Sobre el uso de la sílaba como unidad de síntesis en español

Leopoldo N. Feal Pinto

Departamento de Informática, Universidad de Valladolid, España.
pol@infor.uva.es

Informe Técnico IT-DI-2000-0004

Sobre el uso de la sílaba como unidad de síntesis en español

Leopoldo N. Feal Pinto

Departamento de Informática, Universidad de Valladolid, España.
pol@infor.uva.es

Resumen En este informe se propone y se trata de justificar la elección de la sílaba como unidad de síntesis para el español, dentro del marco de la síntesis de voz por concatenación de unidades. Asimismo, se presenta un estudio de las distintas sílabas que pueden encontrarse en el idioma español, a partir de un corpus de 799281 palabras obtenidas de varios textos. Las palabras se han transcrito fonéticamente según el alfabeto SAMPA[1] , y se han dividido en 1456067 sílabas de síntesis (sílaba normal + hiatos), con el objeto de incorporarlas como unidades sonoras a un sintetizador de voz. En el anexo que acompaña a este informe, se presenta una relación de las sílabas identificadas junto con varios ejemplos de aparición de las mismas.

1 Introducción

En los sintetizadores por concatenación de unidades, uno de los factores que influye decisivamente en la calidad de la voz resultante es la elección de la unidad de síntesis.

A medida que se han ido construyendo sintetizadores para diferentes idiomas, se han propuesto nuevas unidades que se adaptaban mejor a las particularidades del idioma objetivo; es habitual utilizar inventarios que mezclen distintos tipos de unidades, o incluso varias instancias de un mismo fragmento obtenidas en contextos diferentes.

En general, es deseable que la unidad elegida cumpla algunas propiedades básicas[Dut97]:

- Deberá contener tantos efectos coarticulatorios como sea posible.
- Deberá ser fácilmente conectable con el resto de unidades.
- Su número y longitud deberá ser tan pequeño como sea posible.

Las dos primeras propiedades sugieren que unidades más grandes (como palabras o frases) son preferibles, pero entran en conflicto con la tercera propiedad, que aconseja la elección de unidades más pequeñas.

Actualmente, la unidad más popular para la síntesis por concatenación es el *difono*, que comienza y termina en las zonas acústicamente estables de dos

[1] Speech Assesment Methods Phonetic Alphabet

sonidos contiguos, conteniendo así la transición entre ellos. Existe un punto de concatenación por fonema con esta aproximación, por lo que la segmentación deberá ser cuidadosa para que la concatenación de los difonos sea de buena calidad. Sin embargo, el número de difonos presentes en un idioma es pequeño (entre 1200 y 3000), dependiendo del idioma y de las variedades alofónicas consideradas para cada fonema.

Otro tipo de unidad que podría considerarse como adecuada para la síntesis es la *sílaba*. Tradicionalmente, se ha dicho de ella que a pesar de que puede producir buena calidad (ya que contiene como el difono transiciones entre sonidos, e incluso la densidad de puntos de concatenación es menor que en aquel caso), el número de sílabas presentes en un idioma es inmanejable (decenas de miles, dependiendo del idioma). No obstante, se ha utilizado en algunos casos con buenos resultados[PHH96].

En la próxima sección trataremos de justificar por qué en el caso del español, la elección de la sílaba podría ser muy adecuada para la síntesis, dado que, como veremos, su número es muy reducido.

2 La sílaba como unidad de síntesis

Es posible que la elección de unidades más pequeñas que la sílaba en la construcción de sintetizadores de voz para otros idiomas distintos del español haya influido en el tipo de unidad elegido para éste. Lo cierto es que en la documentación de varios sintetizadores por concatenación para el español a los que hemos tenido acceso, se utiliza principalmente el difono[B+98][LG93], a veces junto con otros elementos como trífonos o tetráfonos[F+94], para incrementar la calidad; hasta la fecha, no hemos encontrado ninguna aproximación basada en sílabas para la síntesis del español.

Fundamentalmente son dos los problemas que presenta el uso de la sílaba como unidad de síntesis[PHH96]:

- El gran número de sílabas distintas en un idioma.
- Los posibles efectos de coarticulación entre sílabas, que deberán ser tratados convenientemente.

