MARIÑO, J. B.- NADEU, C.- LLISTERRI, J. (1987) "Síntesis automática del habla", in *Inteligencia artificial: conceptos, técnicas y aplicaciones*. Barcelona: Marcombo (Serie Mundo Electrónico, 13). pp. 157-165. ISBN: 84-267-0639-8.

http://liceu.uab.es/~joaquim/publicacions/Marino_Nadeu_Llisterri_87_Sintesis_Automatica_Habla.pdf

síntesis automática del habla 14

José B. Mariño, C. Nadeu y J. Llisterri

14.1 INTRODUCCION

El hombre realiza sus intercambios de información con el mundo exterior fundamentalmente a través del lenguaje, ya sea oral o escrito. Hasta el presente se puede afirmar que en su comunicación con los ordenadores (y los instrumentos controlados por ellos), el hombre ha hecho uso exclusivo del lenguaje escrito: pulsando las teclas de una consola para proporcionar o pedir información al ordenador o leyendo sobre la pantalla o el papel de la impresora el texto que el ordenador ofrece como respuesta. Resulta natural extender la capacidad de comunicación al mensaje oral.

No es difícil presentar un panorama de los pros y los contras que la comunicación oral puede ofrecer. En primer lugar debe mencionarse que, en la misma, las manos y la vista del usuario quedan liberadas, pudiendo aplicarse a una tarea simultánea a la comunicación. Ello ofrece posibilidades interesantes en el gobierno de sistemas de gran complejidad en que la atención visual sea importante (aeronaves, por ejemplo); permite sustituir la consulta de manuales por un diálogo con un ordenador instructor, manteniéndose la atención en el equipo con el que se está trabajando (adiestrándose en su uso, procediendo a su reparación, etc.); y, en general, facilita gran libertad de movimientos al personal usuario.

Una segunda ventaja proviene de la universalidad de la red telefónica; aunque ésta puede ser aprovechada para la transferencia de información sin acudir al habla, la comunicación oral, al no requerir otro equipo que el teléfono, ofrece una ventaja sustancial. Cualquier aparato telefónico se convierte en un enlace potencial con el ordenador; de este modo el acceso a bases de datos, la reserva y venta de billetes de avión o ferrocarril, las operaciones bancarias, etc., podrían realizarse desde cualquier punto.

En cuanto a los inconvenientes, pueden destacarse el carácter no privado del mensaje oral y los errores e incomodidades que acompañan a los sistemas de comunicación oral actuales, ya que distan (como veremos a lo largo de este capítulo) de ofrecer una conversación natural.

En la tabla 14.1 se ofrece, a modo de resumen, un cuadro ilustrativo de las ventajas e inconvenientes más sobresalientes que sobre la comunicación oral con el ordenador han enunciado diversos autores.

La comunicación es, fundamentalmente, una operación bidireccional: cada interlocutor ha de interpretar el mensaje que recibe y, a su vez, debe ser capaz de generar un mensaje. Aunque pueden citarse ejemplos de comunicaciones unidireccionales, las aplicaciones más interesantes de la comunicación hombre-máquina son aquellas que implican un diálogo. Debemos, por tanto, facultar al ordenador para hablar y para entender lo que se le dice. La primera facultad requiere que el ordenador *sintetice* el habla, lo que exige que disponga en su memoria de una representación acústica (*codificación*) de la voz; como se verá, la síntesis del habla es un problema próximo a obtener una solución plenamente satisfactoria. Por contra, la capacidad de entendimiento constituye hoy en día un horizonte lejano, prácticamente imposible de situar con cierta exactitud en el tiempo; en la comunicación oral, además del soporte físico del mensaje (la voz) están implicadas unas referencias lingüísticas comunes a los interlocutores; entender el habla implica no sólo *reconocer* las distintas palabras del mensaje oral e interpretar los contenidos sintácticos y semánticos, sino también conocer sus condicionamientos pragmáticos. En lo que decimos hay a menudo referencias al contexto y a nuestro conocimiento del mundo y de las reglas que rigen los intercambios conversacionales. Hoy en día solamente se han obtenido resultados parciales en el reconocimiento de la voz y se requieren avances fundamentales en inteligencia artificial para acceder al entendimiento del habla [1].

14.2 ADQUISICION DE LA SEÑAL

El almacenamiento, análisis y síntesis de la voz por el computador requieren como etapa previa la *digitalización* de la misma [2, 3]. Como el computador solamente trata con números (y relaciones lógicas), para que la voz le sea

Ventajas

1) Es la forma natural de la comunicación humana.
2) Universal entre los hombres.
3) Libera las manos y los ojos del usuario.
4) Es factible en la oscuridad.
5) Permite gran movilidad del operador.
6) Es posible la comunicación simultánea con hombres y máquinas.
7) Puede ser más rápida que otros medios de comunicación.
8) Compatible con sistemas de comunicación existentes.

Desventajas

1) No queda (a menos que se haga explícitamente) registro de la comunicación.
2) Falta de privacidad, por lo que puede ser escuchada o grabada por terceros.
3) Puede interferir otras comunicaciones orales.
4) Puede ser interferida por otras señales acústicas.
5) Fatiga, cambios psicológicos o físicos pueden alterar las características de la voz.

Tabla 14.1 Ventajas e inconvenientes de la comunicación oral hombre-máquina.

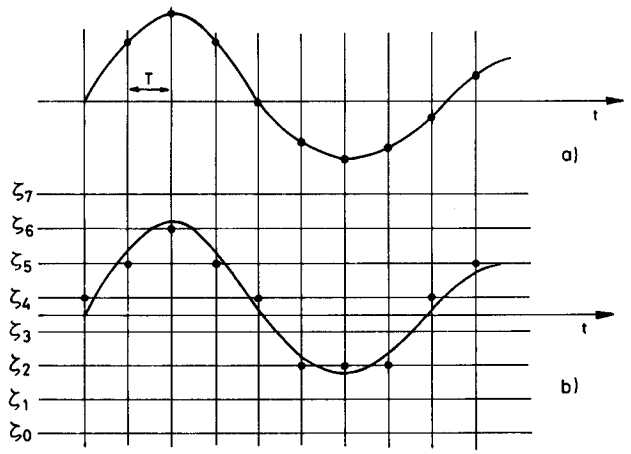

Figura 14.1 Ilustración del muestreo (*a*) y del muestreo y la cuantificación con 3 bits (*b*).

accesible es preciso convertirla en una sucesión de números. Este proceso implica básicamente dos operaciones: el *muestreo* y la *cuantificación*, tal como se ilustra en la figura 14.1. De la señal de voz se retienen únicamente muestras de la misma tomadas a intervalos regulares de tiempo T (período de muestreo); y, aunque el número de posibles valores que puede tomar cada muestra es infinito, dada la limitada capacidad de representación del computador, cada muestra es aproximada por el valor más próximo de entre los que tienen representación (cuantificación).

