

En la figura 2 podemos observar los resultados que proporciona este modelo para un caso concreto ([e], [o]):

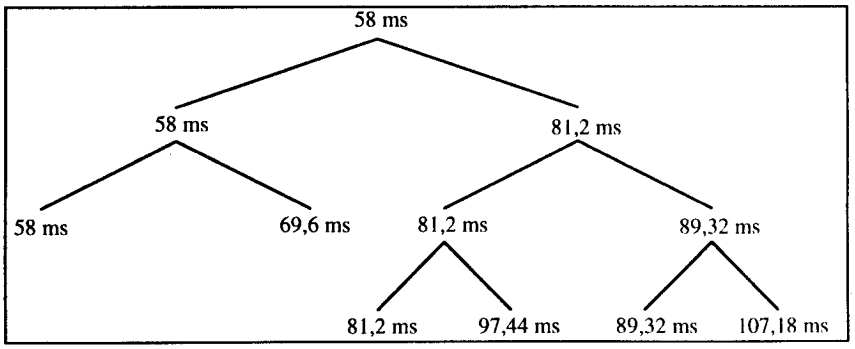

Figura 2. *Modelo de duración vocálica. Datos para un caso concreto ([e] y [o]).*

El hecho de considerar que la duración de base es igual a la duración mínima redunda aun más en beneficio de la simplicidad de nuestro modelo. De esta forma, sólo debemos operar en las ramas derechas del árbol, dejando en

las ramas de la izquierda el valor que aparece en el nodo inmediatamente superior.

3. Evaluación del modelo

3.1. Un primer test de evaluación

Con el fin de comprobar si las predicciones de nuestro modelo son adecuadas, hemos llevado a cabo una primera evaluación. Para ello, hemos contrastado las predicciones de nuestro modelo con los datos extraídos de un nuevo experimento[8].

En la tabla 2 se puede apreciar que no existen grandes diferencias, por lo que respecta a la duración intrínseca y al influjo del acento en la duración de las vocales en posición no prepausal, con respecto a los datos que predice el modelo:

	modelo	1er. test	Dif.
[a], +ac	75,60	80,08	+4,48
[e], +ac	69,60	68,00	-1,60
[i], +ac	64,80	61,76	-3,04
[o], +ac	69,60	71,09	+1,49
[u], +ac	64,80	64,62	-0,18
[a], -ac	63,00	62,26	-0,74
[e], -ac	58,00	60,86	+2,86
[i], -ac	54,00	55,17	+1,17
[o], -ac	58,00	61,95	+3,95
[u], -ac	54,00	61,43	+7,43

Tabla 2. *Valores medios de duración (expresados en milisegundos) de las vocales en posición no prepausal en función del acento.*

[8] Para obtener estos datos hemos entrevistado a un locutor masculino de 23 años de edad, hablante de español peninsular estándar. El corpus de lectura ha sido extraído del editorial de un periódico. El número total de vocales analizadas es de 247.

Diferencias entre los valores que predice el modelo y los valores obtenidos en un nuevo experimento.

Hemos aplicado un *t-test de Student* para cada uno de los pares que aparecen en la tabla 2 y, excepto para [u] en sílaba átona, las diferencias no son estadísticamente significativas.

Por lo que respecta a las vocales en posición prepausal, los resultados son los siguientes:

	modelo	1er. test	Dif.
-ac, cer	81,20	80,33	-0,87
-ac, ab	97,44	83,62	-13,82
+ac, cer	89,34	85,33	-4,01
+ac, ab	107,18	127,00	+19,82

Tabla 3. *Valores medios de duración (expresados en milisegundos) de las vocales en posición prepausal. Diferencias entre los valores que predice el modelo y los valores obtenidos en un nuevo experimento.*

Como se puede observar en la tabla 3, las predicciones del modelo concuerdan considerablemente con los resultados hallados en un nuevo experimento en el caso de las vocales que pertenecen a sílaba cerrada, sean o no acentuadas. No ocurre lo mismo en lo que respecta a las vocales que pertenecen a sílaba abierta, donde las diferencias entre las predicciones del modelo que proponemos y los resultados obtenidos para realizar esta primera evaluación son estadísticamente significativas, como lo demuestra la aplicación de un *t-test de Student.*

Hemos incorporado los nuevos datos obtenidos a los que ya poseíamos y hemos modificado el modelo, con el objetivo de mejorarlo. De este modo, el modelo queda reestructurado de la siguiente forma:

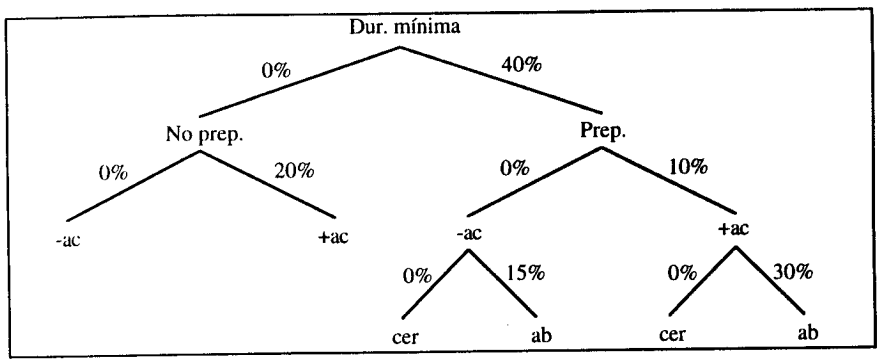

Figura 3. *Primera modificación del modelo de duración vocálica.*

Como se puede observar en la figura 3, sólo hemos debido corregir el modelo en el caso de las vocales en posición prepausal. Para solucionar las diferencias que hallábamos en el caso de [u] en posición no prepausal, será suficiente con modificar la duración base de esta vocal.

3.3. Segundo test de evaluación

Hemos llevado a cabo un segundo experimento[9]. En la tabla 4 establecemos una comparación entre las predicciones del modelo y los resultados que proporciona este nuevo experimento.

[9] En este caso, el informante es uno de los estudiados ya en Marín(en prensa). El corpus es el que utilizan Puigví y Fernández (1992). Este nuevo corpus es literario, mientras que el utilizado hasta ahora es periodístico. Pensamos que de esta forma la contrastación puede resultar más significativa.

	modelo	2° test	Dif.
[a], +ac	75,60	79,73	+4,13
[e], +ac	69,60	73,94	+4,34
[i], +ac	64,80	60,54	-4,26
[o], +ac	69,60	69,30	-0,30
[u], +ac	64,80	66,85	+2,05
[a], -ac	63,00	63,18	+0,18
[e], -ac	58,00	56,39	-1,61
[i], -ac	54,00	58,63	+4,63
[o], -ac	58,00	56,61	-1,39
[u], -ac	54,00	51,36	-2,64

Tabla 4. *Valores medios de duración (expresados en milisegundos) de las vocales en posición no prepausal en función del acento. Diferencias entre los valores que predice el modelo y los valores obtenidos en un nuevo experimento.*

En este caso, por lo que respecta a las vocales en posición no prepausal, la aplicación de un *t-test de Student* demuestra que las diferencias entre las predicciones del modelo y los resultados de estas nuevas mediciones no son significativas en ningún caso. Con lo cual se reafirma la idea de que las predicciones de nuestro modelo parecen ser acertadas. En la tabla 4 se puede observar que el grado de coincidencia es bastante elevado.

En cuanto a las vocales en posición prepausal, podemos observar, como se aprecia en la tabla 5, que no existen grandes diferencias en el caso de las vocales que pertenecen a sílaba cerrada, sean o no acentuadas. No obstante, en lo que respecta a las vocales que aparecen en sílaba abierta, la aplicación de un *t-test de Student* demuestra que las diferencias son significativas.

	modelo	2° test	Dif.
-ac, cer	81,20	87,12	+5,92
-ac, ab	93,38	108,20	+14,82
+ac, cer	89,34	103,33	+13,99
+ac, ab			

Tabla 5. *Valores medios de duración (expresados en milisegundos) de las vocales en posición prepausal. Diferencias entre los valores que predice el modelo y los valores obtenidos en un nuevo experimento.*

Nuevamente, como ocurría en el primer test de evaluación, son las vocales que pertenecen a sílaba abierta las únicas que se alejan de las predicciones que proporciona el modelo de duración vocálica que hemos elaborado.

Esta coincidencia nos hace pensar que las vocales que pertenecen a sílaba abierta y que se hallan en posición prepausal presentan una gran variación y un comportamiento poco homogéneo; características éstas que hacen difícil una aproximación como la que estamos llevando a cabo en este estudio.

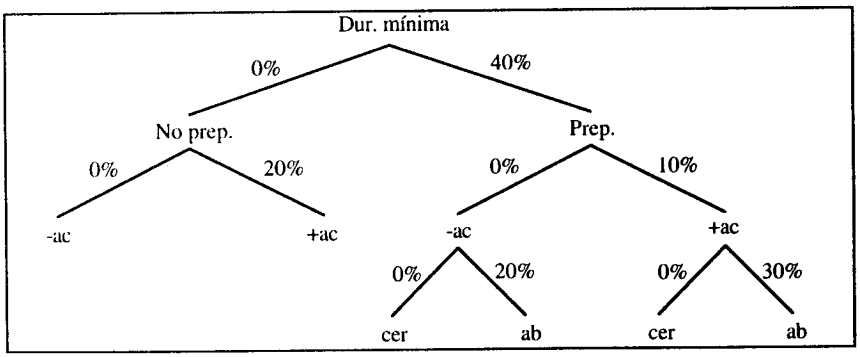

Figura 4. *Segunda modificación del modelo de duración vocálica.*

Como se puede observar en la figura 4, hemos modificado de nuevo el modelo de duración, en este caso con la incorporación de los datos obtenidos en un segundo experimento.