Respecto del segundo problema, cabe decir que en español una sílaba siempre contiene un núcleo vocálico precedido y/o seguido por una o varias consonantes. La distinción entre unas consonantes y otras está asociada principalmente a las transiciones sufridas por los formantes de las vocales a las que acompañan[Qui88]. Teniendo esto en cuenta, sílabas de la forma

```
consonante(s)+vocal(es)+consonante(s)
```

contendrían todos los efectos coarticulatorios producidos al pronunciar esa secuencia de sonidos.

El problema podría darse entonces en la frontera entre sílabas, en casos donde una consonante, precedida y seguida por una vocal pudiera afectar a los formantes de ambas. A este respecto, Quilis[Qui88] describe una experiencia de

síntesis realizada por Bertil Malmberg en 1955, donde se apunta que la percepción de la sílaba en el caso de una consonante explosiva entre dos vocales está asociada a las transiciones experimentadas por los formantes de la vocal que forma parte de la sílaba, cuando se mantienen estables los de la otra vocal. Un efecto parecido se ilustra en la figura 1, donde en el espectrograma de la izquierda (correspondiente a la secuencia */ina/* en */inamovible/*) se observa cómo los formantes de la */i/* permanecen estables en toda su duración, sin presentar apenas efectos coarticulatorios con la consonante que le sigue; según nuestro concepto de sílaba, la secuencia */ina/* se divide en dos sílabas: */i–na/*. Por su parte, el espectrograma de la derecha representa la secuencia de sonidos */ink/* en la palabra */incontable/*. Aquí, los formantes de la */i/* presentan una variación frecuencial a lo largo del tiempo, permitiendo así una suave transición hacia la */n/*, sonido con el que forma sílaba.

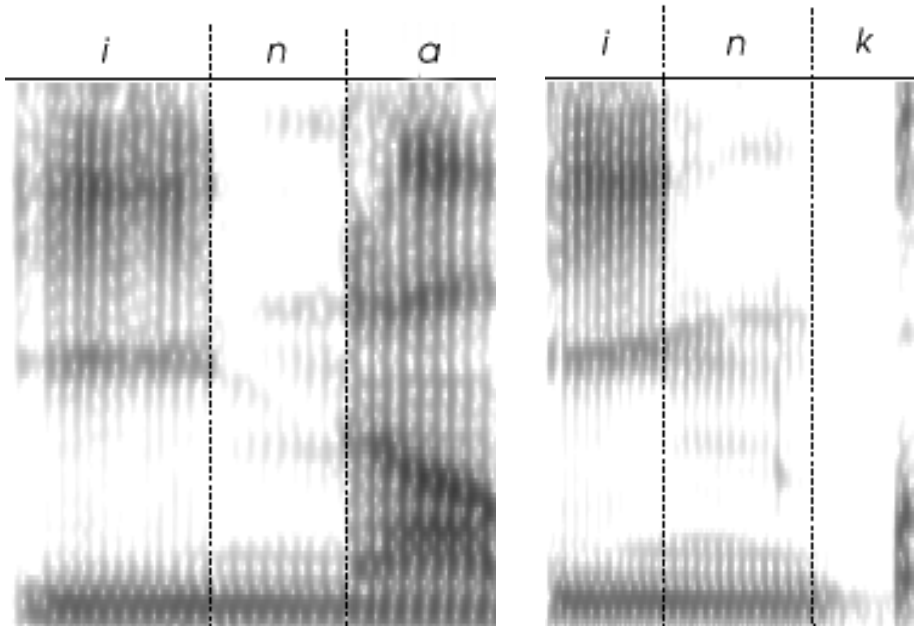

Figura1. Espectrogramas de /ina/ e /ink/

No obstante, el concepto tradicional de sílaba deja fuera combinaciones de sonidos que presentan un fuerte efecto coarticulatorio, como es el caso de los hiatos, donde dos sonidos vocálicos pertenecen a sílabas distintas. Es por eso que preferimos redefinir la estructura de la sílaba española, permitiendo cualquier combinación de vocales como núcleo silábico. A partir de este momento, cuando hablemos de sílabas, estaremos haciendo referencia a esta estructura extendida.

Así, hemos realizado un conteo de las sílabas que pueden encontrarse en español, intuyendo que su número no iba a ser excesivamente grande, a diferencia de otros idiomas como el inglés, francés o alemán. Los resultados, como se verá en la siguiente sección son esperanzadores: el número de sílabas se mantiene parejo al número de difonos para otros idiomas.