Para que este proceso de conversión analógico-digital (A/D) sea realizado adecuadamente, deben elegirse el período de muestreo T y el cuantificador de acuerdo con ciertos criterios.

Consideremos en primer lugar el período de muestreo. Resulta evidente que cuanto mayor sea más pobre resulta la representación de la señal, y parece razonable pensar que, si se sobrepasa cierto valor, la pérdida de información puede ser irrecuperable. Para ilustrar esto, consideremos el muestreo de la señal cosenoidal de frecuencia f

$$x(t) = A \cos(2\pi ft + \Phi)$$

con período de muestreo T, que proporciona la secuencia

$$x(n) = A \cos(2\pi fTn + \Phi)$$

Si tomamos en consideración la relación trigonométrica

$$\cos\beta = \cos(2\pi K \pm \beta) \qquad K = 1,2,...$$

se advierte fácilmente que la secuencia $x(n)$ puede resultar del muestreo de una señal cosenoidal con cualquiera de las frecuencias

$$f' = K\frac{1}{T} \pm f \qquad K = 1,2,...$$

De este modo, para que $x(n)$ represente a $x(t)$ sin ambigüedad, debe verificarse que

$$f < \frac{1}{2}\,\frac{1}{T}$$

Invocando al análisis armónico, podemos representar cualquier señal por un conjunto de sinusoides; de este modo, si Bf es la frecuencia más alta (ancho de banda de la señal), debe elegirse el período de muestreo de modo que

$$1/T > 2Bf = f_N$$

f_N recibe el nombre de frecuencia de Nyquist.

La cuantificación se ilustra en la figura 14.2 como una función no lineal del valor $x(n)$ de la muestra:

$$x_q(n) = \zeta_i \quad \text{si} \quad \eta_i < x(n) < \eta_{i+1}$$

donde los posibles valores de i son

$$i = 0, ..., N = 2^b - 1$$

y b es el número de bits utilizados para representar los valores cuantificados ζ_i. En el caso de que se conozca perfectamente la estadística de $x(n)$, se pueden diseñar los conjuntos de ζ_i y η_i de modo que la distorsión producida por la cuantificación (medida, por ejemplo, mediante el error cuadrático medio) sea mínima. En la práctica, el cuantificador más utilizado es aquel en que ζ_i y η_i se distribuyen uniformemente:

$$\begin{aligned} \zeta_i &= i\Delta + \zeta_0 \qquad i = 0, ..., N-1 \\ \eta_i &= i\Delta + \eta_0 \qquad i = 0, ..., N \end{aligned}$$

Δ recibe el nombre de escalón cuántico y $M_d = \eta_N - \eta_0$ el de margen dinámico; ambos están relacionados con el número de bits b del siguiente modo

$$\Delta = M_d / 2^b$$

Bajo ciertas hipótesis sencillas, la potencia del ruido de cuantificación

$$e_q(n) = x(n) - x_q(n)$$

es

$$\sigma_q^2 = \Delta^2 / 12$$

Si σ_x^2 es la potencia de la señal, la relación señal/ruido del cuantificador resulta ser

$$S/R = 10 \log \frac{\sigma_x^2}{\sigma_q^2} = 6b + 10{,}8 + 20 \log \sigma_x / M_d$$

De esta expresión puede concluirse que cuanto mayor sea

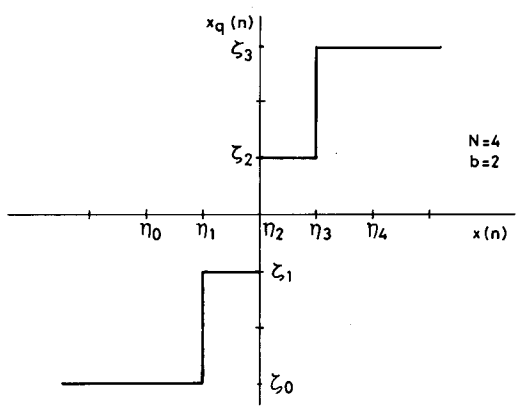

Figura 14.2 Ejemplo de cuantificador de 2 bits.

σ_x mayor será S/R; esto es así ya que el escalón cuántico está fijado y con ello la potencia del ruido de cuantificación; de este modo, cuando la potencia de la señal σ_x es pequeña el ruido es más apreciable que cuando dicha potencia es grande; este comportamiento tiene un límite, ya que al aumentar σ_x aumenta la probabilidad de valores de la señal fuera del margen del cuantificador (sobrecarga); típicamente se suele elegir para la voz

$$M_d = 4\,\sigma_x$$

de este modo, con 12 bits se obtiene una S/R máxima de 70 dB.

Como es bien sabido, la potencia de la señal de voz fluctúa con el tiempo en márgenes que pueden alcanzar 60 dB; ello disminuye el rendimiento del cuantificador. Un modo de equilibrar el comportamiento del mismo para pequeñas y grandes potencias es rediseñar los escalones cuánticos, haciéndolos más pequeños para niveles bajos de $x(n)$ y mayores para los niveles altos. Esto permite obtener un rendimiento equivalente al cuantificador uniforme de 12 bits con un cuantificador logarítmico (log PCM) de 8 bits.

Otro modo de mantener la eficiencia del cuantificador cuando fluctúa la potencia de la señal es mantener la relación σ_x/M_d constante; es decir, adaptar el margen dinámico M_d del cuantificador a la potencia de la señal. La figura 14.3 ilustra una realización práctica de esta idea.

