UN SISTEMA AUTOMÁTICO DE SÍNTESIS DE HABLA MEDIANTE SEMISÍLABAS

Jorge Romano
Cátedra de Proceso de Datos
Universidad Técnica de Munich / R.F.A.

0. RESUMEN

El presente artículo describe un sistema automático de síntesis de habla para palabras castellanas aisladas. Las unidades acústico-fonéticas elegidas para la síntesis a partir de texto de un vocabulario ilimitado con buena calidad son semisílabas; la entrada al sistema es ortográfica. En primer lugar se determina el "inventario" o conjunto de semisílabas que aparecen en palabras castellanas aisladas. Este inventario comprende únicamente unas 650 semisílabas. Seguidamente se indica el método seguido para la obtención y codificación (parametrización) de estas unidades. Finalmente se describe un sencillo sistema de reglas para concatenar semisílabas formando sílabas y para concatenar estas últimas dando lugar a palabras. Estas reglas tienen en cuenta la coarticulación y la acentuación, con lo que se obtiene una voz sintetizada no sólo comprensible sino además de sonido natural.

1. INTRODUCCIÓN

Para sintetizar automáticamente la voz humana a partir de un texto mediante un ordenador existen diversas filosofías. Puede utilizarse un modelo físico-matemático de su producción (considerando las cuerdas vocales como un generador de ondas sonoras y las cavidades bucal y nasal como resonadores que modelan las ondas que las atraviesan; por ejemplo el sintetizador de formantes). También puede sintetizarse el habla tomando como base segmentos de señales sonoras de un hablante humano, segmentos que deben seleccionarse, codificarse y almacenarse previa y adecuadamente. Esta segunda posibilidad es la que permite obtener en general una voz sintetizada de mejor calidad, es decir de sonido más natural.

Para poder sintetizar todas las palabras posibles de una lengua, o sea un vocabulario ilimitado, resulta imposible almacenar en la memoria de un ordenador todas las señales sonoras de estas palabras producidas por una persona, pues, aunque se codificasen adecuadamente a fin de reducir su extensión, ocuparían un espacio de memoria prohibitivo. Por ello resulta necesario sintetizar las diferentes palabras combinando otras unidades menores.

Si elegimos los fonemas como unidades de síntesis obtenemos un inventario muy reducido (24 fonemas en el castellano peninsular). No obstante, aunque las palabras sintetizadas a partir de fonemas puedan resultar comprensibles, difícilmente obtendremos una voz natural. Para ello es necesario tener en cuenta la coarticulación, es decir, la influencia mutua entre fonemas vecinos debida a la limitada velocidad de la articulación. Como es bien sabido, la realización de un fonema determinado depende, debido a la coarticulación, del entorno en el que se presenta, dando lugar a sus diferentes alófonos. Para reproducir estos alófonos a partir de fonemas sería necesario un sistema de reglas muy complicado.

Gran parte de los efectos de la coarticulación se manifiestan dentro de la sílaba, por lo que la sílaba podría ser una unidad adecuada para un sistema de síntesis que tenga en cuenta la coarticulación. Sin embargo, para la síntesis de un vocabulario ilimitado, sería preciso un inventario que contuviera todas las sílabas posibles, en castellano del orden de 4000 probablemente. Este número sigue siendo excesivo para su almacenamiento en memoria, incluso de forma parametrizada. Además sería preciso determinar previamente todas las sílabas que pueden aparecer en la lengua castellana.

Si dividimos una sílaba en dos partes, cortándola por su núcleo vocálico, obtenemos dos semisílabas, una inicial y una final. El número de semisílabas existentes en una lengua es mucho menor que el de sílabas gracias al gran número de posibles combinaciones. Por otro lado, al ser el núcleo vocálico relativamente estable, hace que la coarticulación entre la semisílaba inicial y la final de una misma sílaba sea pequeña. Los restantes fenómenos de coarticulación no contenidos en las semisílabas pueden aproximarse (reproducirse) mediante un sistema de reglas relativamente sencillo. Por todo esto la semisílaba aparece como una unidad muy apropiada para la síntesis del vocabulario ilimitado de una lengua con una voz no sólo inteligible sino además relativamente natural.

2. ESTRUCTURA SILÁBICA DEL CASTELLANO

Para poder determinar el inventario de semisílabas de una lengua debemos estudiar previamente su estructura silábica. En nuestro caso esto es relativamente fácil, pues la estructura silábica del castellano es relativamente sencilla. La mayoría de las sílabas son cortas, de menos de 3 fonemas, y no existen sílabas con más de 5 fonemas. Además el número de fonemas del castellano es reducido: 5 vocales y 19 consonantes.