3 Identificación de las sílabas del español

Cuando se pretende identificar las sílabas de un idioma pueden seguirse dos aproximaciones: un *esquema generativo*, en el cual a partir de la estructura silábica del idioma se generan todas las posibles sílabas, o un *análisis de textos*, donde a partir de un corpus representativo del idioma elegido se dividen en sílabas las palabras que lo forman. La primera aproximación tiene muchos inconvenientes: algunas de las sílabas generadas no se utilizan en el idioma a pesar de tener una estructura válida; asimismo, será difícil encontrar ejemplos de todas ellas. El segundo enfoque evita estas dificultades cuando los textos elegidos representan convenientemente el idioma. No obstante, algunas sílabas se escaparán al análisis, debido a su escasa frecuencia de aparición.

Hemos adoptado la segunda solución, diviendo el proceso en cuatro etapas: *obtención de textos, extracción de palabras, transcripción fonética y división en sílabas*. En una etapa posterior se ha procesado la información obtenida para la elaboración del anexo.

3.1 Obtención de textos

Los textos sobre los que se ha efectuado el análisis se han obtenido de servidores web que los ofrecen en formato electrónico, como por ejemplo:

Proyecto Gutenberg: `http://www.promo.net/pg/`
Obras de Miguel de Cervantes: `http://cervantes.alcala.es/obras.htm`
Cervantes Project 2001: `http://www.csdl.tamu.edu/cervantes/english/ctxt/sb/`

Cabe destacar que no es fácil encontrar obras contemporáneas en formato electrónico, por lo que se han empleado principalmente obras que, aunque escritas hace más de dos siglos, utilizan una ortografía casi idéntica a la actual, aspecto muy importante de cara a conseguir sílabas empleadas en el español contemporáneo.

A continuación se muestra la lista de textos utilizados:

- *Don Quijote de la Mancha*, de Miguel de Cervantes.
- *La fuerza de la sangre*, de Miguel de Cervantes.
- *La Galatea*, de Miguel de Cervantes.
- *La gitanilla*, de Miguel de Cervantes.

- *El amante liberal*, de Miguel de Cervantes.
- *El casamiento engañoso*, de Miguel de Cervantes.
- *El celoso extremeño*, de Miguel de Cervantes.
- *El coloquio de los perros*, de Miguel de Cervantes.
- *La señora Cornelia*, de Miguel de Cervantes.
- *Las dos doncellas*, de Miguel de Cervantes.
- *La ilustre fregona*, de Miguel de Cervantes.
- *El licenciado vidriera*, de Miguel de Cervantes.
- *Los trabajos de Persiles y Segismunda*, de Miguel de Cervantes.
- *Rinconete y Cortadillo*, de Miguel de Cervantes.
- *La Tebaida*, de Publio Papinio Estacio, sobre el texto publicado en 1888.
- *Declaración de los Derechos Humanos.*
- *El lazarillo de Tormes*, anónimo.
- *Rimas y leyendas*, de Gustavo Adolfo Becquer.

3.2 Extracción de palabras

En esta etapa se han extraido las palabras presentes en los textos, entendiendo como palabras aquellas cadenas de caracteres que confrontan la siguiente expresión regular:

```
[a-zA-ZáéíóúÁÉÍÓÚñÑüÜ]+
```

Posteriormente, se procedió a la normalización de las cadenas resultantes, consistiendo ésta en la supresión de tildes y paso a minúsculas de cada uno de los caracteres.

Como resultado, se obtuvieron un total de 799281 palabras, de las cuales 36859 eran distintas.

3.3 Transcripción fonética de las palabras

En esta etapa se realizó una transcripción fonética de las palabras, conforme al alfabético fonético SAMPA (Speech Assesment Methods Phonetic Alphabet), que utiliza caracteres ASCII entre el 33 y el 127 para la codificación de fonemas, facilitando su lectura por parte de las computadoras (más información en `http://www.phon.ucl.ac.uk/home/sampa/home.htm`) . Este alfabeto en su versión para el castellano define los fonemas descritos en la tabla 1.

No obstante, la transcripción fonética llevada a cabo utiliza un subconjunto de estos fonemas, prescindiendo de las fricativas B,D,G y de las semivocales, pues resulta difícil distinguir en la práctica estos sonidos de sus correspondientes oclusivos y vocálicos, respectivamente. En general, la estrategia utilizada ha sido la de minimizar el número de alófonos considerados para mantener pequeño el número final de sílabas distintas.

