Procesamiento del Lenguaje Natural, num. 29 (2002), pp.147-152 recibido 09-05-2002; aceptado 02-07-2002

Modelo de duración para conversión de texto a voz en euskera

Eva Navas
Universidad del País Vasco
Alda. Urquijo s/n
eva@bips.bi.ehu.es

Inmaculada Hernáez
Universidad del País Vasco
Alda. Urquijo s/n
inma@bips.bi.ehu.es

Juan María Sánchez
Universidad del País Vasco
Alda. Urquijo s/n
ion@bips.bi.ehu.es

Resumen: En este artículo se presenta el trabajo realizado en el modelado de la duración de los fonemas en euskera estándar, para ser utilizado en conversión de texto a voz. El modelado estadístico se ha llevado a cabo mediante árboles binarios de regresión utilizando un corpus de 57.300 fonemas. Se han realizado varios experimentos de predicción testeando diferentes factores de influencia. El resultado obtenido en la predicción de la duración tiene un RMSE de 22.23 ms.
Palabras clave: Duración, conversión de texto a voz, modelado prosódico

Abstract: This paper presents the modelling of phone durations in standard Basque, to be included in a text-to-speech system. The statistical modelling has been done using binary regression trees and a large corpus containing 57.300 phones. Several experiments have been performed, testing different sets of predicting factors. The result when predicting durations with this model has a RMSE of 22.23 ms.
Keywords: Duration, text to speech conversion, prosody modelling.

1 *Introducción*

Con el fin de obtener un sistema de conversión de texto a voz (CTV) de gran calidad, la prosodia es un elemento al que se le ha de prestar especial atención. Por ello, todos los sistemas CTV incluyen un módulo de prosodia, en el cual se modelan normalmente la entonación, la duración de los sonidos y la energía. En este artículo se describe el trabajo realizado en el modelado de la duración de los sonidos con el fin de ser incluido en el CTV para el euskera basado en concatenación de difonemas, AhoTTS (Hernáez et al., 2001).

En el habla natural la duración de los sonidos es muy dependiente del contexto. El objetivo de un buen modelo de duración para CTV es estimar acertadamente esta variación basándose en información que se pueda obtener únicamente a partir del texto. Entre los factores contextuales con mayor influencia en la duración de los sonidos que se consideran habitualmente se encuentran la posición de la palabra en la frase, el fonema de que se trata y el contexto fonético en que se encuentra el sonido.

El modelo de duración más extendido es el sistema secuencial de reglas, propuesto por Klatt (1976), en el que se parte de un valor intrínseco de duración para cada sonido que es modificado posteriormente de acuerdo a ciertas reglas. Éste es el modelo de duraciones utilizado con anterioridad en AhoTTS.

En el siguiente apartado se describe la base de datos empleada en el estudio de las duraciones de los sonidos en euskera. Seguidamente, en el apartado 3 se detalla el análisis estadístico realizado sobre estos datos, en el apartado 4 se describen los experimentos realizados y los resultados obtenidos y finalmente, en el apartado 5 se presentan las conclusiones y líneas futuras de este trabajo.

2 *Base de datos*

Para el estudio de la duración segmental se ha utilizado la base de datos *Julen*, formada por 1.757 frases aisladas de longitud muy variada escritas en euskera estándar y leídas por un locutor nativo de Lesaka a velocidad normal y con entonación neutra.

El propósito de la grabación de esta base de datos fue obtener las unidades de síntesis necesarias para realizar conversión de texto a

ISSN 1135-5948

voz en euskera y por ello no se consideró como objetivo en el diseño del corpus el que tuviera gran variedad sintáctica. Tampoco se procuró incluir el mayor número de contextos fonéticos posibles para evitar la dispersión y escasez de datos que suele producirse en los estudios de duración (van Santen, 1994). A pesar de todo esto, la base de datos fue leída a velocidad constante y por lo tanto es adecuada para el estudio de las duraciones de los sonidos.

Para comprobar que la velocidad de lectura se mantuvo constante durante toda la grabación, la base de datos se dividió aleatoriamente en dos partes y se comparó el número medio de alófonos por segundo, así como la media y desviación típica de la duración de cada alófono, en cada una de ellas. En la Figura 1 se muestra el número medio de alófonos por segundo en cada parte en que se dividió la base de datos, así como en la base de datos completa, pudiéndose comprobar que no existe diferencia apreciable entre los valores obtenidos.