La sílaba se compone de ataque, núcleo y coda. En castellano tanto el ataque como la coda se componen de ninguna, una o como máximo dos consonantes; existen 32 posibles ataques silábicos pero únicamente 17 codas distintas (teniendo en cuenta las neutralizaciones; cf. Alcina & Blecua). Además la coda más frecuente es la vacía, consecuencia de la tendencia del castellano a la sílaba abierta. El núcleo silábico comprende una, dos o tres vocales. Existen en castellano 24 posibles núcleos vocálicos (ver tabla 1).

ataques silábicos	nulo:	/-/
	simples:	/p,t,k,b,d,g,s,f,θ,x,m,n,ñ,j,l,λ,r,r̄,c/
	compuestos:	/tr, pr, kr, dr, br, gr, fr, pl, kl, bl, gl, fl/
núcleos silábicos	vocales:	/a,e,i,o,u/
	diptongos: crecientes:	/ia, ie, io, iu, iu, ua, ue, ui, uo/
	diptongos: decrecientes:	/ai, ei, oi, au, eu, (ou)/
	triptongos:	/iai, iei, uai, uei, (iau)/
codas silábicas	vacía:	/-/ (muy frecuente)
	simples:	/B,D,G,N,s,l,r,θ,f,x/
	compuestas:	/Bs, Ds, Gs, Ns, ls, rs/

TABLA 1: Componentes de la sílaba castellana

La separación silábica del castellano puede describirse también mediante reglas sencillas, incluso en el caso de varias consonantes intervocálicas (cf. Alcina & Blecua, p.405). Únicamente la determinación de la frontera silábica entre vocales presenta ciertos problemas (cf. Esbozo). En nuestro sistema suponemos que si por lo menos una de las vocales es débil y no va acentuada forma diptongo, en caso contrario hiato.

La sencilla estructura silábica del castellano junto con las pocas reglas que rigen la separación silábica, la sencilla correspondencia entre grafemas y fonemas y las claras reglas de acentuación ortográfica facilitan el diseño de un sistema automático de síntesis de habla a partir de texto para la lengua castellana.

3. EL INVENTARIO DE SEMISÍLABAS

3.1 Definición de la semisílaba

Como ya se ha indicado, al dividir una sílaba efectuando un corte a través de su núcleo vocálico se obtienen dos semisílabas, una inicial (SSI) y una final (SSF). Las semísilabas ya han sido utilizadas eficazmente para la síntesis de habla (p. ej. Fujimura para el inglés y Dettweiler para el alemán) y también para el reconocimiento (Fujimura, Ruske & Schotola).

Tanto Fujimura como Dettweiler sugieren que el corte a través del núcleo vocálico sea asimétrico, de modo que aproximadamente el primer tercio del núcleo quede en la SSI y el resto en la SSF. Como la coarticulación se manifiesta sobre todo hacia atrás, es decir, desde un fonema a sus anteriores, con esta definición del corte se consigue que la mayor parte de la coarticulación de la coda silábica sobre el núcleo quede contenida en la SSF. Por otro lado, la menor duración de la influencia coarticulatoria del ataque silábico sobre la vocal siguiente queda dentro de la SSI. Esta definición del corte tiene en castellano una ventaja adicional: reduce el tamaño del inventario a almacenar pues acorta la duración de las SSI, que en castellano son mucho más numerosas que las SSF, al contrario por ejemplo que en alemán.

3.2 Inventario fonológico

El inventario mínimo para la síntesis de una lengua deberá contener todas las semisílabas que presenten oposición fonológica. Por tanto los pares de fonemas que se neutralizan en posición de coda silábica sólo se tienen en cuenta una vez.

Si el corte se realiza a través de la vocal fuerte, que es el núcleo silábico propiamente dicho, las transiciones de los diptongos crecientes (semiconsonantes) quedarán dentro de las SSI, y las de los decrecientes (semivocales), en las SSF. Por tanto el número de posibles SSI es igual al producto de posibles ataques por la suma de las cinco vocales más los 8 diptongos crecientes:

32 ataques x (5 vocales + 8 dipt.crec.) = 416 SSI −4 = 412 SSI

(Se restan 4 pues la realización de la semivocal /i/ tras un ataque vacío es igual que la de la consonante /j/.)