Con esta aproximación, la transcripción fonética puede llevarse a cabo a través de simples sustituciones, como muestra el código para la herramienta SED[D+97] de la tabla reftab:trans (la entrada son palabras, una en cada línea, y deben aplicarse las sustituciones sucesivamente).

Simbolo	Palabra	Transcripción
Consonantes		
Oclusivas		
p	padre	"paDre
b	vino	"bino
t	tomo	"tomo
d	donde	"donde
k	casa	"kasa
g	gata	"gata
Africadas		
tS	mucho	"mutSo
jj	yate	"jjate
Fricativas		
f	fácil	"faTil
B	cabra	"kaBra (= /b/)
T	cinco	"Tinko
D	nada	"naDa (= /d/)
s	sala	"sala
x	mujer	mu"xer
G	luego	"lweGo (= /g/)
Nasales		
m	mismo	"mismo
n	nunca	"nunka
J	año	aJo
Líquidas		
l	lejos	"lexos
L	caballo	ka"baLo
r	puro	"puro
rr	torre	"torre
Semivocales		
j	rei	rrej
	pie	pje
w	deuda	"dewDa
	muy	mwi
Vocales		
i	pico	"piko
e	pero	"pero
a	valle	"baLe
o	toro	"toro
u	duro	"duro

Tabla1. Alfabeto fonético SAMPA

Comando	**Ejemplo**
s/^h//	huerta→uerta
s/\([^c]\)h/\1/g	tahur →taur
s/^y$/i/	y→**i**
s/y$/i/	ley→le**i**
s/x/ks/g	máximo→má**ks**imo
s/j/x/g	paja→pa**x**a
s/^y/jj/g	yate→**jj**ate
s/y\([aeiou]\)/jj\1/g	Mayo→ma**jj**o
s/y/i/g	Mayka→ma**i**ka
s/ge/xe/g	general→**x**eneral
s/gi/xi/g	ginebra→**x**inebra
s/gue/ge/g	guerra→**g**erra
s/gui/gi/g	guisante→**gi**sante
s/v/b/g	vaca→**b**aca
s/ca/ka/g	laca→la**ka**
s/co/ko/g	corazón→**ko**razón
s/cu/ku/g	cuervo→**ku**ervo
s/cl/kl/g	tecla→te**kl**a
s/cr/kr/g	cráter→**kr**áter
s/que/ke/g	queso→**k**eso
s/qui/ki/g	taquilla→ta**ki**lla
s/ch/tS/g	pecho→pe**tS**o
s/w/u/g	whisky→**u**hisky
s/ce/Te/g	cereza→**Te**reza
s/ci/Ti/g	tacita→ta**Ti**ta
s/c/k/g	roca→ro**k**a
s/za/Ta/g	caza→ca**Ta**
s/zo/To/g	cazo→ca**To**
s/zu/Tu/g	tozudo→to**Tu**do
s/ze/Te/g	zener→**Te**ner
s/zi/Ti/g	herzio→her**Ti**o
s/z/T/g	azteca→a**T**teca
s/ñ/J/g	buñuelo→bu**J**uelo
s/ll/L/g	calle→ca**L**e
s/güe/gue/g	agüero→a**gue**ro
s/güi/gui/g	pingüino→pin**gui**no
s/^r\([^r]\)/rr\1/	rapaz→**rr**apaz
s/\([lmnsT]\)r\([^r]\)/\1rr\2/g	Enrique→En**rr**ique
s/ee/e/g	leed→led
s/aa/a/g	
s/ii/i/g	
s/oo/o/g	
s/uu/u/g	

Tabla2. Reglas de transcripción fonética

3.4 División en sílabas

La última etapa del proceso corresponde a la división en sílabas de las palabras ya transcritas fonéticamente. Para ello, se ha implementado en LEX[L+92] un autómata finito determinista que, a partir de una palabra invertida, obtiene las sílabas que la forman.

