RESLA, 14 (2000-2001), 7-22

LA ELABORACIÓN DE UN DICCIONARIO DE UNIDADES ACÚSTICAS PARA LA SÍNTESIS DEL HABLA. UNA APROXIMACIÓN METODOLÓGICA

LOURDES AGUILAR CUEVAS
Universitat Autònoma de Barcelona
RAFAEL MARÍN GÁLVEZ
Planeta Actimedia, Barcelona

RESUMEN. *Un sistema de conversión de texto a habla requiere el diseño y la conexión de módulos de procesamiento lingüístico y de tratamiento de señal. En aquellos sistemas basados en la concatenación de segmentos de voz, la calidad de la voz resultante depende en buena medida del diseño del diccionario de unidades acústicas que se han grabado y almacenado. En este trabajo se presentan las opciones metodológicas utilizadas más comúnmente en la constitución de un inventario de unidades acústicas para la síntesis de habla. Las etapas (definición de las unidades de síntesis, diseño del corpus de grabación, selección del locutor) se abordan teniendo en cuenta la descripción del sistema fonético y fonológico del español, de tal modo que se demuestra que aunque sea una tarea eminentemente aplicada debe conjugarse con la reflexión teórica.*

PALABRAS CLAVE. *Síntesis por concatenación de unidades, difonemas, base de datos de unidades acústicas.*

ABSTRACT. *Text-to-speech synthesis involves the computation of a speech signal from input text. Accomplishing this requires a system that consists of several components, from linguistic analysis to speech coding. In concatenative synthesis, the quality is highly dependent on the source units. In this article, an overview of the methodological options in the task of creating an acoustic inventory is presented. The topics introduced (the size and type of the units stored, the structure of the recorded database, the selection of the speaker) are tackled from the knowledge of the Spanish phonetics and phonology, showing that an applied task must be guided by the theoretical reflection.*

KEYWORDS. *Concatenative synthesis, diphones, acoustic units database.*

1. INTRODUCCIÓN

El ámbito de estudio de las denominadas tecnologías del habla, fundamentalmente aplicado, y dentro de él, la conversión de texto a habla, constituye un área de investigación que ha despertado –y sigue haciéndolo– un interés generalizado en disciplinas tan (aparentemente) dispares como la ingeniería, la informática o la lingüística.

En este trabajo se revisan las opciones metodológicas utilizadas más comúnmente en la constitución de un diccionario de unidades acústicas para la síntesis de habla, a la vez que se describe un inventario de unidades para el español[1] (en adelante, DIU). Un diccionario de unidades de síntesis puede concebirse como el conjunto de señal almacenada necesaria para generar habla mediante un sistema de conversión de texto a habla basado en la concatenación de segmentos acústicos.

Habitualmente, la elaboración de un diccionario de unidades de síntesis incluye las siguientes etapas: 1) el diseño de un inventario de unidades; 2) la definición de un corpus de grabación; 3) la selección del locutor y la grabación del corpus, y 4) la segmentación, etiquetado y almacenamiento de las unidades del corpus que serán utilizadas para producir habla una vez asociadas con el resto de módulos del sistema, y principalmente con el de transcripción fonética.

Por lo general, en el desarrollo de un sistema de conversión de texto a habla se distingue entre la tarea responsable del procesamiento lingüístico y la encargada del tratamiento de la información acústico-fonética. En el caso de que el procedimiento de síntesis se base en la concatenación de segmentos de voz, el éxito de esta segunda tarea depende en buena medida del diseño del diccionario de unidades, diseño que ha de estar fundamentado teóricamente, como pretendemos demostrar a lo largo de este trabajo. En relación con esto, cabe recordar que, como cualquier tarea que incorpore conocimientos fonéticos a las tecnologías del habla, el desarrollo de un diccionario responde al compromiso entre los condicionamientos técnicos y la adecuación a principios lingüísticos generales (véase Llisterri, Aguilar y Garrido 1997).

2. DISEÑO DE UN INVENTARIO DE UNIDADES

El diseño de un inventario de unidades acústicas requiere la definición y selección de: 1) las unidades de síntesis; y 2) los alófonos.

2.1. *Definición de las unidades de síntesis*

La elección del tipo de unidades de síntesis viene determinada básicamente por criterios relativos al coste de almacenamiento, la flexibilidad en el procesamiento de texto no restringido, y la calidad del habla obtenida (para una revisión, cf. Llisterri 1988). Así, por ejemplo, mediante la grabación de unidades largas como las palabras se consigue un sistema con un nivel óptimo de calidad, pero con un coste de almacenamiento muy elevado y sin posibilidad de generar palabras que no hayan sido guardadas previamente. En

el otro extremo, la utilización del fonema (o del alófono) como unidad acústica supone un gran ahorro en el almacenamiento, a la vez que permite convertir en voz cualquier secuencia de texto; sin embargo, la calidad del habla dista mucho de ser aceptable: al ser la transición entre sonidos zonas acústicamente inestables, la concatenación en sus límites resulta muy difícil.

Elementos como el difonema, a medio camino entre la palabra y el alófono, pese a estar alejados de los enfoques lingüísticos al uso, se convierten en las unidades de síntesis más utilizadas (véase Rodríguez *et al.* 1993; Böeffard *et al.* 1993; Tzoukermann 1994). Se denomina difonema al segmento que abarca desde la mitad de la zona estable de un sonido hasta la mitad de la zona estable del siguiente: de esta manera, la transición entre los dos sonidos queda contenida en la unidad y se evitan problemas de concatenación. Veamos en (1) la segmentación de una frase en difonemas (para una mejor comprensión del ejemplo, por el momento hacemos caso omiso de las diferencias alofónicas).