El número de SSF posibles debería ser el producto de la suma de vocales y diptongos decrecientes por el número de codas, pero como la coda compuesta implica un núcleo simple, este número se reduce a:

(5 vocales x 17 codas) + (6 dipt.decrec. x 11 codas simples) = 151 SSF

Los triptongos se realizan uniendo un diptongo creciente con uno decreciente. Obtenemos así un total de 563 semisílaba en el inventario. Este número podría reducirse si se demuestra que algunas combinaciones resultan imposibles (p.ej. /ji/+vocal = SSI). Para la síntesis de habla continua es posible que fuera necesario ampliar el inventario para dar cabida a nuevos "diptongos" (sinalefas).

3.3 Inventario fonético

Para que la voz sintetizada suene más natural es necesario introducir algunas variaciones alofónicas en el inventario. Sin embargo, cuantos más alófonos se distingan mayor será el número de semisílabas resultantes, y por consiguiente la capacidad de la memoria necesaria. Por tanto se trata de alcanzar un compromiso, incluyendo en el sistema únicamente los alófonos más importantes, es decir los más diferenciados y frecuentes, e intentando reproducir los restantes a partir de los fonemas respectivos mediante reglas que tengan en cuenta el entorno. Este es el caso, por ejemplo, en los alófonos vocálicos nasalizados, que se dejan reproducir con buena calidad interpolando entre la vocal y la nasal correspondiente.

Después de algunos experimentos se consideró necesario incluir en el inventario los alófonos fricativos de las plosivas sonoras (/β/, /ð/ y /ɣ/), que aparecen frecuentemente en el interior de palabra. Es necesario pues aumentar el número de ataques en 3 simples y 5 compuestos, con lo que se obtiene un total de 40. Por otro lado, aunque la fonología castellana sólo conoce un archifonema nasal /N/ en posición postnuclear, debido a sus grandes variaciones alofónicas se consideró necesario diferenciar por lo menos dos alófonos nasales en esta posición para que la articulación sonara mejor.

Tras estas consideraciones podemos estimar la extensión del inventario fonético de semisílabas:

(32 + 8) ataques x (5 + 8) núcleos - 4	= 516 SSI	678 semisílabas
((5+6) núcleos x (11+1) codas simples) + (5 vocales x 6 codas compuestas)	= 162 SSF	

3.4 Reducción del inventario

Dado que en castellano la coda compuesta consta de una consonante + /s/ y de que ésta presenta una coarticulación mínima con los fonemas precedentes, resulta posible la descomposición de todas las codas compuestas en la correspondiente coda simple más un "sufijo" /s/ sin pérdida perceptible de calidad. De este modo se logra reducir el número de SSF en 30, quedando un total de 132. El inventario fonético reducido comprende pues un total de 648 semisílaba más 1 sufijo.

Desgraciadamente, debido a la fuerte coarticulación de los fonemas en posición prenuclear, no es posible una descomposición análoga de los ataques compuestos en prefijo más ataque simple. Por tanto si quisiéramos reducir aún más el inventario sería preciso disminuir el número de núcleos vocálicos.

Hasta ahora hemos supuesto que a ambos lados del corte de una sílaba se encontraban segmentos vocálicos de la misma vocal. Una posible solución para reducir el inventario sería definir el corte de las semisílabas en la región de transición de los diptongos, quedando únicamente en el inventario núcleos vocálicos simples y sintetizando los diptongos mediante interpolación de los parámetros de dos vocales diferentes. Mediante experimentos se comprobó que la interpolación de los diptongos decrecientes en principio es posible. Pero para ello es necesario que el segmento vocálico de las SSI sea más largo que hasta ahora. Por tanto, aunque el número de SSF se reduciría en 60 habría que prolongar las más de 500 SSI, con lo que en resumidas cuentas se precisaría mayor espacio en memoria.

La reproducción de los diptongos crecientes parece, sin embargo, un problema difícil, debido a la rápida transición de las semiconsonantes. No obstante de este modo podría lograrse una drástica reducción del inventario, quedando únicamente 200 SSI (40 ataques x 5 vocales), por lo que parece que valdría la pena seguir estudiando este aspecto, así como la forma más adecuada de sintetizar los triptongos en este caso.

3.5 Obtención del inventario de semisílabas

El inventario de semisílabas se obtiene a partir de las señales sonoras digitalizadas y parametrizadas de un locutor, a fin de obtener una voz de sonido natural. La calidad de los registros debe ser lo más uniforme posible, con articulación clara pero no excesiva. Para ello las semisílabas no se obtienen de sílabas aisladas, sino de sílabas acentuadas de palabras bisilábicas sin sentido (logotomas) con un mínimo de coarticulación entre sus sílabas; se eligieron por ello las formas "SSI-pa" y "tap-SSF", pues la consonante labial /p/ presenta una articulación neutra ("posición fonética normal", cf. Quilis, p. 159).