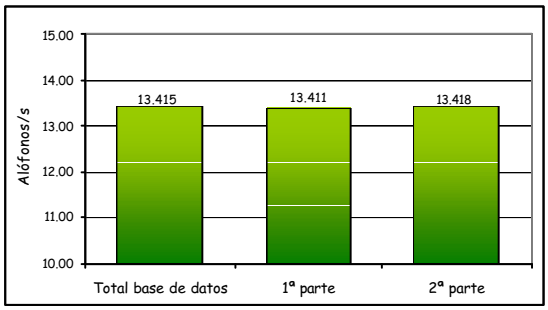

Figura 1: Velocidad de lectura en la base de datos *Julen*.

Las principales características de la base de datos *Julen* se detallan en la Tabla 1.

Tipo	Oral y textual
Propósito	Ud. de síntesis
Nº fonemas	57.300
Nº palabras	12.391
Nº frases	1.757

Tabla 1: Características de la base de datos *Julen.*

Esta base de datos fue marcada automáticamente a nivel de fonema utilizando la transcripción fonética que proporciona el módulo de procesado lingüístico de AhoTTS. Posteriormente las marcas y las etiquetas automáticas fueron revisadas manualmente por dos etiquetadores expertos, en los casos en que se detectaron errores de gravedad.

El inventario de sonidos utilizado es el del SAMPA para el euskera, que puede consultarse en http://bips.bi.ehu.es/SAMPA.html y que incluye 5 sonidos vocálicos y 28 consonánticos.

Los datos se dividieron aleatoriamente en un conjunto de entrenamiento (75% del total con 42.779 alófonos) y uno de test (25% del total con 14.521 alófonos).

3 Análisis estadístico

Para predecir la duración de los sonidos en conversión de texto a voz se han utilizado diversos métodos de análisis estadístico, principalmente redes neuronales (Riedi, 1995), árboles binarios de clasificación y regresión (Riley, 1992), (Lee y Oh, 1999), modelos multiplicativos (Shih y Ao, 1996) y sumas de productos (van Santen, 1992), aunque también se han usado otros modelos estadísticos (Riedi, 1997), (Goubanova, 2001).

En este trabajo se han elegido los árboles binarios de regresión (Breiman et al., 1984) como método de estudio estadístico de las duraciones de los sonidos, porque permiten manejar al mismo tiempo variables tanto de naturaleza continua como discreta, son capaces de seleccionar automáticamente los factores que tienen mayor influencia en la predicción de la variable objetivo y finalmente porque producen diagramas muy fáciles de interpretar.

3.1 Definición de la variable objetivo

El primer problema que se plantea a la hora de predecir la duración de los sonidos es la elección de la unidad de duración: tradicionalmente se ha utilizado el fonema como unidad fundamental (Riedi, 1998), (Febrer, Padrell y Bonafonte, 1998), pero también se han hecho estudios utilizando difonemas (O Shaughnessy et al., 1988), sílabas (Campbell, 1992) o palabras (Bouzon y Hirst, 2002). En este trabajo se ha elegido el fonema por ser la que más fácilmente podía ser integrada en el CTV AhoTTS que controla la duración a nivel segmental.

Otro problema que también hay que considerar es la distribución de las duraciones de los sonidos, que es típicamente una distribución log-normal (Huber, 1990). La mayoría de los métodos estadísticos asumen una distribución normal, por lo que en muchos casos se hace necesaria una codificación de la

duración que consiga acercar más su distribución a la normalidad.

En este trabajo se ha considerado la *z-score* o puntuación típica cuya expresión se muestra en la fórmula (1).

$$z-score=\frac{dur-m_k}{\sigma_k} \quad (1)$$

donde dur es la duración del sonido, m_k es la duración media de ese sonido en toda la base de datos y σ_k su desviación típica.

En la Figura 2 se muestra la distribución de las duraciones de los sonidos en la base de datos *Julen* y en la Figura 3 la distribución de su *z-score.* Como puede observarse en ellas, la distribución de la duración tiene un fuerte sesgo a la izquierda (media 75 ms., desviación estándar 30.7 ms.), mientras que la de la *z-score* se aproxima más a una distribución normal (media 0, desviación estándar 1).

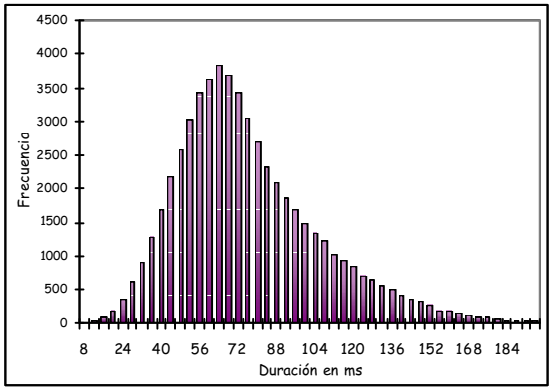

Figura 2: Distribución de la duración de los sonidos en la base de datos *Julen*.