(1) Malena tiene pena de bandoneón
SILENCIO-M / MA / AL / LE / EN / NA / AT / TI / IE / EN / NE / EP / PE / EN / NA / AD / DE / EB / BA / AN / ND / DO / ON / NE / EO / ON / N-SILENCIO

Los difonemas se obtienen a partir de la combinación de las categorías 'vocal', 'consonante' y 'silencio' del siguiente modo: vocal-vocal, vocal-consonante, consonante-vocal, consonante-consonante, silencio-vocal, vocal-silencio, silencio-consonante, consonante-silencio. La categoría 'silencio', a pesar de no haber sido descrita lingüísticamente, es necesaria para marcar el inicio y el fin de las secuencias de habla.

Además de los difonemas, en un diccionario de unidades de síntesis se pueden incluir polifonemas (fragmentos de voz que abarcan más de dos segmentos), con el fin de evitar la concatenación en las fronteras entre los sonidos breves o acústicamente inestables, como las consonantes aproximantes o la vibrante simple, en el caso del español. También se han propuesto inventarios mixtos, integrados por demisílabas y difonemas (Portele, Höfer y Hess 1997) o el uso de unidades distintas según el contexto prosódico (Campbell y Black 1997).

En el diseño de DIU, se parte del difonema como unidad básica de síntesis, si bien se incluyen también polifonemas, integrados por las combinaciones vocal-aproximante-vocal, y los grupos consonánticos homosilábicos del español.

2.2. *Selección de los alófonos*

Como hemos mencionado, los difonemas se forman a partir de la combinación de las categorías 'vocal', 'consonante' y 'silencio', pero antes es necesario definir qué segmentos se incluyen en las categorías 'vocal' y 'consonante'.

Para la constitución del inventario de alófonos resulta necesario, en primer lugar, identificar aquellos segmentos fónicos que en una lengua dada pueden afectar en mayor medida a la calidad del habla generada por el sistema, aunque no incidan en la inteligibilidad (por ejemplo, la diferenciación entre hiatos y diptongos en español).

En segundo lugar, debe establecerse una relación correcta entre la unidad acústica (segmento de voz etiquetado y almacenado en el diccionario) y la unidad fonética (representación fonética propuesta por el módulo de transcripción de grafía a fonema). En este sentido, y siguiendo con el problema de los hiatos y diptongos, si no se ha diseñado el sistema de conversión de texto a habla teniendo en cuenta las interacciones entre los diferentes módulos, podemos encontrarnos con dos situaciones anómalas: en una, disponemos de unidades acústicas distintas para cada diptongo y cada hiato, pero el transcriptor no es capaz de discriminarlos a partir del texto –piénsese en piezas léxicas como *du.e.to* y *due.lo*–; en la otra, el transcriptor dispone de un diccionario de excepciones que permite reconocer los hiatos no marcados ortográficamente, pero en el diccionario de unidades acústicas no se ha considerado la diferencia entre grupos vocálicos pertenecientes a una misma sílaba y grupos vocálicos separados por un margen silábico. Ni en un caso ni en otro el tratamiento de los hiatos y diptongos será adecuado.

Empecemos por establecer el inventario de alófonos vocálicos y consonánticos del español para luego seleccionar aquellos que formarán parte del inventario de segmentos del diccionario de unidades.

En cuanto a los fonemas vocálicos, Navarro Tomás (1918) describe las realizaciones recogidas en la Tabla 1. No incluimos las variantes relajadas, que aparecen normalmente en posición final, o entre dos acentos, por pertenecer al habla coloquial o familiar, registro que no se suele incorporar en los sistemas de conversión de texto a habla. Tampoco hacemos referencia a las vocales tónicas o átonas, pero basta mencionar que cualquiera de los fonemas vocálicos /i e a o u/ puede aparecer en sílaba acentuada o inacentuada.

Fonema	Alófono	Contexto de aparición
i	cerrado	En sílaba libre
	abierto	En sílaba trabada Antes y después de [r] Antes de [x]
	semivocal	En la posición final de un diptongo
	semiconsonante	En la posición inicial de un diptongo
e	cerrado	En sílaba libre En sílaba trabada por [m] [n] [s] y [d] Seguido de "x"
	abierto	Seguido y precedido de [r] Delante de [x] En el diptongo [ei] En sílaba trabada por cualquier consonante que no sea [m], [n], [s], [d] y [s] o seguida de "x"
a	medio	En sílaba acentuada
	palatal	Ante consonantes palatales En el diptongo [ai]

Fonema	Alófono	Contexto de aparición
	velar	En el diptongo [au] y ante una [u] acentuada Ante la vocal [o] En sílaba trabada por [l] Delante de [x]
o	cerrado	En sílaba libre acentuada
	abierto	En contacto con [r] Delante de [x] En el diptongo [oi] En sílaba trabada por cualquier consonante En posición acentuada entre una [a] precedente y una [r] o una [l]
u	cerrado	En sílaba libre acentuada
	abierto	En contacto con [r] Delante de [x] En sílaba trabada
	semivocal	En la posición final de un diptongo
	semiconsonante	En la posición inicial de un diptongo

Tabla 1. *Fonemas vocálicos, realizaciones alofónicas y contextos de aparición*

El inventario de alófonos consonánticos del español es el expuesto en la Tabla 2 a partir de la descripción de Navarro Tomás (1918) y Machuca (2000).