4. Conclusiones

A partir de los resultados obtenidos podemos extraer las siguientes conclusiones:

Las características que presentaban los datos de que partíamos nos han llevado a proponer un modelo de tipo multiplicativo y nos han permitido utilizar como sistema de representación de este modelo un diagrama arbóreo.

El modelo de duración vocálica que hemos elaborado ofrece unas predicciones que funcionan razonablemente bien por lo que respecta a la duración intrínseca de cada vocal y a la influencia de la posición en la frase y del acento. Los dos tests de evaluación que hemos llevado a cabo así lo demuestran.

Por lo que respecta a la influencia de la estructura silábica en las vocales que aparecen en posición prepausal, cabe decir que, si bien las predicciones para las vocales que se insertan en sílaba trabada parecen ser acertadas, no ocurre lo mismo en lo referente a las vocales que aparecen en sílaba abierta. En este último caso, los dos tests de evaluación que hemos realizado demuestran que nuestras predicciones parecen ser erróneas. No obstante, cabe señalar que la duración de estas vocales presenta un alto grado de variación.

Bibliografía

BARTKOVA, K. and SORIN, C. (1987). "A model of segmental duration for speech synthesis in French", *Speech Comunication*, 6, 245-260.

CAISSE, M. (1982). "Context-induced vowel duration change and intrinsec vowel duration", *JASA*, 72(S1), S65(A).

CAMPBELL, W. N. (1992). "Sillable-based segmental duration", in *Talking Machines: Theories, Models and Designs*, ed. by G. Bailly, C. Benoit and T. R. Sawallis (North-Holland, Amsterdam), pp. 211-224.

CARLSON, R. (1991). "Duration models in use", in *Proceedings XIIth International Congress on Phonetic Sciences*, pp. 243-246.

CARLSON, R. and GRANSTRÖM, B. (1986). "A search for durational rules in a real-speech database", *Phonetica*, 43, 140-154.

KLATT, D. H. (1973). "Interaction beetween two factors that influence vowel duration. *JASA*, 54, pp. 1102-1104.

KLATT, D. H. (1975). "Vowel lengthening is syntactically determined in a connected discourse", *Journal of Phonetics*, 3, pp. 129-140.

KLATT, D. H. (1976). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", *JASA*, 59. pp. 1208-1222.

MACARRÓN, A. *et alia.* (1991). "Generation of duration rules for a Spanish text-to-speech synthesizer, *Proceedings of Eurospeech 91,* pp. 617-620.

MARÍN, R. (en prensa). "La duración vocálica en español", *Estudios de Lingüística de la Universidad de Alicante.*

MORTAMET, L. (1991). "Implementing duration expert rules into a text-to-speech synthesis system", *Proceedings of Eurospeech 91,* pp. 621-624.

PUIGVÍ, D. y FERNÁNDEZ, J. M. (1992). "Estudi de la durada i la situació de les pauses en castellà", trabajo no publicado, Departamento de Filología Española, Universidad Autónoma de Barcelona.

RILEY, M. D. (1992). "Tree-based modelling of segmental durations", in *Talking Machines: Theories, Models and Designs,* ed. by G. Bailly, C. Benoit and T. R. Sawallis (North-Holland, Amsterdam), pp. 265-273.

SANTOS, A. *et alia.* (1988). Diseño y evaluación de reglas de duración en la conversión de texto a voz, *SEPLN,* boletín nº 6, pp. 71-92.

VAN SANTEN, J. P. H. and OLIVE, J. P. (1989). "Diagnostic tests of segmental duration models", *JASA,* 85, S43 (Q1).

VAN SANTEN, J. P. H. and OLIVE, J. P. (1990). "The analysis of contextual effects on segmental duration", *Computer Speech and Language,* Vol. 4, pp. 359-391.

VAN SANTEN, J. P. H. (1990). "Deriving text-to-speech durations from natural speech", *Proceedings of the ESCA Workshop on Speech Synthesis,* pp. 157-160.

VAN SANTEN, J. P. H. (1992a). "Contextual effects on vowel duration", *Speech Communication,* 11, pp. 513-546.

VAN SANTEN, J. P. H. (1992b). "Deriving text-to-speech durations from natural speech", in *Talking Machines: Theories, Models and Designs,* ed. by G. Bailly, C. Benoit and T. R. Sawallis (North-Holland, Amsterdam), pp. 275-286.