Las reglas del silabicador son las siguientes:

```
consonante [pbtdkgSjfTsxmnJlLr]
vocal [aeiou]
conNoRNiLNijNiS [pbtdkgfTsxmnJL]
%%


{consonante}*{vocal}+{conNoRNiLNijNiS}? {Añadir sílaba;}
{consonante}*{vocal}+r[pbtdkgf]? {Añadir sílaba;}
{consonante}*{vocal}+rr {Añadir sílaba;}
{consonante}*{vocal}+l[pbfgkt]? {Añadir sílaba;}
{consonante}*{vocal}+jj {Añadir sílaba;}
{consonante}*{vocal}+St {Añadir sílaba;}

. {;}
\n {Imprimir palabra dividida en sílabas;}
```

Como resultado de esta etapa se obtuvieron 1456067 sílabas, de las cuales 1894 eran distintas. Algunas de estas sílabas no podían darse en español pues procedían de errores ortográficos, palabras extranjeras o marcas en el texto. Por eso, tras realizarse una revisión manual de los resultados, se redujo la cifra a 1641 sílabas distintas. Es importante destacar que este conjunto de sílabas no es suficiente para construir un texto arbitrario; a menudo, las sílabas contenidas en nombres propios (de persona, poblaciones, accidentes geográficos, medicinas, ...), sólo aparecen en ese contexto, y lo mismo puede decirse de la pronunciación española de palabras extranjeras. Parece razonable entonces utilizar un inventario auxiliar, basado en unidades más pequeñas que la sílaba (con los 24 fonemas contemplados, tendríamos un total de 552 difonos) para hacer ilimitado el vocabulario del sintetizador. El uso de este inventario auxiliar también podría solventar algunos problemas de concatenación de segmentos entre palabras cuando hay fuertes efectos coarticulatorios. Aun así, el número de unidades sonoras sigue siendo razonablemente pequeño.

4 Conclusiones

Se ha tratado de motivar en este informe la elección de la sílaba como unidad de síntesis para el español, presentando algunas ideas que parecen indicar que para este idioma la calidad del sistema resultante puede incrementarse respecto de otras aproximaciones tradicionales.

El número de sílabas distintas existentes en español, aún teniendo en cuenta que se incluyen los hiatos, se mantiene pequeño (resultado que posteriormente se ha contrastado[Gue83]), permitiendo evitar uno de los mayores problemas de los sistemas basados en sílabas.

No obstante, quedan algunas dificultades por resolver, como la aparición de vocales contiguas en palabras distintas, que según nuestra aproximación pertenecerían siempre a sílabas distintas, presentando fuertes efectos coarticulatorios. Una posible solución pasaría por considerar como palabra cualquier combinación de caracteres entre dos pausas, lo que conduce, según hemos comprobado, a un gran aumento del número de unidades a considerar. Otra solución estaría basada en la utilización de un inventario auxiliar de unidades diseñado para corregir este problema, y que podría estar basado en difonos.

Los resultados presentados en este informe conducirán a la grabación de un corpus de sílabas que será aplicado a un motor de síntesis PSOLA [Dut97], lo que permitirá determinar si la calidad del sistema final aumenta.

Referencias

[B+98] A. Bonafonte et al. The UPC text-to-speech system for Spanish and Catalan. In *5th International Conference on Spoken Language Processing, ICSLP'98*, November 1998.

[D+97] Dale Dougherty et al. *Sed & Awk.* O'Reilly & Associates, 1997.

[Dut97] T. Dutoit. *An introduction to text-to-speech synthesis.* Kluwer Academic Publishers, 1997.

[F+94] Castejón F. et al. Un conversor texto-voz para español. In *Comunicaciones de Telefónica I+D*, volume 5. Telefónica I+D, Julio-Diciembre 1994.

[Gue83] Rafael Guerra. Estudio estadístico de la sílaba en español. In *Estudios de Fonética*, pages 9–112. Consejo Superior de Investigaciones Científicas, 1983.

[L+92] John R. Levine et al. *Lex & Yacc.* O'Reilly & Associates, 1992.

[LG93] Eduardo López Gonzalo. *Estudio de técnicas de procesado lingüístico y acústico para sistemas de conversión texto-voz en español basados en concatenación de unidades.* PhD thesis, Departamento de Señales, Sistemas y Radiocomunicaciones. Universidad Politécnica de Madrid, 1993.

[PHH96] T. Portele, F. Höfer, and Wolfgang J. Hess. A mixed inventory structure for German concatenative synthesis. In *Progress in Speech Synthesis*, pages 263–277. Springer Verlag, October 1996.

[Qui88] Antonio Quilis. *Fonética acústica de la lengua española.* Editorial Gredos, 1988.