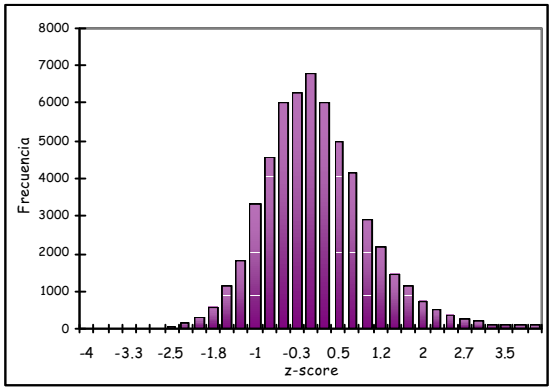

Figura 3: Distribución de la *z-score* de la duración de los sonidos en la base de datos *Julen.*

3.2 Información de predicción

Para la predicción de la duración segmental se han estudiado factores que influyen en la duración de los sonidos de distintas lenguas (Venditti y van Santen, 1998), (Córdoba et al., 1999), (Sproat, 1998). El factor que habitualmente es considerado el principal es la identidad del propio fonema, además de su contexto fonético, la posición del fonema en la palabra y la frase y su nivel de acento.

En este estudio el contexto fonético se ha reducido en principio a una ventana de longitud cinco en torno al fonema en estudio y por ello sólo se ha proporcionado al árbol información sobre el sonido actual, los dos anteriores y los dos siguientes. En concreto, las características que se dan para los fonemas incluidos en la ventana de estudio son las siguientes:

- Identidad del fonema: es un código numérico que indica el sonido de que se trata. Esta variable tiene 33 valores diferentes para el caso del fonema actual y para el caso de fonemas anteriores o siguientes, es necesario añadir un valor de inicio y fin de frase y un código para el silencio.
- Carácter vocálico o consonántico del fonema. En función del valor de esta variable se proporcionan unos u otros datos sobre los fonemas:
- Para las vocales se facilitan:
  - Altura de la vocal, que indica la mayor o menor apertura de la boca para pronunciar la vocal. Los posibles valores de esta variable son: abierta, media y cerrada.
  - Frontera de la vocal, que indica el punto de articulación de la vocal, es decir, la parte de la cavidad bucal en la que se forma el sonido. Para esta variable se han considerado los valores: anterior, central y posterior.
- Para las consonantes:
  - Tipo de consonante, que indica cómo se origina el sonido. Los valores que puede tomar esta variable son: oclusiva, fricativa, africada, nasal, líquida, aproximante y vibrante.
  - Punto de articulación, que se relaciona con el movimiento de la cavidad bucal y de los labios para generar el sonido. Los valores considerados para esta variable son: bilabial, alveolar, palatal, labio-

dental, dorso-alveolar, velar, palatal e interdental.

- Carácter de sonoridad de la consonante. Los valores que puede tomar son: sonoro y sordo, según vibren o no las cuerdas vocales en la producción del sonido.

También se tendrá en cuenta la posición léxica del sonido cuya duración queremos calcular, en concreto se estudiará la influencia de su posición dentro de la sílaba y de la palabra. Estas características se refieren sólo al sonido cuya duración se quiere predecir:

- Posición relativa del fonema en la sílaba. Los valores que puede tener son: inicial, final, medio y sílaba de un único fonema.
- Posición relativa en la palabra de la sílaba a la que pertenece el sonido. Esta variable puede tomar los valores: inicial, final, medio y monosílabo.
- Acentuación del fonema que indica para los sonidos vocálicos si el fonema es acentuado o no. Esta característica se extrae automáticamente de la información proporcionada por el módulo de procesado lingüístico de AhoTTS, y no se realiza ninguna corrección de los errores que pudiera haber.
- Pertenencia del sonido a la última sílaba o última palabra de la frase.
- Número de sílabas y palabras entre signos de puntuación.

4 Experimentos y resultados

Considerar una variable categórica con tantos niveles diferentes como la identidad del fonema, conlleva un problema de dispersión de los datos. A pesar del tamaño de la base de datos, es imposible incluir en ella todas las combinaciones lingüísticamente posibles de diferentes valores de los factores (van Santen, 1994). Al tener en cuenta el factor identidad del fonema, las combinaciones teóricas diferentes, considerando una ventana de 5 fonemas, son más de 39 millones, y aunque algunas de ellas son lingüísticamente imposibles, no todas las que pueden producirse realmente se pueden cubrir en la práctica aumentando el tamaño de la base de datos.