Las logotomas conteniendo las semisílabas del inventario se grabaron en cinta magnética y se digitalizaron a continuación con un convertidor analógico-digital de 12 bits a 10 kHz. Mediante un programa interactivo en un terminal gráfico se extraen y copian finalmente los segmentos de señal correspondientes a las semisílabas deseadas, generando simultáneamente un fichero-índice, que contiene sus transcripciones fonéticas, tipos (SSI ó SSF) y direcciones de los cortes. Resulta preciso comprobar posteriormente la calidad de las semisílabas , p. ej. debido a la presencia de ruidos perturbadores, y corregir algunas direcciones, o sea la duración de algunos segmentos.

3.6 Parametrización y almacenamiento del inventario de semisílabas

Dado que el almacenamiento directo de las señales digitalizadas de todas las semisílabas ocuparía excesivo espacio de memoria, previamente a la síntesis se procede a la parametrización del inventario, lo que además resulta imprescindible para poder concatenar las semisílabas sin discontinuidades y para aplicar las reglas que tienen en cuenta la coarticulación y acentuación.

La parametrización se llevó a cabo mediante el método de predicción lineal (análisis LPC; cf. Markel & Gray). Cada 10 ms se obtiene un grupo de 12 coeficientes de reflexión, que se almacenan junto con la energía e información sobre la frecuencia del fundamental ("pitch"). Esta última información es de suma importancia para sintetizar una voz natural, pues al conservar la micromelodía, p.ej. en las transiciones entre fonemas, no sólo se consigue una mayor naturalidad sino que además se aumenta la inteligibilidad.

A continuación es necesario normalizar la curva de frecuencia de las semisílabas, entre otros motivos para obtener un valor constante en todos los cortes intravocálicos (cf. Dettweiler 1980). También puede resultar preciso corregir la curva de la frecuencia a posteriori para eliminar discontinuidades indeseadas.

4. REGLAS PARA LA SÍNTESIS DE PALABRAS AISLADAS

Las semisílabas parametrizadas no pueden utilizarse directamente para la síntesis; para evitar saltos bruscos en los cortes o uniones es preciso suavizar (alisar) estas discontinuidades. Esto se consigue mediante una interpolación adecuada de los parámetros. Además es preciso establecer reglas que modifiquen los parámetros teniendo en cuenta la coarticulación y la acentuación. (Nótese que los valores absolutos de la duración de la interpolación o del recorte se citan a modo indicativo, pudiendo depender fuertemente de la velocidad del habla con la que se genere el inventario.)

4.1 Síntesis de monosílabos

Para sintetizar una palabra monosílabica basta con concatenar una SSI con la correspondiente SSF (y en caso necesario además el "sufijo"; p. ej. "vals"). Según lo expuesto más arriba, ambas semisílabas deben presentar el en corte segmentos de la misma vocal. No obstante, si las uniéramos sin más, podrían producirse ruidos molestos en el corte, sobre todo si los dos grupos de parámetros a ambos lados del corte difieren mucho entre si. A fin de eliminar estas perturbaciones se modifican los grupos de parámetros en un intervalo alrededor del corte, interpolándolos entre los valores en los extremos de este intervalo. Los coeficientes de reflexión (PARCOR) modificados se obtienen mediante interpolación lineal de las correspondientes funciones del área ("log area ratios", cf. Markel & Gray). La energía y la frecuencia ("pitch") se interpolan linearmente (cf. Dettweiler 1980).

Este tipo de interpolación resulta también adecuado cuando es necesario prolongar o acortar las semisílabas en el corte, p. ej. en polisílabos, pues permite obtener una curva más natural de la melodía. En general basta con interpolar en un intervalo de 30 ms, simétrico respecto al corte, tal como indica Dettweiler para el alemán.

En sílabas con coda nasal la interpolación se realiza sólo del corte hacia atrás, o sea en el interior de la SSI, a fin de nasalizar también ligeramente el correspondiente segmento vocálico. Por otro lado ya hemos indicado que en el caso del sufijo /s/ no es preciso modificar los parámetros.