SONIDOS	Labial	Labiodental	Dental	Alveolar	Palatal	Velar
Sonoridad	+ -	+ -	+ -	+ -	+ -	+ -
Oclusivas	b p		d̪ t̪+t̪		ɟ	g k
Aproximantes	β		ð̞	ɹ	j	ɣ w
Fricativas		f̞ f	θ̞ θ	s̞ s	ʝ̞	x̞ x
Africadas					ʤ ʧ	
Nasales	m	ɱ	n̪- n̪+	n	ɲ	ŋ
Laterales			l̪-l̪+	l	ʎ	
Vibrantes				ɾ - r		

Tabla 2. *Alófonos consonánticos del español*

Parece claro que incluir todos los alófonos –quince vocálicos y cuarenta consonánticos– generaría un diccionario de unidades demasiado extenso, con 3.135 unidades, según aparece en la Tabla 3.

vocal-vocal	15 x 15	225
vocal-consonante	15 x 40	600
consonante-vocal	40 x 15	600
consonante-consonante	40 x 40	1.600
silencio-vocal	1 x 15	15
vocal-silencio	15 x 1	15
silencio-consonante	1 x 40	40
consonante-silencio	40 x 1	40
Total		3.135

Tabla 3. *Posible inventario de difonemas*

Por otro lado, además de las restricciones de tamaño del inventario que pueden imponer las aplicaciones, debemos considerar un problema de adecuación teórica: la simple combinación cartesiana de los alófonos vocálicos con los consonánticos no responde a la realidad de la lengua, ya que muchas de las variantes fonéticas son contextuales. Por citar un caso, la nasal labiodental [ɱ] únicamente puede aparecer precediendo a una consonante labiodental ([f]), por lo que son imposibles en español las combinaciones restantes. Puede argumentarse en contra de este enfoque que las restricciones fonotácticas se dan generalmente en el marco de la palabra y que muchas de las combinaciones imposibles en ese dominio se dan entre palabras, especialmente si aparecen nombres extranjeros. No obstante, pretender dar cuenta de todas esas posibilidades es tarea vana, dado que la procedencia lingüística de las palabras foráneas es diversa y variable, y como consecuencia, el tipo de problemas que ocasionan, difícilmente abarcable en su totalidad; sin olvidarnos de que se necesitaría incluir alófonos extranjeros en el inventario.

Llegados a este punto, cabe mencionar que la preocupación por el tratamiento de los nombres extranjeros, muy presente en el campo de la síntesis de habla, donde cualquier texto debe poder ser oralizado, no es exclusiva de esta área de investigación y desarrollo: de hecho, los problemas de locución en los medios de comunicación audiovisuales son similares. Los libros de estilo recomiendan pronunciar las palabras y los nombres extranjeros tal como se pronuncian en sus lenguas originarias, por lo menos en lo que se refiere a las más frecuentes (inglés, francés) y a las otras lenguas oficiales en el territorio español (Radio Televisión Madrid 1993; Mendieta 1993; Tubau 1993). En otras ocasiones, en cambio, se recomienda ajustar la pronunciación a las reglas fonéticas del español, intentando que ésta se parezca a la original, como es el caso en Grijelmo (1998).

Si adoptamos la recomendación de hispanizar los nombres extranjeros en la salida de un conversor de texto a habla, no será necesario incorporar alófonos extranjeros y se podrán respetar las restricciones de distribución. De este modo, *squash* se pronunciará como 'esquach', y *Generalitat* como 'yeneralitat', siendo la hispanización tarea de las

reglas de transcripción grafía-fonema. En el caso de que se quiera respetar la pronunciación extendida de ciertas palabras con sonidos ajenos al sistema fonético español, puede optarse por la grabación de algunas unidades acústicas con estos sonidos o bien se puede pensar en un diccionario de excepciones que contenga las palabras más frecuentes grabadas en su totalidad.

En el inventario que describimos aquí, no se incluyen alófonos ajenos al sistema fonético español, y se deja que sean otros módulos del sistema de conversión quienes decidan la solución correcta en cada caso.

Pero aun esquivando el problema de los nombres extranjeros, quedan otras cuestiones por examinar: nos referimos a las glides[2] y a los procesos relativos al modo de articulación, punto de articulación y sonoridad en las consonantes del español. En cada caso, recordaremos la clasificación y distribución de los sonidos establecidas en manuales de descripción fonética del español (Navarro Tomás 1918; Gil 1988; Machuca 2000), marco de referencia para argumentar las elecciones en la constitución del diccionario.

2.2.1. *Las glides*

Denominamos glides a los segmentos vocálicos que forman parte de un diptongo y que no ocupan la posición de núcleo silábico, como los elementos resaltados en (2).

(2) a. t*i*erra, c*u*atro
b. pa*i*sano, ca*u*sa

En la tradición fonética del español, se conoce como semiconsonsonante el segmento /i u/ que aparece en posición inicial del diptongo, (2a), y como semivocal, el que lo hace en posición final, (2b).

Al margen de la discusión sobre la naturaleza fonémica de las glides en español (véase una revisión de las diferentes interpretaciones en Aguilar, en prensa), estos segmentos plantean dos problemas para un conversor de texto a habla: la distinción entre hiato y diptongo, y la consonantización de las semiconsonantes en posición de ataque silábico.

Por un lado, si bien debemos convenir que la distinción hiato-diptongo está léxicamente condicionada –nada permite diferenciar entre el diptongo de *duelo* y el hiato de *dueto*–, produce diferencias significativas en palabras como: *ahí/hay*, *pie/pié/píe*. Además, es un factor que incide en gran medida en la percepción como natural de una voz. Acústicamente, las diferencias entre hiato y diptongo están establecidas (Ren 1986; Aguilar 1999) pero introducir las glides en el inventario aumenta considerablemente el tamaño final de la base de datos.

La solución adoptada en DIU consiste en incluir combinaciones vocálicas en hiato y diptongos en la categoría de difonemas vocal-vocal. La selección correcta de un tipo de combinación vocálica u otra corresponde al módulo de transcripción grafía-fonema del sistema.