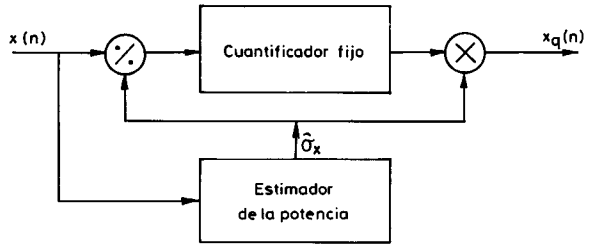

Figura 14.3 Cuantificador adaptativo: se estima la potencia de la señal, con lo que se normaliza ésta; se cuantifica con un cuantificador fijo diseñado para $\sigma_x = 1$ y se desnormaliza.

14.3 CODIFICACION DE LA VOZ

Si consideramos que, para preservar la calidad telefónica en la conversión A/D, se requieren 12 bits en un cuantificador uniforme y una frecuencia de muestreo normalizado en telefonía de 8 kHz, se deduce la necesidad de 96 kb/s para digitalizar la voz; esta cantidad resulta excesiva a efectos prácticos.

Por otro lado, la lengua es un procedimiento codificado para el intercambio de información, y los estudios sobre percepción y producción del habla sugieren la existencia de unidades abstractas subyacentes a la casi infinita gama de sonidos que el aparato fonatorio humano es capaz de producir. Dentro de esta gama hay elementos que poseen un valor distintivo, es decir, que se utilizan para distinguir palabras con significado diferente como /*s*/ y /*b*/ en *cabo* y *caso*. Tales unidades reciben el nombre de fonemas —por convención se transcriben entre barras inclinadas / /— y se pueden caracterizar mediante una matriz de rasgos articulatorios o acústicos [4]. Si contemplamos el habla como una sucesión en el tiempo de fonemas conectados o separados por pausas, 72 bits/s bastarían para su codificación, ya que la cadencia normal de pronunciación no excede en promedio los 12 fonemas/s y con 6 bits se pueden representar todos los fonemas de la mayoría de las lenguas —según E. Alarcos, el español posee 23 y casi nunca hay más de 40 en ninguna lengua—. Por supuesto, esta representación del habla no tendría en cuenta las características propias del locutor (los rasgos específicos de su voz, el influjo de las emociones, etc.).

Ahora bien, la conversión sin restricciones de la señal analógica en su correspondiente representación fonológica (reconocimiento) o viceversa (síntesis), no resulta factible al menos con los conocimientos actuales. La dificultad estriba en que en el habla natural los elementos abstractos que hemos denominado fonemas son realizados mediante la acción coordinada del conjunto del aparato fonatorio, por lo que sus propiedades acústicas —resultado de la configuración de dicho aparato en el momento de la producción— pueden variar en función del entorno. En el habla no hay fonemas, sino sus realizaciones sonoras, llamadas *alófonos* o segmentos y éstos —transcritos convencionalmente entre corchetes [ ]— están sometidos a un alto grado de variación.

En conjunto, la variabilidad de los alófonos puede atribuirse a tres factores: las diferencias entre los hablantes, las diferencias en un mismo locutor, y el contexto. Las realizaciones sonoras de cada individuo están condicionadas, en primer lugar, por la anatomía de su aparato fonatorio, que difiere de una persona a otra, pero también por características de tipo sociolingüístico como la variedad dialectal —piénsese, por ejemplo, en las diferencias de pronunciación entre el andaluz y el madrileño o entre las distintas variedades del español de América—, la clase social —la nasalización de las vocales asociada a un cierto «status»—, o las variaciones de tipo estilístico como la cadencia de pronunciación o la afectividad de los enunciados que puede modificar la duración de determinados alófonos (por ejemplo en una exclamación).

Además de las variaciones motivadas por las diferencias entre los hablantes, debe tenerse en cuenta que el mismo locutor repitiendo el mismo enunciado en un contexto idéntico producirá resultados acústicos diferentes debidos a los condicionamientos mecánicos del aparato fonatorio y a factores de tipo paralingüístico como el estado psicológico, la intención o la afectividad contenida en el mensaje.

Finalmente, las realizaciones de un fonema dependen también del entorno en que se encuentra. Por ejemplo, las características acústicas de las vocales responden a las de las consonantes adyacentes y viceversa, en un fenómeno de influencia mutua que se conoce como coarticulación, y la presencia o ausencia de una pausa condiciona la aparición en la cadena hablada de determinados alófonos —en español las características de la realización de /*b*/ en *bota* [bota] y en *cabo* [*kaβo*] son distintas: en el primer caso se trata de un cierre del paso del aire y en el segundo de una aproximación de los articuladores—. Además, los rasgos intrínsecos de los alófonos pueden verse modificados por los elementos llamados suprasegmentales —acento y entonación— que afectan a más de un segmento y que se perciben en el habla como cambios de intensidad, duración y altura tonal en correlación con los parámetros físicos correspondientes [5]. Así, por decirlo de otra manera, la parte de señal correspondiente a un fonema depende en gran manera del contexto fonológico y sintáctico en el que éste se encuentra en el momento de su realización en el habla [6, 7].

La gran distancia que media entre los 96 kb/s producidos por la digitalización y los 72 bits/s requeridos por la

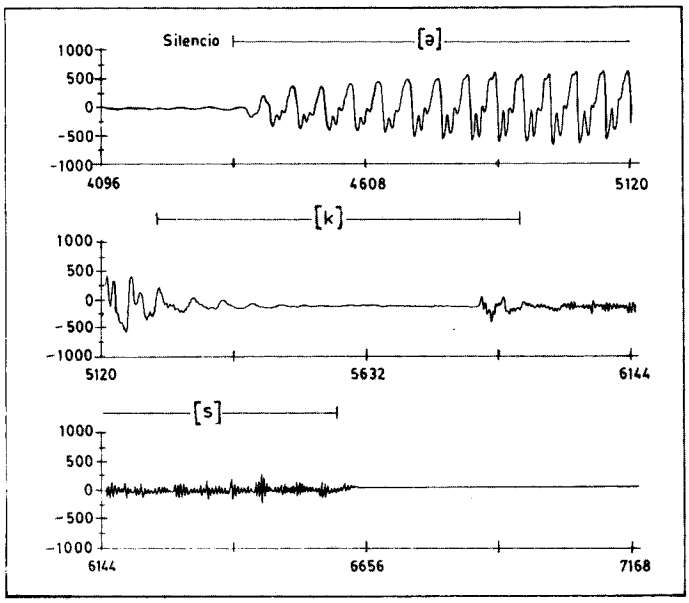

Figura 14.4 Señal de voz correspondiente a la primera sílaba de la palabra Expotrónica pronunciada en catalán. Sobre la figura se indica la extensión aproximada de cada alófono. La numeración en el eje temporal indica el número de muestras; la frecuencia de muestreo es de 8 kHz.