Aunque la duración del núcleo vocálico no es relevante en palabras monosílabicas, en las sílabas con diptongos crecientes se reduce ligeramente la duración de la vocal de la SSF, lo que permite obtener un ritmo silábico más uniforme en polisílabos. Además en todos los monosílabos se superpone a la micromelodía una rampa creciente de frecuencia, obteniéndose así monosílabos acentudados más naturales, pues la curva de pitch de las SSF fue normalizada al generar el inventario.

4.2 Síntesis de polisílabos: tratamiento de las fronteras silábicas

La síntesis de palabras polisílabicas presenta nuevos problemas; además de unir pares de semisílabas para formar sílabas es preciso concatenar éstas, con lo que se obtiene un nuevo tipo de uniones (SSF-SSI), donde es necesario tener en cuenta la coarticulación. Debemos diferenciar tres tipos de fronteras silábicas: vocal/consonante (v/c), vocal/vocal (v/v), y consonante/consonante (c/c); no existen en castellano apenas casos de frontera "consonante/vocal" debido a la tendencia a la sílaba abierta.

4.2.1 Frontera silábica v/c

La correcta síntesis de esta frontera es de gran importancia al ser la más frecuente en castellano. Como las SSF se obtuvieron de sílabas finales de las logotomas, presentan al final una clara relajación de la articulación, que, sin embargo, no debe aparecer en el interior de una palabra. Por ello, como regla general para la síntesis de esta frontera silábica, se cortan los 60 ms finales de la SSF y se prolonga ligeramente su segmento vocálico en el corte. Dependiendo de la consonante de la SSI se aplican además las siguientes reglas:

a) Si la SSI empieza con plosiva sorda (no existen plosivas sonoras en esta posición, sino alófonos fricativos!) o con africada, las dos semisílabas pueden unirse directamente sin necesidad de más modificaciones.

b) Si la SSI empieza con fricativa, lateral o nasal se cortan sus primeros 20 ms y se interpola 40 ms alrededor del corte (en el caso de nasal sólo hacia atrás), a fin de suavizar las transiciones de los formantes.

c) En el caso de las vibrantes la interpolación sólo dura 20 ms para no perder el primer golpe de la lengua. Además como la vibrante simple prolonga la duración de la vocal precedente, en este caso se aumenta la duración del segmento vocálico en el corte (después de haber cortado su porción final como en todos los casos). No obstante la reproducción de este fonema es difícil si la obtención de las correspondientes SSI no se realiza con sumo cuidado.

4.2.2 Frontera silábica v/v

Esta frontera implica la existencia de un hiato, o sea de una transición relativamente rápida de los formantes. Por ello la interpolación ha de ser breve (30 ms), a fin de evitar que el hiato se convierta en diptongo. Además se acortan ligeramente ambas semisílabas junto al corte para que las vocales no suenen excesivamente relajadas.

4.2.3 Frontera silábica c/c

En este frontera es también preciso distinguir varios casos, según el tipo de consonantes que lo formen.

a) Si la SSF acaba con plosiva o vibrante, las dos semisílabas pueden unirse directamente, sin interpolación ni acorte.

b) Si la SSF acaba con fricativa se cortan sus últimos 80 ms, pues las fricativas en interior de palabra son más breves que al final. Si la SSI empieza también con fricativa debe acortarse aún más la SSF.

c) Si la SSF presenta una coda compuesta, se sintetiza mediante la correspondiente SSF con coda simple, a la cual se le cortan los 80 ms finales, más el "sufijo" /s/, que debe oirse claramente. Si la SSI empieza también con fricativa, se cortan también 50 ms del sufijo.

d) Si la SSF acaba con nasal o lateral se cortan sus últimos 80 ms y se interpola 40 ms hacia atrás debido a la asimilación de la posición articulatoria del ataque. Si la SSI empieza además con plosiva sonora se acorta ésta en 50 ms. Sin embargo la SSF acabada en nasal es el caso que presenta mayores problemas, siendo quizás necesario introducir nuevos alófonos en el inventario para reproducir mejor la coraticulación.

4.3 Reglas para reproducir la acentuación

El castellano es una lengua con acento libre, que además presenta carácter distintivo. Aunque en el habla continua existan palabras tónicas y otras átonas, el presente artículo se limita al estudio de las primeras. Para la síntesis de palabras átonas en una habla continua debería tenerse en cuenta la existencia de "unidades melódicas" o "grupos acentuales" (cf. Alcina & Blecua, p. 253 y p. 455 ss).