Por otro lado, no en todos los estudios del sistema fónico del español se consideran [j] y [w] como alófonos. Al contrario, en diversos estudios fonológicos se entiende

que las variantes consonánticas que aparecen en palabras como *hierro* o *huerta* están fonológicamente emparentadas con las glides: las glides se marcan léxicamente dado su carácter impredecible, y las variantes consonánticas se derivan de éstas mediante una regla de consonantización (Harris 1969; Harris 1989; Hualde 1991).

Sin embargo, al menos en lo que respecta al campo de la síntesis de habla, es preferible considerar como alófonos consonánticos las realizaciones [j w], opción adoptada en DIU. La similitud fonética entre la consonante aproximante palatal y la glide palatal, y la consonante aproximante labiovelar, y la glide velar –traducida en los dobletes ortográficos del tipo *hierro/yerro*– no es motivo suficiente para derivar las unas de las otras. En Aguilar (en prensa) se aducen varios argumentos: las consonantes se comportan de manera diferente a las glides frente a fenómenos de asimilación tales como la sonorización de [s] y la asimilación de punto de articulación de nasales y de laterales (*desierto/deshielo, desuello/deshueso*); no todas las palabras en las que aparece [j] o [w] se relacionan formalmente con otras en las que aparezca una glide o una vocal (*mayo, alcahueta*); entre otros motivos.

En el inventario de alófonos de DIU no se incluye en cambio la consonante africada sonora, a la que en las descripciones del español se le atribuye una relación de distribución complementaria con [j]: la variante africada aparece en las llamadas posiciones fuertes, es decir, en posición inicial absoluta y tras [n] o [l] (Navarro Tomás 1918; Canellada y Madsen 1987), y la aproximante (considerada fricativa en los estudios citados) en el resto de contextos. La razón de no incluir dicho alófono tiene que ver con la variedad de pronunciación de [ʝ] en función de los estilos de habla. La variante africada es un reforzamiento articulatorio, que se da en determinados contextos y para ciertos hablantes, pero que aparece al lado de otras variantes, como la oclusiva palatal, la fricativa o la aproximante (Aguilar 1997; véase una nueva aproximación en Martínez Celdrán y Fernández Planas 2001).

2.2.2. *Procesos fonológicos relativos al modo de articulación*

En español, las consonantes /b/, /d/, /g/ en posición de ataque silábico se realizan como oclusivas sonoras en posición inicial absoluta y detrás de una consonante nasal; en el caso de /d/, también cuando es precedida por una consonante lateral. En el resto de posiciones se realizan fonéticamente como aproximantes.

Este proceso fonológico plantea, con respecto al inventario de alófonos, la disyuntiva de elegir entre dos posibilidades: a) incluir el alófono oclusivo y el alófono aproximante como unidades; b) establecer un representante de los dos alófonos.

En el primer caso, se adopta un enfoque fonético en la selección de unidades del inventario, asegurando una correcta generación de las consonantes oclusivas sonoras y aproximantes; es decir, para obtener una palabra como *baba* se concatenarán los difonemas *silencio-b, b-a, a-β, β-a, a-silencio*. El problema de este tratamiento radica en que, al considerar los alófonos oclusivos sonoros y los aproximantes como unidades del inventario, tanto unos como otros deberán combinarse con el resto de los alófonos del

español, y aparecerán las restricciones fonotácticas antes ejemplificadas con las consonantes nasales: es decir, encontraremos una [m] combinada con una [β].

En el segundo caso, se asume que las diferencias entre las variantes oclusivas y aproximantes aparecen de forma natural al ser pronunciadas las palabras del corpus. Dicho de otro modo, tales diferencias ya estarán presentes en el fragmento de voz grabado y almacenado. Esta solución reduce el número de alófonos del inventario, y por tanto, el número final de difonemas. Sin embargo, plantea el problema de unir segmentos que son acústicamente diferentes: siguiendo con el ejemplo anterior, para generar *baba*, se concatenarán *silencio-B, B-a, a-B, B-a, a-silencio*, siendo la primera B el segmento inicial de una consonante oclusiva, y la segunda B el segmento final de una consonante aproximante. La estructura acústica de ambas consonantes es suficientemente distinta para aconsejar el abandono de este enfoque (véase Machuca 1997).

En el diccionario de unidades DIU, se ha optado por incluir las consonantes oclusivas sonoras [b d g] y las consonantes aproximantes [β δ γ] aunque no en los mismos contextos, sino respetando las restricciones fonotácticas de la lengua: corresponde al módulo de transcripción grafía-fonema seleccionar la variante contextual correcta en cada caso.

2.2.3. *Procesos de asimilación del punto de articulación*

En posición de distensión silábica, la consonante nasal toma el punto de articulación de la consonante siguiente: así hallamos nasales bilabiales (*enviar*), labiodentales (*énfasis*), interdentales (*onza*), dentales (*cantar*), palatales (*concha*) y velares (*anca*).

Del mismo modo, la lateral /l/ presenta diferentes alófonos en función del punto de articulación de la consonante siguiente, si bien la asimilación en este caso excluye las zonas de articulación extrema: labial (bilabial y labiodental) y velar (velar y labiovelar). Por consiguiente, encontramos una realización interdental (*alzar*), dental (*alta*) y palatal (*colcha*).

Por otro lado, en posición final de sílaba, la /s/ puede dentalizarse o interdentalizarse cuando aparece ante una consonante dental (*hasta*) o interdental (en *asceta*), respectivamente.

A la hora de seleccionar los alófonos para el diccionario, vuelve a aparecer la disyuntiva planteada en el epígrafe anterior: a) considerar cada una de las variantes como unidad del inventario; b) considerar un representante para las variantes nasales, uno para las laterales y uno para la fricativa alveolar. En el segundo caso, las variaciones debidas a los procesos asimilatorios aparecerán en la pronunciación del difonema correspondiente: así, el alófono [l] del difonema *l-t* se realizará como dentalizado, el alófono [n] del difonema *n-g* se velarizará, y el alófono [s] del difonema *s-z* se interdentalizará.