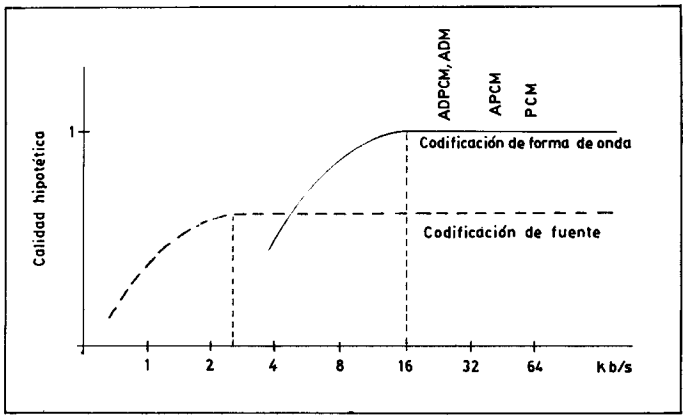

Figura 14.5 Relación calidad-compresión de los dos tipos de codificación.

codificación —actualmente utópica— de tipo fonológico hace claramente patente la gran redundancia presente en la señal de voz. La misma evolución temporal de la señal muestra ya dicha redundancia. Obsérvese, por ejemplo, en la figura 14.4, cómo la parte de señal correspondiente a la vocal es casi periódica y una gran parte de la consonante oclusiva [*k*] no es más que silencio.

Todos los sistemas de codificación explotan de algún modo la redundancia con el objetivo de reducir los requerimientos de memoria o velocidad de transmisión de la simple codificación PCM lineal, atendiendo al mismo tiempo a las demandas contrapuestas de sencillez del sistema y de mantenimiento de la calidad del habla (inteligibilidad, naturalidad, etc.).

Los sistemas de codificación se suelen agrupar en dos clases [8]. En la primera, la finalidad primaria es mantener la forma de la evolución de la señal en el tiempo (*codificación de forma de onda*). Son codificaciones que consiguen una alta calidad del habla, pero sólo superan la barrera de los 16 kb/s en ciertos sistemas experimentales bastante complejos. Los sistemas de la segunda clase consiguen eliminar redundancia basándose en un modelo de producción del habla y codificando la evolución temporal de sus parámetros significativos (*codificación de fuente*). En la figura 14.5 se muestra la diferencia de comportamiento de ambas clases de sistemas por lo que se refiere a la relación que establecen entre compresión de información y calidad del habla.

14.3.1 Codificación de forma de onda

Las características estadísticas de la señal de voz evolucionan continuamente en el tiempo. Sin embargo, debido al lento movimiento de los órganos del aparato fonador humano, pueden considerarse estacionarias localmente, es decir, durante cortos intervalos de tiempo (10-30 ms). Dicha propiedad es explotada en lo que se refiere a la amplitud por los cuantificadores adaptativos ya mencionados.

Otra característica que presentan la mayoría de sonidos del habla es una concentración de energía a bajas frecuencias del espectro, lo cual implica un alto grado de correlación temporal entre muestras de la señal. Los *codificadores diferenciales* (DPCM), en su forma más simple, aprovechan esta propiedad cuantificando la diferencia de amplitudes entre muestras consecutivas, lo cual permite juntar más los niveles de cuantificacion y, por tanto, disminuir el error para un determinado número de bits. Otra técnica muy simple es la denominada *modulación delta* (DM), la cual hace aún más pequeña la diferencia anterior a base de muestrear la señal a una frecuencia varias veces superior a la frecuencia de Nyquist, posibilitando así la utilización de un cuantificador de dos niveles (1 bit) que suele ser *adaptativo* (ADM).

El DPCM puede interpretarse como un sistema que cuantifica el error entre la muestra a codificar y una predicción de la misma (la muestra anterior). Esta idea puede extenderse prediciendo en base a una combinación lineal de cierto número de muestras (*predicción lineal*). Los coeficientes del predictor están estrechamente vinculados a la correlación de la señal y, por tanto, a su espectro. Este tipo de codificación utiliza siempre un cuantificador adaptativo y en su forma más completa adapta el predictor a las variaciones del espectro (ADPCM). En algunas —pero no todas— de las modalidades de ADPCM y de otros sistemas adaptativos de codificación se hace necesario que, adicionalmente a la secuencia de bits generados por el cuantificador, se añada el escalón cuántico y los coeficientes del predictor codificados (*información lateral*).

En la figura 14.6 se muestra el esquema básico ADPCM que engloba también las demás codificaciones descritas. En él se ha añadido un bloque conformador espectral del ruido o error de cuantificación, el cual trata de conseguir que la relación señal/ruido sea aproximadamente constante en función de la frecuencia a fin de mejorar la percepción del habla.

Existen otras técnicas más complejas que no responden al esquema de la figura 14.6, tales como las que pasan la señal a un dominio transformado y allí la cuantifican (codificación en subandas, ATC, etc.), o las que trabajan con vectores de muestras en lugar de muestras sueltas (*cuantificación vectorial*) [3]. Ninguna de dichas técnicas ha salido aún del laboratorio.