La acentuación prosódica consiste primordialmente en destacar la sílaba tónica. Las palabras castellanas presentan en general una única sílaba tónica (a excepción de los adverbios terminados en -mente). Según Quilis (p. 330) y Solé (cap. 4, II) esto se consigue sobre todo aumentando, o en general variando la frecuencia de la voz, y en segundo lugar prolongando la duración de la vocal. La intensidad (energía) resulta irrelevante en la mayoría de los casos.

Por ello, para sintetizar una sílaba tónica superponemos a su micromelodía normalizada una rampa de frecuencia creciente. Si la sílaba tónica no es la última, a la siguiente se le superpone una rampa decreciente. Estas rampas presentan un ascenso del 20%, equivalente a unos 3 semitonos. Rampas más pronunciadas destacan aún más la presencia de un acento pero suenan menos naturales.

Además de variar la frecuencia se prolonga la duración de la sílaba tónica dilatando su SSF. También se prolonga ligeramente la SSI siguiente para destacar su ataque. Es preciso no obstante distinguir entre palabras agudas, llanas y esdrújulas para obtener una acentuación natural. En estas últimas, por ejemplo, la sílaba postónica es muy breve, por lo que al sintetizarla se la "comprime" adecuadamente.

Es además necesario tener en cuenta que la duración de la sílaba depende del número de sílabas de la palabra correspondiente; las palabras con varias sílabas presentan sílabas más cortas. Esto carece de importancia si se sintetizan palabras aisladas, pero deberá tenerse en cuenta al sintetizar habla continua.

Debe quedar claro, finalmente, que en el habla natural existen muchas posibilidades de realizar la acentuación. La aquí indicada es únicamente una de ellas, que no obstante permite reconocer correctamente el acento.

5. EL SISTEMA EXPERIMENTAL DE SÍNTESIS

El cometido del sistema experimental de síntesis es generar los parámetros LPC para la síntesis de una palabra a partir de la ortografía de dicha palabra. Mediante un sintetizador LPC se obtiene a continuación la senal acústica correspondiente. En la figura podemos ver el diagrama de flujo del sistema completo.

El primer paso del sistema es la trascripción fonética, o sea convertir la entrada ortográfica, tecleada p. ej. en un terminal, en la correspondiente escritura fonética, que ya tiene en cuenta los alófonos introducidos en el inventario (p. ej. los fricativos de las plosivas sonoras y los nasales). Simultáneamente se comprueba si existe un acento gráfico y cúal es su posición.

Seguidamente se transforma la secuencia de sonidos, es decir de fonemas y alófonos, en una secuencia de semisílabas. Para ello se determinan las fronteras silábicas y además en cada sílaba el corte semisilábico. En caso de que la palabra en cuestión no llevara acento gráfico se determina la sílaba tónica en función del último fonema.