Ahora bien, a diferencia de los alófonos oclusivos y aproximantes, las distinciones relativas al punto de articulación afectan en grado menor a la estructura acústica de los sonidos, por lo que el problema de concatenar un difonema *a-l* y *l-t*, siendo alveolar la primera *l* y dental la segunda puede solventarse más fácilmente. Adoptando pues la opción (b), el tamaño del inventario no se verá incrementado en exceso.

Como resultado de las consideraciones aducidas, en el diccionario DIU, no se han incluido las variantes asimiladas de las consonantes nasales, laterales ni fricativas.

2.2.4. *Procesos de asimilación de la sonoridad*

Como es sabido, las consonantes fricativas sordas se sonorizan en posición final de sílaba cuando preceden a una consonante sonora (*Lisboa, Afganistán, juzgar, reloj de arena*), y las diferencias acústicas radican en la presencia de zonas de frecuencia bajas que se asemejan en su configuración a los formantes vocálicos.

Para estas variantes, hemos considerado, al igual que en el caso de las asimiladas en el punto de articulación, que el alófono [s] se sonoriza cuando forma parte del difonema *s-δ* y que es posible concatenar un alófono *vocal-s* y *s-δ* sin gran pérdida de calidad, pese a las diferencias de sonoridad.

Por todo ello, las variantes sonoras de las consonantes fricativas no se han incorporado en el inventario de unidades de DIU.

2.3. *Inventario de alófonos*

De acuerdo con los criterios establecidos a lo largo del anterior apartado, el inventario de alófonos está constituido por 31 unidades, a saber:

—1 alófono de silencio

—7 alófonos vocálicos: [a], [e], [i̯] [i], [o], [u̯], [u].

—23 alófonos consonánticos: [p], [t], [k], [b], [d], [g], [β], [ð], [ɣ], [ʝ], [w], [m], [n], [ɲ], [l], [ʎ], [ɾ], [r], [f], [θ], [s], [x], [ʧ].

2.4. *Inventario de unidades de síntesis*

El inventario de unidades de síntesis en DIU no consiste en la combinación de cada uno de los alófonos con el resto, sino que se han incorporado algunas de las restricciones distribucionales de los segmentos fónicos del español.

Por otro lado, los trifonemas y cuatrifonemas se conciben como parte del inventario, no como un complemento, de ahí que no se repitan secuencias en un formato y otro (King, Portele y Hofer 1997).

2.4.1. *Conjunto de difonemas*

- Vocal-Vocal: 100

Las combinaciones vocal-vocal incorporan la variable del acento con el fin de disponer de secuencias homosilábicas y heterosilábicas.

- Vocal-Consonante: 85
- Consonante-Vocal: 85

Se consideran 17 de los alófonos consonánticos descartando las consonantes aproximantes y la vibrante simple, que se encuentran en el conjunto de polifonemas.

- Consonante-Consonante: 361

Se ha respetado la distribución de las variantes fonéticas de /b, d, g/ y de /r, ɾ/. La misma solución se ha adoptado en las combinaciones entre 'silencio' y 'consonante'.

- Silencio-Consonante: 19
- Consonante-Silencio: 19

2.4.2. *Conjunto de trifonemas*

- Vocal-Consonante aproximante-Vocal: 105
- Vocal-[ɾ]-Vocal: 25
- Vocal-[ɾ]-Consonante: 95
- Consonante oclusiva-[ɾ]-Vocal: 30
- [f]-[ɾ]-Vocal: 5
- [p, k, b, g]-[l]-Vocal: 20
- [f]-[l]-Vocal: 5

2.4.3. *Conjunto de cuatrifonemas*

- Vocal-[β δ γ]-[ɾ]-Vocal: 75
- Vocal-[β δ]-[l]-Vocal: 50

En total, el inventario está constituido por 1.079 unidades de síntesis.

3. DISEÑO DE UN CORPUS DE GRABACIÓN

Una vez diseñado el inventario de unidades para la síntesis, ha de elaborarse un corpus de grabación, definido como el conjunto de secuencias que el locutor va a oralizar y de donde posteriormente se van a extraer las unidades acústicas que, como ya hemos mencionado, servirán para generar habla en un sistema de síntesis por concatenación.

Por un lado, hay que decidir el entorno en el que aparece el difonema: entre otras posibilidades, palabras, frases, textos. Por otro, es necesario decidir si los contextos han de ser idénticos para todas las unidades, es decir, si la vocal o consonante que precede o sigue a la unidad de síntesis debe ser la misma en todo el inventario; y si se requieren unas pautas de lectura del corpus para controlar, por ejemplo, las variaciones prosódicas del locutor.

3.1. *Entorno de la unidad*

Las unidades de síntesis pueden extraerse de un corpus de textos leídos, como se describe por ejemplo en Ding y Campbell (1997). Ahora bien, ya que en el enfoque metodológico que estamos describiendo no se parte de una grabación previa, vamos a centrarnos en la posibilidad de utilizar palabras (según el concepto fonológico) o logatomas (secuencias que aceptan cualquier combinación de sonidos independientemente de su plausibilidad fonológica).

Si escogemos la palabra como contexto de aparición de los difonemas, encontramos al menos dos inconvenientes: primero, ciertas combinaciones imposibles en el inte-

rior de palabra pueden darse en las fronteras; segundo, determinadas combinaciones posibles en la lengua no aparecen en ninguna palabra: son los llamados "vacíos léxicos", formas que respetan los condicionamientos fonológicos pero que por motivos idiosincrásicos no se han incorporado al caudal léxico de la lengua.