Con el fin de controlar la dispersión de datos que pueda producir el considerar este factor, se han realizado varios experimentos de predicción. Con la parte del corpus reservada para entrenamiento se han construido tres árboles de regresión para predecir la *z-score* de la duración, uno utilizando todas las variables predictoras (árbol I), otro en el que el tamaño de la ventana de fonemas considerados se ha reducido a tres (árbol II) y finalmente otro en el que no se proporciona la identidad de los fonemas, sino únicamente el resto de las variables articulatorias (árbol III).

4.1 Análisis de la influencia de los factores

La importancia dada por el árbol I a cada uno de los factores se muestra en la Tabla 2. Al factor que más peso tiene se le asigna una importancia del 100% y el resto de los valores se expresan de forma relativa a éste. Como puede verse en la Tabla 2, la identidad de los cinco fonemas contenidos en la ventana de estudio considerados conjuntamente es la característica más importante a la hora de determinar la duración y entre ellos los fonemas que más peso tienen son los más cercanos al actual. Además tienen influencia la sonoridad y el carácter vocálico o consonántico del fonema siguiente. Con estos 7 factores queda prácticamente determinada la duración, y en este caso no se tienen en cuenta ninguno de los factores relacionados con la posición del sonido en la palabra o frase.

Factor	Importancia
fonema siguiente	100.00%
siguiente sordo/sonoro	65.10%
fonema anterior	58.81%
siguiente vocal/cons.	51.75%
fonema actual	43.87%
fonema anterior al anterior	34.33%
fonema siguiente al siguiente	29.13%

Tabla 2: Importancia de las variables en la predicción de la *z-score* con el árbol I.

El árbol II considera estas mismas variables en la predicción de la duración, pero sustituye la identidad de los fonemas de los que carece por el carácter consonántico o vocálico del fonema anterior y el tipo de consonante tanto anterior, como actual y siguiente. En la Tabla 3 pueden verse los factores que el árbol II ha considerado más importantes para la predicción de la duración y la importancia que les ha asignado.

Factor	Importancia
fonema siguiente	100%
fonema anterior	72.22%
siguiente sordo/sonoro	58.82%
fonema actual	53.67%
anterior vocal/cons.	47.48%
tipo cons. anterior	36.44%
tipo cons. actual	22.61%
tipo cons. siguiente	22.02%

Tabla 3: Importancia de las variables en la predicción de la *z-score* con el árbol II.

En el caso del árbol III, es necesario considerar muchos más factores para determinar la duración de los sonidos, en concreto hace falta tener en cuenta 20 factores para llegar al mismo nivel de influencia que en los casos anteriores. En la Tabla 4 se muestran los ocho factores de mayor peso seleccionados por el árbol. El factor con mayor influencia es el carácter de sonoridad del fonema siguiente, aunque también el del fonema actual y el del anterior tienen gran importancia. Además tienen mucho peso el tipo de consonante actual, anterior y siguiente y el punto de articulación del sonido actual.

Factor	Importancia
siguiente sordo/sonoro	100%
tipo cons. actual	89.9%
siguiente vocal/consonante	86.89%
pto. articulación actual	65.36%
tipo cons. anterior	63.11%
anterior sordo/sonoro	59.47%
actual sordo/sonoro	52.86%
tipo cons. siguiente	49.12%

Tabla 4: Importancia de las variables en la predicción de la *z-score* con el árbol III.

En los tres experimentos realizados las características más influyentes se refieren al fonema anterior y al siguiente, y no al actual. Esto es debido a que la codificación de la duración utilizada, *z-score*, ya incluye esta influencia.

El acento, que habitualmente es considerado un factor influyente en la duración de los sonidos, no ha sido seleccionado en ningún caso por estos árboles como importante. La causa puede ser las discrepancias existentes entre la realización del acento por el locutor y las etiquetas asignadas automáticamente.

Así, según estos resultados, está claro que el contexto fonético es determinante para la estimación de la duración de un sonido. Para aumentar la precisión en la predicción, y permitir así que los árboles consideren otras variables referidas a contextos no fonéticos (como la posición en la palabra o en el grupo prosódico), se debería realizar una clasificación previa de los sonidos y habría que construir árboles específicos para cada grupo.

4.2 Resultados de predicción

La duración de los fonemas de la parte de la base de datos reservada para test se predijo usando los tres árboles construidos. En la Tabla 5 se muestran los resultados obtenidos en cada experimento. Como se observa en ella, a pesar de la dispersión de los datos que puede producirse en el primer caso, la predicción es mucho más acertada considerando la identidad de los fonemas, aunque mejora ligeramente al utilizar una ventana que incluye únicamente tres fonemas.