14.3.2 Codificación de fuente o paramétrica

Tal como se ha dicho, este tipo de codificación presupone

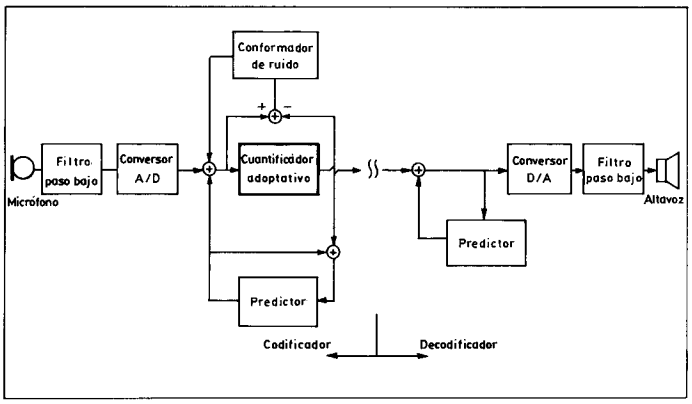

Figura 14.6 Esquema de un sistema completo de codificación y decodificación ADPCM. El filtro de entrada limita el espectro de la señal a la mitad de la frecuencia de muestreo a fin de evitar el solapamiento («aliasing») en la digitalización.

un modelo paramétrico de generación de la señal. En el caso del habla, el modelo básico que se emplea es el de la figura 14.7; en él, la señal de voz es la respuesta de un filtro lineal variante con el tiempo a una excitación consistente en una secuencia cuasi-periódica de pulsos (presencia de *sonoridad*), un ruido de banda ancha (ausencia de sonoridad) o una combinación de ambas [2]. Los pulsos (*el tono*) son atribuidos a la vibración de las cuerdas vocales y el ruido de banda ancha al paso del aire a través de una constricción. El filtro modela la acción del tubo acústico (*tracto vocal*) situado entre la glotis y los labios. Los polos del filtro originan picos en el espectro que simulan las resonancias del tubo acústico (los *formantes*). Por ejemplo, la vocal [ə] de la figura 14.4 es un alófono sonoro, por lo cual le corresponderá un espectro tal como el de la figura 14.8*a*, es decir, con una estructura de formantes bien definida y un detalle fino del espectro consistente en armónicos que son debidos a la cuasi-estacionariedad del tono. En cambio, la consonante fricativa [s] es sorda y tendrá asociado un espectro tal como el de la figura 14.8*b* en el cual el detalle fino tiene un cariz errático o ruidoso [6, 7].

Los codificadores paramétricos (*vocoders*) tienen en cuenta aspectos relevantes de la señal desde el punto de vista perceptual. Dado que el oído es relativamente insensible a la distorsión de fase, estos codificadores ponen el énfasis en el modelado eficiente de las partes del espectro con amplitud alta (por ejemplo, los formantes) a las cuales es más sensible el oído.

Una de las dos técnicas más utilizadas parametriza la envolvente del espectro por medio de los coeficientes de predicción lineal (*modelo LPC*) [9], lo cual equivale a modelar el tracto vocal con un filtro cuya función de

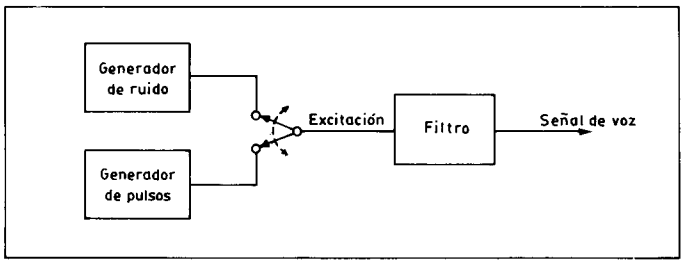

Figura 14.7 Modelo digital de producción de la voz.

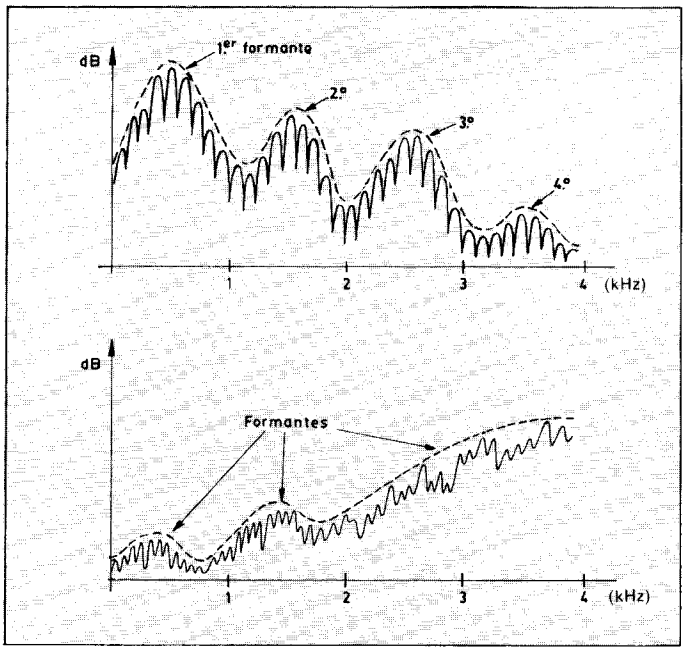

Figura 14.8 Espectros típicos de voz sonora (*a*) y sorda (*b*).

transferencia, que es la inversa de la del filtro de la figura 14.9, no presenta ceros, sino sólo polos. Los coeficientes se estiman minimizando la energía del error de predicción; el procedimiento que permite calcularlos a partir de la señal es directo y eficiente.

En la codificación LPC, contrariamente a lo que sucede en ADPCM, no se conserva el residuo o error de predicción sino —únicamente— el valor de la energía de la señal, la indicación del tipo de excitación, y, cuando hay sonoridad, el valor del período del tono (se considera que la conformación de los pulsos es llevada a cabo también por el filtro).

Los sintetizadores LPC utilizan habitualmente para la realización del filtro la estructura denominada red en celosía, cuyos parámetros (PARCOR o coeficientes de reflexión) tienen sentido físico en términos de reflexión acústica en el interior del tracto vocal y presentan indudables ventajas por lo que a su cuantificación se refiere. Además, pueden ser interpolados linealmente en las transiciones entre segmentos (de 10-30 ms) de la señal sin que de ello resulten filtros inestables [2].