Una buena solución para sortear los obstáculos mencionados es integrar los ejemplos en un conjunto de frases, de tal modo que algunos difonemas se extraigan del margen de palabras. No obstante, a pesar de presentar la ventaja de una lectura poco forzada, la combinación de contextos intraléxicos e interléxicos puede ocasionar problemas, dado que ciertos procesos fonológicos tienen comportamientos diferentes en función del dominio de aplicación (Mohanan 1986).

La segunda opción, esto es, el empleo de logatomas, conlleva todos los inconvenientes derivados de la falta de naturalidad como, por ejemplo, la dificultad de pronunciación por parte del locutor y la consiguiente lectura hiperarticulada, pero es un buen método para obtener todas las combinaciones de alófonos, siendo además posible igualar el contexto de aparición de la unidad de síntesis.

3.2. *El control de la estructura de la forma léxica*

Por control de la estructura de la forma léxica entendemos la homogeneización de la longitud de la palabra, la posición del acento, el contexto fonético y el lugar de la unidad en la palabra para todas las piezas del corpus. Como resultado, se obtiene una regularidad rítmica que puede ser tediosa para el lector.

Parece claro que renunciando a dicho control, la lectura resulta menos artificial, pero en contrapartida, los entornos de las unidades de síntesis no son homogéneos, además de que la naturalidad conlleva variaciones de frecuencia, intensidad y duración poco deseables para el procesamiento acústico posterior de la señal. Por el contrario, si se restringen las posibilidades de configuración de la unidad, los contextos de extracción serán semejantes y el locutor podrá mantener una velocidad de elocución y un tono de voz constante.

La posibilidad de automatización del proceso de obtención de datos, con la consiguiente facilidad para crear nuevos diccionarios de voces, rige en muchas ocasiones las decisiones sobre el corpus final de grabación. En el desarrollo de sistemas de conversión de texto a habla para diferentes lenguas, se encuentran ejemplos de distintos métodos: Boëffard *et al.* (1993) extraen las unidades de síntesis de logatomas para desarrollar el sistema de conversión de texto a habla multilingüe del CNET; Rodríguez *et al.* (1993) optan por las palabras con sentido en la constitución del diccionario de Telefónica I+D; Tzoukermann (1994), por su parte, graba las unidades tanto en logatomas como en palabras y elige a continuación las mejores unidades para la versión francesa del sistema multilingüe de AT&T.

Para DIU se ha elaborado un corpus de logatomas que incluye todas y cada una de las unidades de síntesis. Se han impuesto, además, unas restricciones de aparición de dichas unidades en lo que se refiere al contexto precedente y siguiente del difonema o polifonema, el número de sílabas de la forma y el lugar de aparición del acento léxico.

Las unidades de síntesis aparecen en las sílabas interiores de palabras trisílabas, a menos que se hayan combinado con la categoría 'silencio'. El acento de palabra recae en la penúltima sílaba, reflejando así la tendencia a las palabras llanas del español, descrita en los tratados de gramática. Las consonantes que integran las sílabas precedente y siguiente a la unidad de síntesis son oclusivas sordas.

4. SELECCIÓN DEL LOCUTOR Y CRITERIOS DE GRABACIÓN

La elección del locutor que va a grabar el corpus, esto es, la voz del sistema de conversión de texto a habla, es una etapa importante en la constitución del diccionario, dado que de ella depende en buena medida la aceptabilidad del sistema por parte del usuario.

Para realizar esta selección habitualmente se acude a pruebas de generación de habla sintética con voces de diferentes sujetos. La decisión final suele basarse, además, en las preferencias manifestadas por un grupo de oyentes en un test de percepción y en las impresiones perceptivas de los propios investigadores.

En el caso que nos ocupa, los locutores interesados en grabar el corpus leyeron un fragmento de texto, y los logatomas (según los criterios especificados anteriormente relativos al inventario de alófonos, de unidades de síntesis y estructura de los logatomas) necesarios para generar las frases. Se segmentaron las unidades, se concatenaron siguiendo el sistema descrito en Conkie e Isard (1997) y a la voz sintética se le superpuso los valores prosódicos de duración y de frecuencia fundamental de los segmentos de las frases leídas. La selección se basó en los juicios de cuatro investigadores habituados a oír voz sintética.

Para la grabación se contó con dos locutores masculinos y uno femenino.

En cuanto a los criterios de grabación del corpus, se puede pensar en una sesión dirigida, donde el locutor recibe instrucciones de lectura, o en una sesión no dirigida, en la que el locutor realiza los sonidos del corpus de acuerdo con sus intuiciones lingüísticas, y siguiendo las convenciones ortográficas de la lengua.

En el primer caso, se obtiene una total correspondencia entre las unidades del corpus y los fragmentos de voz grabados, mientras que en el segundo, se considera que cualquier proceso fonético que afecte a la unidad de síntesis estará incluido en su pronunciación y que, por tanto, el resultado será más natural.

En las sesiones de grabación[3], se instó al locutor a seguir unas pautas de lectura muy precisas, con el fin de asegurar la realización fonética prevista para cada una de las unidades.

5. CONCLUSIONES

Hasta aquí hemos descrito el diseño de un inventario de unidades acústicas para un sistema de síntesis basado en la concatenación. Las etapas de selección del inventario y del corpus de grabación se han abordado teniendo en cuenta las explicaciones lingüísticas de que disponemos, además de la solución adoptada en cada caso para el diccionario-

rio DIU. Así, por ejemplo, la distinción entre hiatos y diptongos, y el tratamiento de las variaciones fonéticas contextuales se han ubicado en el contexto general de los procesos fonológicos del español. De este modo, podemos demostrar que es posible conjugar el desarrollo de una tarea eminentemente aplicada con su fundamentación teórica.