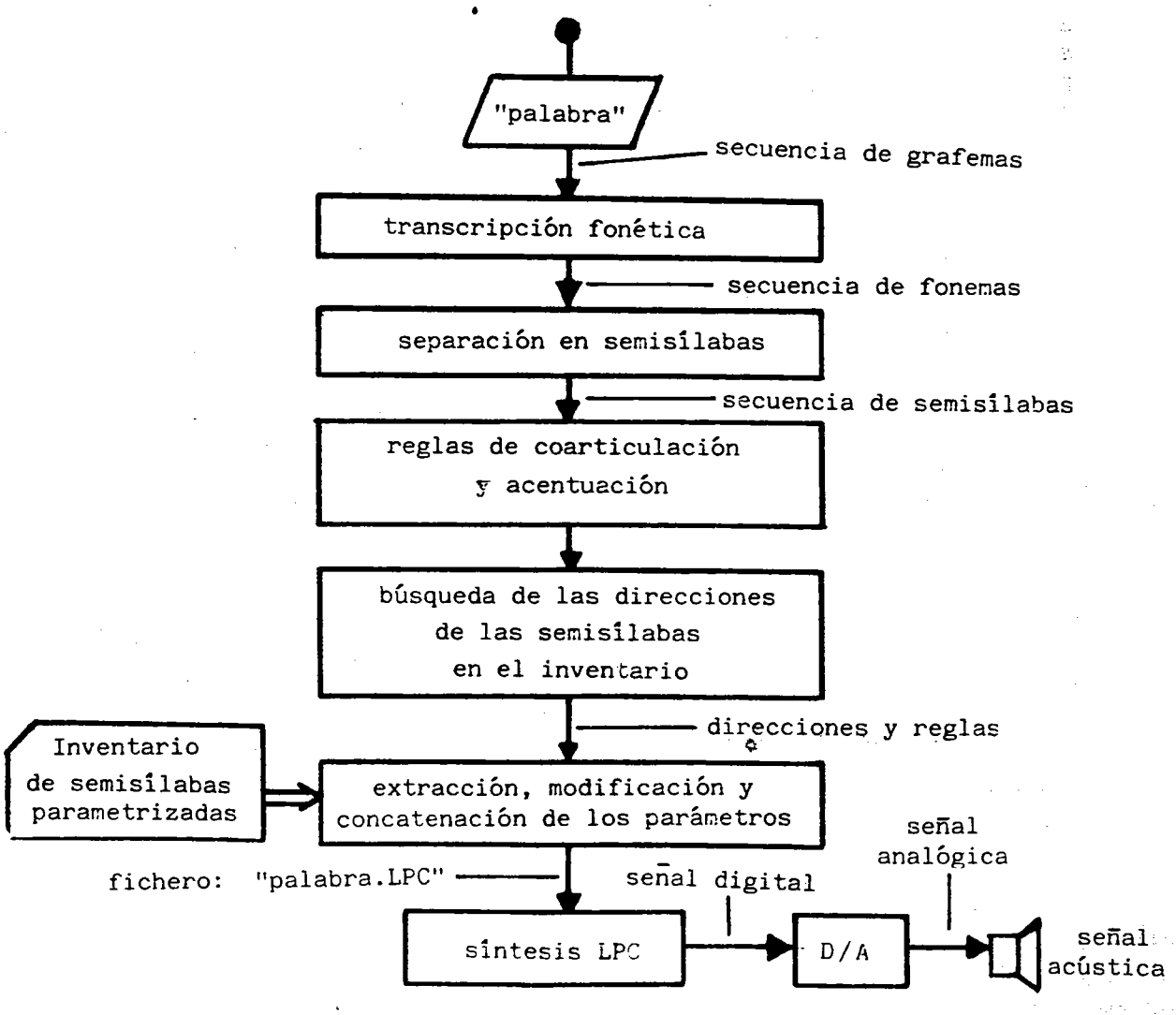

DIAGRAMA DE FLUJO DEL SISTEMA EXPERIMENTAL DE SÍNTESIS

El módulo siguiente determina las reglas de coarticulación y acentuación a aplicar. Acto seguido se determinan las direcciones en el inventario de los parámetros de las semisílabas necesarias, se extraen dichos parámetros y se concatenan adecuadamente, modificándolos conforme a las reglas previamente determinadas. De este modo se obtiene un fichero de tipo ".LPC", que sirve de entrada a un programa sintetizador LPC convencional (cf. Markel & Gray). Mediante un convertidor digital-analógico y el correspondiente altavoz puede oírse la voz sintetizada.

6. CONCLUSIONES

La impresión general de la voz sintetizada es buena; suena relativamente natural y es bastante comprensible. En un experimento piloto los oyentes entendieron correctamente el 75% de las palabras sintetizadas ofrecidas. La comprensión de las mismas palabras naturales, es decir, ofrecidas a los oyentes después de pasar por un análisis y síntesis LPC de las mismas características pero sin segmentar en semisílabas, fue del 81%, o sea solamente un 6% más. La voz del locutor con la que se generó el inventario puede reconocerse claramente.

No obstante, como la obtención de un buen inventario conlleva un gran número de problemas, no se utilizó el inventario completo sino solamente una parte de él, en el que sin embargo estaban contenidas todos los posibles ataques, núcleos y codas silábicos. De este modo es posible comprobar para todas las posibles combinaciones de fonemas la validez de las reglas de síntesis establecidas, reglas con las que se obtuvieron en general muy buenos resultados

La parametrización con el vocoder utilizado también presentó algunos problemas. Posiblemente si se aumentara la frecuencia de muestreo a 16 kHz y se incrementara el orden del vocoder, se podrían distinguir mejor algunos sonidos, por ejemplo la /s/ y la /f/, y comprobar con más rigor las reglas de síntesis establecidas. También sería entonces posible decidir sobre la conveniencia de introducir nuevos alófonos en el inventario.

Como el sistema experimental se implementó en un ordenador con sistema operativo a tiempo compartido, no se pueden dar datos sobre la velocidad de la síntesis, pues depende de la carga momentánea del ordenador. Además el sistema experimental no se ha optimizado en este sentido; simplemente se intentó que fuera lo más flexible posible, o sea fácil de modificar.