Se suelen emplear filtros con 10-12 coeficientes que son actualizados cada 10-30 ms y cuantificados normalmente con un promedio de 4 bits por coeficiente. Aparte, con la

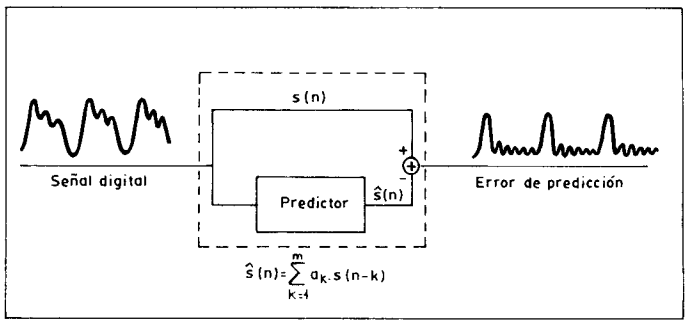

Figura 14.9 Filtro de predicción lineal; $s(i)$ es la muestra de la señal en el instante i.

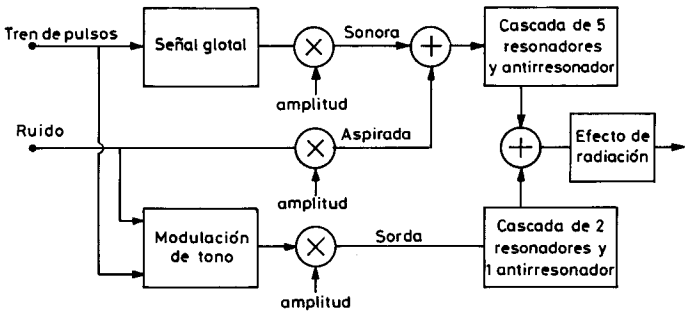

Figura 14.10 Esquema general de un sintetizador de formantes.

misma periodicidad, se codifican también los parámetros asociados a la excitación, es decir, el período del tono y el factor de amplitud, cada uno de ellos con 5-6 bits. En total resulta un mínimo del orden de 1200 bits / s, que aún puede ser reducido hasta 800 bits/s si se cuantifican vectorialmente los parámetros.

El máximo resulta ser, con los valores anteriores, de 6000 bits/s, velocidad que es bastante inferior a la de los codificadores de forma de onda más eficientes. Sin embargo, en comparación con ellos, se obtiene un habla menos natural, de *calidad sintética*, lo cual es atribuible principalmente a la excesiva simplicidad del modelo de excitación. En 1978, el gobierno norteamericano definió un estándar (LPC-10) con un filtro de predicción de orden 10, y un requerimiento de 2,4 kb/s [10].

La segunda técnica de codificación paramétrica utiliza directamente los valores de frecuencia y ancho de banda de los formantes. Los sintetizadores más sencillos suelen constar de cuatro filtros (*resonadores*) digitales de segundo orden, colocados normalmente en cascada, sintonizado cada uno de ellos a un formante [11]. Los sintetizadores más complejos disponen de un extenso conjunto de filtros en serie y paralelo que, junto con su forma de combinar los dos tipos de excitación, les confiere gran flexibilidad en el control de los factores que inciden en la producción del habla. Ello permite, por ejemplo, modelar fácilmente los ceros espectrales de los sonidos nasales. El principal inconveniente de este tipo de codificación es la dificultad del proceso de análisis; la obtención de los formantes y su ancho de banda es una tarea llena de dificultades, y no bien resuelta todavía: los formantes presentan un amplio margen de variación en frecuencia y, a menudo, tienen lugar confusiones entre formantes cercanos. Otro inconveniente de esta técnica es su dificultad para modelar variaciones rápidas de los formantes.

En la figura 14.10 se muestra un posible esquema de un sintetizador por formantes semejante al de Klatt [12]. El tracto vocal correspondiente a voz sonora y aspirada es modelada mediante la conexión en cascada de 4 resonadores; en las realizaciones más complejas, tanto los formantes como sus anchos de banda evolucionan con el tiempo; versiones más sencillas no ejercen control sobre los anchos de banda, lo que provoca una degradación más apreciable en las consonantes nasales. En los sonidos sordos se localiza la excitación en la cavidad bucal; el menor tamaño del resonador acústico concentra la energía en las altas frecuencias; de este modo, uno o dos resonadores son suficientes para modelar estos sonidos. Un tratamiento adecuado de las nasales exige la introducción de un par adicional resonador/antirresonador para modelar un formante adicional provocado por la mayor longitud del tracto vocal y el cero espectral más bajo de los introducidos por el acoplamiento de la cavidad nasal.

14.4 SINTESIS DEL HABLA

Como es lógico suponer, la produccion del habla con un ordenador requiere el almacenamiento previo, en forma codificada, de la señal acústica correspondiente a unidades lingüísticas elementales (fonemas, sílabas, palabras, etc.) y se realiza concatenando segmentos de voz obtenidos a partir de ellas [11, 13].

Así pues, el tipo de unidad elemental utilizada es una característica definitoria de todo sistema de síntesis. Cuanto mayor es el tamaño de la unidad, mejor es la calidad del habla, debido a que el proceso de concatenación reviste menos dificultad. Sin embargo, la memoria precisada para el almacenamiento de todas las unidades crece exponencialmente con su tamaño. En efecto, mientras que el número de fonemas que posee una lengua no llega al medio centenar, su léxico comprende centenares de miles de palabras.

Este compromiso entre maximización de la calidad del habla sintetizada y minimización de la memoria requerida también aparece en la codificación, puesto que la aplicación de los métodos más eficientes en cuanto al ahorro de memoria (codificación de fuente) introduce una cierta degradación de la calidad del habla. No obstante, el espacio de memoria no es el único factor a minimizar en un sistema de síntesis, sino también la complejidad del sintetizador y, especialmente, la extensión del conjunto de reglas que gobiernan la concatenación y la dificultad de creación del vocabulario de unidades elementales. Por supuesto, la elección de una u otra opción dependerá de cada aplicación específica.