No obstante, es necesario mencionar que este artículo no abarca en su totalidad la constitución de un diccionario de unidades de síntesis, ya que éste no está completo hasta que se ha procesado acústicamente, tarea que pasa normalmente por las etapas de colocación de marcas de segmentación entre los sonidos que forman parte de la unidad de síntesis[4] y de conexión con un determinado sistema de concatenación de unidades[5].

Por último, una vez obtenido el diccionario completo de unidades de síntesis, es decir, una vez almacenados los segmentos útiles para la generación de habla, con las informaciones necesarias asociadas, se requiere una evaluación de la calidad de la voz sintética obtenida (véase Aguilar *et al.* 1994ab).

NOTAS

1. La estructura del diccionario de unidades acústicas que se va a describir en el artículo es el resultado del trabajo en común de Lourdes Aguilar, Alistair Conkie, David Casacuberta y Rafael Marín (en orden alfabético) en el *Departament de Filologia Espanyola* de la *Universitat Autònoma de Barcelona* durante el periodo comprendido entre el 1 de enero y el 1 de julio de 1996, con el objetivo de disponer de herramientas de dominio público para la síntesis multilingüe del habla. De la grabación completa del inventario con una voz masculina, se procesó la base de datos que se distribuye con el sistema de síntesis MBROLA4, disponible en http://tcts.fpms.ac.be/synthesis/mbrola.html. La cesión está garantizada por una licencia firmada por un representante legal de la UAB y uno de la *Faculté Polytechnique de Mons*, de la que el *Departament de Filologia Espanyola* dispone de una copia, el *Laboratoire de Théorie des Circuits et de Traitement du Signal, Faculté Polytechnique de Mons*, de otra, y la *Oficina de Transferència Tecnològica*, de la *Universitat Autònoma de Barcelona*, de la tercera.
   Las tareas de procesamiento acústico fueron realizadas por Alistair Conkie, en aquellos momentos investigador en la UAB, y Vincent Pagel, de la *Faculté Polytechnique de Mons*. La información sobre la versión procesada puede consultarse en los documentos que acompañan a la base de datos para su distribución a través del proyecto de síntesis multilingüe MBROLA, que se lleva a cabo en el TCTS, *Laboratoire de Théorie des Circuits et de Traitement du Signal, Faculté Polytechnique de Mons*; las especificaciones no coinciden necesariamente con lo descrito aquí, al haberse incorporado a un sistema de síntesis concreto.
   Para obtener información sobre el inventario completo, o la disponibilidad de otras voces, son, indistintamente, personas de contacto: Alistair Conkie (adc@cstr.ed.ac.uk), Lourdes Aguilar (Lourdes.Aguilar@uab.es), Rafael Marín (rafa@sumi.es), David Casacuberta (David.Casacuberta@uab.es).
2. Utilizamos el término *glide,* recogido en el *Diccionario de lingüística*, R. Cerdà, para hablar indistintamente de semiconsonantes y semivocales.
3. La grabación tuvo lugar en las instalaciones de la *Universitat Politècnica de Catalunya*, y se utilizó un laringógrafo, del *Laboratori de Fonètica* de la *Universitat Autònoma de Barcelona*, para obtener la señal laringográfica.
4. Los esfuerzos se dirigen a conseguir la automatización del proceso de segmentación de unidades, si bien para ciertos elementos fonéticos se precisa una corrección manual (Boëffard *et al.* 1993; Ljolje, Hirchsberg y van Santen 1997; Whightman y Talkin 1997). En el caso del inventario que describimos y en la base de datos derivada para el proyecto MBROLA, no se utilizaron técnicas automáticas: usando las herramientas que ofrece el sistema Waves+, instalado en una SunSparc del *Departament de Filologia Espanyola* de la UAB, Alistair Conkie hizo una primera segmentación, que fue revisada posteriormente por Lourdes Aguilar.
5. En el caso de la versión que distribuye el proyecto MBROLA, la información sobre el sistema aparece en las páginas del mismo proyecto (cf. Dutoit 1997).

REFERENCIAS

Aguilar, L., J. M. Fernández, J. Garrido, J. Llisterri, A. Macarrón, L. Monzón y M. A. Rodríguez. 1994a. "Evaluation of a Spanish Text-to-Speech System". *Conference Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis.* 207-210.

Aguilar, L., J. M. Fernández, J. Garrido, J. Llisterri, A. Macarrón, L. Monzón y M. A. Rodríguez. 1994b. "Diseño de pruebas para la evaluación de habla sintetizada en español y su aplicación a un sistema de conversión de texto a habla". *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural.* Córdoba, 20-22 de julio de 1994.

Aguilar, L. 1997. *De la vocal a la consonante.* Santiago de Compostela: Servicio de Publicaciones de la Universidad de Santiago de Compostela.

Aguilar, L. 1999. "Hiatus and Diphthong: Acoustic Cues and Speech Differences". *Speech Communication* 28 (1): 57-74.

Aguilar, L. en prensa. "A vueltas con el problema de las semiconsonantes y las semivocales". *Verba.*

Boëffard, O., B. Cherbonnel, F. Emerard, S. White. 1993. "Automatic segmentation and quality evaluation of speech unit inventories for concatenation-based, multilingual PSOLA text-to-speech systems". *Proceedings of Eurospeech'93.* 1449-1452.

Campbell, N. y A. Black. 1997. "Prosody and the Selection of Source Units for Concatenative Synthesis". Eds. J. P. H. van Santen, R. W. Sproat, J. P. Olive, J. Hirschberg. *Progress in Speech Synthesis*, Nueva York: Springer-Verlag. 279-292.