Arbol	RMSE	Coef. correl.
I	22.24 ms	0.701
II	22.23 ms	0.702
III	27.25 ms	0.627

Tabla 5: Resultados de los árboles de predicción de duración.

5 Conclusiones y trabajos futuros

En este trabajo se ha modelado estadísticamente la duración de los fonemas en euskera estándar mediante árboles binarios de regresión, con el fin de utilizar el modelo obtenido en un conversor texto a voz. El mejor modelo implementado tiene un RMSE de 22.23 ms. y utiliza la identidad del fonema siguiente, anterior y actual como factores más importantes en la predicción.

Es necesario realizar más experimentos con el fin de mejorar el RMSE obtenido, utilizando diferentes árboles para distintos grupos de sonidos, sobre todo para separar sordos de sonoros, ya que ésta es la característica articulatoria que mayor influencia ha tenido en todos las pruebas realizadas.

6 Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia y Tecnología (TIC2000-1005-C03-03 y TIC2000-1669-C04-03).

Bibliografía

Bouzon, C. y D. Hirst. 2002. The influence of prosodic factors on the duration of words in British English. En *Proceedings of 1st International Conference on Speech Prosody*, páginas 689-701, Aix-en-Provence.

Breiman, L., J.H. Friedman, R.A. Olsen y C.J. Stone. 1984. *Classification and Regression Trees*. Chapman & Hall, Boca Raton, Florida.

Campbell, W.N. 1992. Syllable-based segmental duration. *Talking machines: teories, models and designs*. páginas 211-224. Elsevier, Amsterdam.

Córdoba, R., J.A. Vallejo, J.M. Montero, J. Gutiérrez-Arriola, M.A. López y J.M. Pardo. 1997. Automatic modeling of duration in a Spanish text-to-speech system using neural networks. En *Proceedings of Eurospeech 99*, páginas 1619-1622, Budapest.

Febrer, A., J. Padrell y A. Bonafonte. 1998. Modeling Phone Duration: Application to Catalan TTS. En *Proceedings of 3rd International Workshop on Speech Synthesis*, Jenolan Caves, Australia.

Goubanova, O. 2001. Predicting segmental duration using Bayesian belief networks. En *Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Edimburgo.

Hernáez, I., E. Navas, J.L. Murugarren y B. Etxebarria. 2001. Description of the AhoTTS System for Basque Language. En *Proceedings of 4th ISCA Tutorial & Research Workshop on Speech Synthesis*, Edimburgo.

Huber, K. 1990. A statistical model of duration control for speech synthesis. En *Proceedings of the 5th European Signal Processing Conference (EUSIPCO 90)*, páginas 1127-1130, Barcelona.

Klatt, D. H. 1976. Linguistics uses of segmental duration in English: Acoustic and perceptual evidence *J. Acoust Soc. Am,*. 59:1209-1221.

Lee, S. H. y Y. H. Oh. 1999. Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems. *Speech Communication*, 28: 283-300.

O Shaughnessy, D. L. Barbeau, D. Bernardi y D. Archambault. 1988. Diphone Speech Synthesis. *Speech Communication*, 7: 55-65.

Riedi, M. 1995. A Neural-Network-Based Model of Segmental Duration for Speech Synthesis. En *Proceedings of Eurospeech 95*, páginas.599-602, Madrid.

Riedi, M. 1997. Modeling Segmental Duration with Multivariate Adaptive Regression Splines. En *Proceedings of Eurospeech 97*, páginas 2627-2630, Rhodes, Greece.

Riedi, M. 1998. Controlling segmental duration in speech synthesis systems. Tesis doctoral.

Riley, M.D. 1992. Tree-based Modelling of Segmental Durations *Talking Machines, Theories, Models, and Designs* North-Holland, Elsevier Science Publishers, Amsterdam, páginas 265-273.

Shih, C. y B. Ao. 1996. Duration study for the Bell laboratories mandarin text-to-speech system. *Progress in Speech Synthesis*. páginas 383-400. Springer, Berlín.

Sproat, R. 1998. *Multilingual text-to-speech synthesis: the Bell Labs approach*. Kluver Academic Publishers, Dordrecht, Holanda.

van Santen, J. P. H. 1992. Contextual effects in vowel duration. *Speech Communication*, 11:513-546.

van Santen, J. P. H. 1994. Assignement of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8:95-128.

Venditti, J.J. y J. P. H. van Santen. 1998. Modeling segmental durations for Japanese text-to-speech synthesis. En *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, páginas 31-36, Jenolan Caves, Australia.