Los sistemas prácticos de síntesis automática del habla se pueden dividir, según el tipo de aplicación, en dos clases. Por un lado, los sistemas de *respuesta oral* requieren un análisis previo a la síntesis, trabajan a nivel de palabras y utilizan primordialmente técnicas de procesado de señal. Por otro lado, los sistemas de *conversión de texto a voz* utilizan

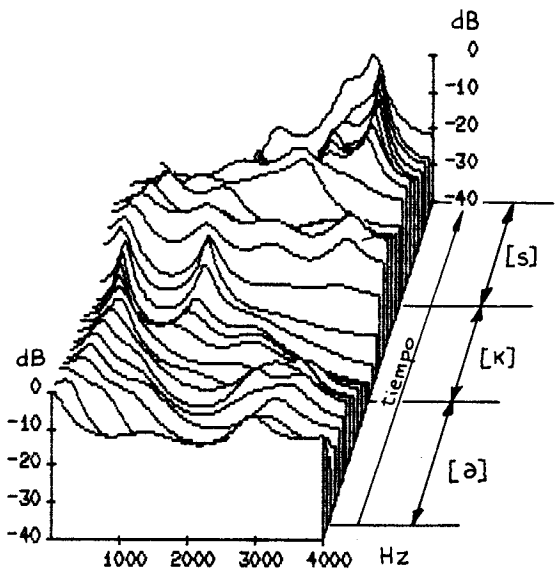

Figura 14.11 Secuencia de espectros correspondientes a la señal mostrada en la figura 14.4. Los espectros se presentan en una versión alisada (LPC), de forma que se han perdido las fluctuaciones debidas a la excitación. La separación entre ellos es de 15 ms.

unidades elementales de menor tamaño y requieren un extenso procesamiento lingüístico [14].

14.4.1 Dificultades de la concatenación

Los problemas de la concatenación son evidentes si se observa atentamente una representación acústica del habla y se consideran las dificultades que plantea su segmentación en elementos discretos: las características temporales y frecuenciales de la señal evolucionan a lo largo del tiempo en una sucesión de transiciones y estados estacionarios, como puede apreciarse en la figura 14.11. La concatenación de los elementos aislados sin respetar las influencias mutuas entre los sonidos del habla —la coarticulación—, lleva a un producto cuyos cambios bruscos degradan en gran manera su inteligibilidad.

Es preciso considerar también otros factores, pues los valores de los parámetros físicos directamente relacionados con el acento y la entonación (frecuencia fundamental, energía y duración de las partes estacionarias) varían según se trate de un alófono acentuado o no, de un enunciado afirmativo, interrogativo, imperativo, etc. Los efectos de algunas de estas modificaciones —como la entonación— se extienden a más de un alófono, tal como indica el término suprasegmental con que se las designa en lingüística [6, 7].

Así pues, las unidades elementales almacenadas en memoria que han de ser concatenadas se deben modificar convenientemente a fin de considerar tanto la coarticulación como los elementos suprasegmentales. Ahora bien, la codificación de forma de onda no es adecuada cuando se requiere dicha modificación, puesto que la coarticulación, debida al paso gradual de una posición a otra del tracto vocal, no puede ser simulada por simple interpolación de la forma temporal de la señal y, aunque la entonación puede variarse quitando o añadiendo muestras a cada período, las alteraciones producidas sobre el espectro originan una perceptible degradación de la calidad del habla. Por consiguiente, para poder utilizar la codificación de forma de onda se hace necesario el almacenamiento de todas las unidades modificadas susceptibles de ser empleadas en la concatenación.

Contrariamente, la codificación de fuente permite realizar las mencionadas modificaciones sobre cada unidad elemental guardada en memoria. La razón está en la flexibilidad de los sintetizadores paramétricos que, al efectuar la separación de excitación y filtro permiten modificar independientemente la frecuencia fundamental, la energía o la duración (parámetros asociados a la excitación), o realizar la interpolación o alisado de los parámetros del filtro que simule la coarticulación.

14.4.2 Sistemas de respuesta oral

La forma más familiar de producir respuesta oral automática con técnicas analógicas es la que utiliza grabadora de cinta. Mas, como sucede también con las demás formas de almacenamiento de la señal analógica, los sistemas resultantes son caros y están muy sujetos a fallos mecánicos, debido al obligado acceso aleatorio a las distintas partes de voz almacenadas. En cambio, la grabación en forma digital posibilita un acceso rápido a cualquier segmento de voz, a un relativamente bajo coste y sin deterioro apreciable en la calidad, siempre que se haga uso de la codificación de forma de onda.

Los sistemas más simples de producción de habla graban enteras las expresiones orales. Así, por ejemplo, los denominados sistemas de almacenamiento y reproducción (*store and forward*) de mensajes hablados, que tanta aplicación empiezan a encontrar en el mundo de la oficina. En ellos se utilizan codificaciones sencillas tales como ADM o ADPCM que permiten comprimir la señal a 32 kb/s manteniendo la calidad de voz telefónica. Otra forma de obtener un ahorro substancial de memoria consiste en evitar el almacenamiento de señal en las pausas o silencios presentes en el habla, codificando tan sólo su duración con el fin de poder regenerarlos al reproducir el mensaje. Un simple detector de voz basado en la energía de la señal puede ser suficiente para separar la voz del ruido ambiental presente en los silencios. El ahorro real de memoria depende de la longitud media del mensaje; para mensajes cortos, las pausas inicial y final ocupan una parte significativa del tiempo total y el ahorro puede ser perfectamente de un 50 % [15].

Cuando la memoria requerida para la grabación de los mensajes enteros resulta excesiva, se hace indispensable la concatenación de pequeñas partes almacenadas separadamente. Parece razonable, entonces, la partición en palabras (o frases muy cortas), ya que se ven mucho menos afectadas por la coarticulación que las unidades lingüísticas de menor tamaño. Esta es, en la actualidad, una forma muy utilizada de síntesis de habla, pues aunque el número de palabras fonéticamente distintas sea muy elevado, la mayoría de las aplicaciones emplean un vocabulario que no sobrepasa unos pocos centenares de palabras [16].

Generalmente, la reproducción se lleva a cabo sin producir ninguna alteración a las palabras que se concatenan. Aunque, como ya se ha mencionado, la codificación de fuente permite modificar con facilidad las características de la voz, los intentos realizados con palabras no han producido resultados satisfactorios. Así pues, las frases se forman yuxtaponiendo varias palabras grabadas por separado o en un contexto diferente, sin modificar su ritmo ni su entonación, lo cual hace que se perciban con falta de naturalidad y que disminuya su inteligibilidad. No obstante, ello no significa un inconveniente grave en muchas aplicaciones: las expresiones invariantes pueden almacenarse enteras; si una palabra ha de insertarse en una frase, se puede grabar con las características adecuadas y, si la misma palabra ha de utilizarse en contextos diferentes y la falta de fluidez no resulta tolerable, pueden grabarse dos o más versiones de la palabra, una para cada contexto.