Canellada, M. J. y J. K. Madsen. 1987. *Pronunciación del español.* Madrid: Castalia.

Cerdà, R. 1986. *Diccionario de lingüística.* Madrid: Anaya.

Conkie, A. y S. Isard. 1997. "Optimal coupling of diphones". *Progress in speech synthesis.* Eds. van Santen, J., Sproat, R., Olive, J. y J. Hirschberg. Springer Verlag, 293-305.

Ding, W. y N. Campbell. 1997. "Optimising Unit Selection with Voice Source and Formants in the CHATR Speech Synthesis System". *Proceedings Eurospeech'97.* vol. 2: 537-540.

Dutoit, T. 1997. *An Introduction to Text-to-Speech Synthesis.* Dordrecht: Kluwer.

Gil, J. 1988. *Los sonidos del lenguaje.* Madrid: Síntesis.

Grijelmo. A. 1998. *Defensa apasionada del idioma español.* Grijelmo, Barcelona.

Harris, J. W. 1975 (1969). *Fonología generativa del español.* Barcelona: Planeta.

Harris, J. W. 1989. "Our present understanding of Spanish syllable structure". *American Spanish Pronunciation.* Eds. P. C. Bjarkman y R. M. Hammond. Washington: Georgetown Univ. Press. 151-169.

Hualde, I. 1991. "On Spanish syllabification". *Current Studies in Spanish Linguistics.* Eds. H. Campos y F. Martínez Gil. Washington: Georgetown Univ. Press. 475-493.

King, S., T. Portele y F. Hofer. 1997. "Speech Synthesis Using Non-Uniform Units in the Verbmobil Project". *Proceedings Eurospeech'97.*vol. 2. 569-572.

Ljolje, A., J. Hirchsberg y J. P. H. van Santen. 1997. "Automatic Speech Segmentation for Concatenative Inventory Selection". Eds. J. P. H. van Santen, R. W. Sproat, J. P. Olive, J. Hirschberg. *Progress in Speech Synthesis*, Nueva York: Springer-Verlag. 305-312.

Llisterri, J. 1988. "La síntesis del habla: estado de la cuestión". *Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural* 6: 17-41.

Llisterri, J., L. Aguilar, J. M. Garrido. 1997. "Incorporación de conocimientos fonéticos a las tecnologías del habla". *Panorama de la investigació lingüística a l'Estat Espanyol. Actes del I Congrés de Lingüística General.* Eds. E. Serra, B. Gallardo, M. Veyrat, D. Jacques y A. Alcina. Valencia: Univ. de Valencia. 5-13.

Machuca Ayuso, M. J. 1997. *Las obstruyentes no continuas del español: relaciones entre las categorías fonéticas y fonológicas en habla espontánea.* Tesis Doctoral. Departament de Filologia Espanyola, Universitat Autònoma de Barcelona.

Machuca Ayuso, M. J. 2000. "Articulación y pronunciación del español". *La expresión oral.* Ed. S. Alcoba. Barcelona: Ariel. 35-69.

Martínez Celdrán, E. y A. M. Fernández Planas. 2001. "Propuesta de transcripción para la africada palatal sonora del español". *Estudios de fonética experimental* XI: 175-190.

Mendieta, S. 1993. *Manual de estilo de TVE.* Barcelona: Labor.

Mohanan, K. P. 1986. *The Theory of Lexical Phonology.* Dordrecht: Reidel.

Navarro Tomás, T. 1985 (1918). *Manual de pronunciación española.* Madrid: CSIC.

Portele, T., F. Höfer y W. J. Hess. 1997. "A Mixed Inventory Structure for German Concatenative Synthesis". Eds. J. P. H. van Santen, R. W. Sproat, J. P. Olive, J. Hirschberg. *Progress in Speech Synthesis.* Nueva York: Springer-Verlag. 263-278

Radio Televisión Madrid 1993. *Libro de estilo de Telemadrid.* Madrid: Telemadrid.

Ren, H. 1986. *On the structure of diphthongal syllables.* PhD dissertation. Ann Arbor International Microfilms.

Ríos, A. 1998. *La transcripción fonética automática del Diccionario Electrónico de Formas Simples Flexivas del Español: un estudio fonológico en el léxico.* Barcelona: Universitat Autònoma de Barcelona [Disponible en http://elies.rediris.es/elies4].

Rodríguez, M. A., J. G. Escalada, A. Macarrón y L. Monzón. 1993. "AMIGO: Un conversor texto-voz para español". *Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural* 13: 389-400.

Tubau, I. 1993. *Periodismo oral.* Barcelona: Paidós.

Tzoukermann, E. 1994. "Text-to-speech for French". *Proceedings of the 2nd ESCA/ IEEE Workshop on Speech Synthesis.* New York. 179-182.

Whightman, C. W. y D. T. Talkin 1997. "The Aligner: Text-to-Speech Alignment Using Markov Models". Eds. J. P. H. van Santen, R. W. Sproat, J. P. Olive, J. Hirschberg. *Progress in Speech Synthesis.* Nueva York: Springer-Verlag. 313-324.

AGRADECIMIENTOS

Este trabajo no hubiera sido posible sin el esfuerzo de colaboración entre investigadores del Departament de Filologia Espanyola de la Universitat Autònoma de Barcelona, el Departamento de Teoria del Senyal i les Comunicacions de la Universitat Politècnica de Catalunya y de la Faculté Polytechnique de Mons durante los años 1996-1997.

Los resultados se presentaron y discutieron en el *XV Congreso Nacional de la Asociación Española de Lingüística Aplicada (AESLA)* celebrado en la Universidad de Zaragoza, del 15 al 18 de abril de 1997. Gracias a los asistentes por su interés.

Agradecemos, por último, las puntualizaciones de un evaluador anónimo.