Finalmente podemos estimar la capacidad de memoria necesaria para el almacenamiento del inventario. Si suponemos que la duración media de las SSI es de 180 ms y que la de las SSF es de 250 ms, teniendo en cuenta que tenemos 516 SSI y 132 SSF, la duración de las señales acústicas a parametrizar es de unos 126 segundos. Si el análisis LPC se realiza con 6 kbit/s, equivalente a un vocoder convencional, el inventario ocuparía un total de 95 kByte de memoria, lo que permitiría implementarlo en un futuro próximo en un sistema de microordenador.

En definitiva, la síntesis de palabras castellanas aisladas utilizando semisílabas como unidades de síntesis se revela como un excelente método para sintetizar un vocabulario ilimitado con una voz de sonido natural además de comprensible. Con este método puede sintetizarse también habla continua, para lo cual sería necesario estudiar la acentuación de unidades melódicas y de frases completas, y la formación de sinalefas que obligaran a [illegible]uir nuevas semisílabas en el inventario.

7. BIBLIOGRAFIA

ALCINA-FRANCH, J. & BLECUA, J.M.: Gramática espanola. Ed. Ariel. Esplugues de Llobregat (Barcelona) 1975.

ESBOZO de una nueva gramática de la Lengua Espanola. Real Academia Espanola. Ed. Espasa-Calpe. Madrid 1977.

DETTWEILER, H.: Versuche zur Sprachsynthese deutscher Wörter mit Halbsilben. Fortschritte der Akustik, DAGA 80. VDE-Verlag. Berlin 1980, p.703-706

DETTWEILER, H.: An approach to demisyllable speech synthesis of german words. Proc. IEEE-ICASSP 81, Atlanta (Georgia), p.110-113.

DETTWEILER, H.: Sprachsynthese deutscher WÖrter mit Halbsilben. Fortschritte der Akustik,DAGA 82. VDE-Verlag, Berlin 1982, p. 1039-1042.

FUJIMURA, O.: Syllable as a unit of speech recognition. IEEE-Trans.ASSP 1975, Vol.23, p. 82-87.

FUJIMURA, O. et al.: Demisyllables and affixes for speech synthesis. 9th International Congress on Acoustics. Madrid 1977. Vol.1, p.513.

MARKEL, J.D. & GRAY, A.H.: Linear prediction of speech. Springer Verlag. Berlin etc. 1976.

QUILIS, A.: Fonètica acústica de la lengua espanola. Ed. Gredos. Madrid 1981.

RUSKE, G. & SCHOTOLA, T.: The efficiency of demisyllable segmentation in the recognition of spoken words. Proc. IEEE-ICASSP 81. Atlanta (Georgia), p. 971-974.

SOLÉ-SABATER, M.J.:Fonètica experimental. Domini, objectius i mètodes. Tesis doctoral. Fac. de Filología. Universidad de Barcelona 1982.

Sociedad para el Procesamiento del Lenguaje Natural

Con fecha 21 Marzo 1984, la Sociedad Española para el procesamiento del Lenguaje Natural ha sido inscrita en los correspondientes Registros Públicos, conforme a la Ley de 24 de Diciembre de 1964.

Los socios que aparecen en el Acta de constitución son los siguientes:

Ramón Almela
Manuela Alvarez
Xavier Artola
Joan Bastardas i Parera
Maria Teresa Cabré
Nicolás Campos
Ramón Cerdá
Arantza Díaz de Ilarraza
Jordi Fortuny
Ernesto García Camarero
Jorge García Sanz
Mª Carmen González Paez
Dolores González
Coloma Lleal i Galceran
Mª Antonia Martí
Eugenio Martínez Celdrán
Monserrat Meya (Promotora)
Honorino Mielgo
Francisco Milla Lozano
José Pérez García
Richard Pérez
Marc Antoní Pérez
M. Josefa Postigo
Pere J. Quetelas i Nicolau
Joaquín Rafel
Jorge Romano
Horacio Rodríguez
Kepa Sarasola
M. Felisa Verdejo (Promotora)

Esta lista se completa con las siguientes solicitudes recibidas posteriormente:

Josse de Kock

Angela Ferioli

Carlos Martín Vide

F. Miguel Martínez Marín

Hernan Urrutia Cárdenas

Se propone la celebración de la próxima asamblea de la Sociedad en Madrid el 2 de Noviembre. La reunión tendrá dos objetivos, uno de carácter organizativo (elección de la Junta Directiva, establecimiento de cuotas, difusión de la Sociedad) y otro de carácter científico (actividades, boletín, etc.).