Estos sistemas de respuesta oral que yuxtaponen palabras grabadas previamente, resultan adecuados en una gran variedad de aplicaciones (sistemas de información por teléfono, sistemas de alarma, juguetes parlantes, etc.) en los que el texto es limitado y el vocabulario reducido. A menudo recurren a la *codificación de fuente* —LPC, sobre todo— porque suele permitirse cierta tolerancia en la calidad del habla y los requerimientos de memoria lo hacen aconsejable. Este tipo de codificación demanda que el análisis automático de la voz para estimar los valores de los parámetros vaya seguido por la acción de un experto, a fin de corregir dichos valores en los intervalos de tiempo donde dicho análisis sea incorrecto o codificar más eficientemente contenidos concretos sin introducir degradación perceptual; en la figura 14.12 se ilustran los procesos de análisis y síntesis implicados en esta clase de sistemas.

14.4.3 Síntesis por reglas

En aplicaciones más generales en las que el texto es ilimitado (conversión de texto a voz) es necesario recurrir a

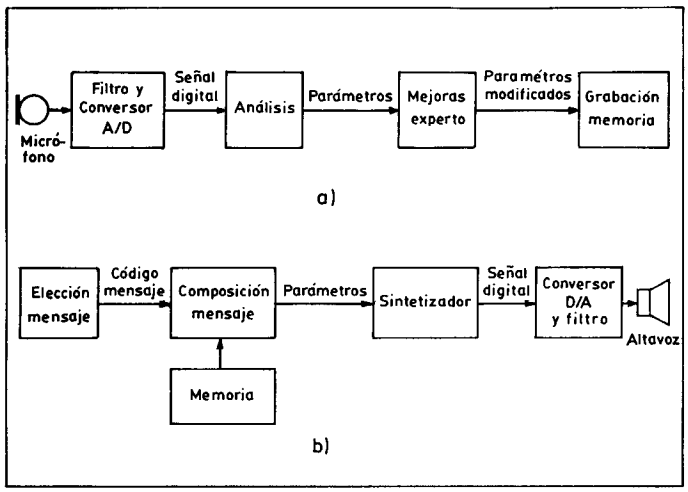

Figura 14.12 Proceso de análisis (*a*) y síntesis (*b*) en un sistema de respuesta oral con codificación de fuente.

unidades más pequeñas que las palabras, tales como fonemas o partes de sílabas. Ahora bien, la síntesis a partir de la concatenación de segmentos sonoros aislados produce resultados inaceptables por las razones que hasta aquí se han ido exponiendo; cuanto más se reduce el tamaño de la unidad utilizada, más se agudizan las dificultades de concatenación.

El objetivo de un sistema de conversión de texto a voz es la obtención automática de la realización sonora de cualquier mensaje escrito [17]. Para reproducir este proceso, que cualquier persona alfabetizada lleva a cabo con suma facilidad, parece lógico recurrir a los conocimientos adquiridos en el estudio de la actividad lingüística de los hablantes. Puede pensarse que las unidades de la síntesis serán las mismas que subyacen a los procesos de producción y percepción del habla —es decir, los fonemas— y caracterizarlas a partir de las propiedades acústicas que utiliza nuestro sistema perceptivo para identificarlas. Así es posible generarlas artificialmente utilizando rasgos de tipo subfonémico extraídos del análisis del habla natural. Desde esta perspectiva, adoptada en los primeros trabajos de síntesis por reglas llevados a cabo en los laboratorios Haskins a finales de los años cincuenta [18], las reglas que definen los rasgos de los fonemas y las modificaciones que en éstos produce el contexto en que se realizan son también útiles en sistemas de reconocimiento que efectúan la operación inversa a la síntesis.

El problema que presenta tal estrategia reside en el tratamiento del alto grado de variación contextual que se da en la realización de los fonemas. Una posible solución consiste en incluir reglas que especifiquen las características acústicas de cada fonema o de cada conjunto de fonemas en función de un contexto determinado y que definan adecuadamente las transiciones entre los segmentos, de vital importancia si queremos reproducir los efectos coarticulatorios del habla real. Es posible también almacenar varios alófonos para cada fonema, con lo cual deben introducirse reglas que indiquen cuál de ellos debe seleccionarse en cada caso.

En contrapartida, el número de fonemas de una lengua es, como hemos visto, reducido; incluso en lo que respecta a los alófonos, su número también es pequeño: para el español, en el sistema de transcripción de la *Revista de Filología Española* se contabilizan 53.

Otras formas de síntesis por regla simplifican el modelado de las transiciones utilizando unidades de mayor tamaño como el difonema y la semisílaba. El difonema se define como el segmento de habla comprendido entre los centros de dos fonemas contiguos, suponiendo para el tratamiento de las transiciones que en el centro del fonema se da una estabilidad que permite fácilmente la concatenación. Este método suele producir buenos resultados y ofrece un sistema de síntesis relativamente flexible, pero presenta problemas en el tratamiento de los elementos suprasegmentales, pues las variaciones de amplitud, duración y frecuencia fundamental o tono deben ajustarse. Las semisílabas, tal como su nombre indica, abarcan la primera o la segunda mitad de una sílaba. Aventajan a los difonemas en la naturalidad conseguida en el tratamiento de los grupos consonánticos. Respecto a las sílabas como unidades de síntesis, poseen la ventaja de ser mucho menos numerosas [19].

Las reglas empleadas en la síntesis —entendiendo por regla cualquier especificación sobre la naturaleza de un elemento segmental o suprasegmental— se basan en el conocimiento de la estructura de la lengua y suelen formalizarse como reglas de reescritura del tipo $A \rightarrow B / C\_\_D$, es decir, A se convierte en B a condición de que esté precedido de *C* y seguido de *D*; tales reglas pueden ser obligatorias u optativas y contienen además elementos optativos ( ) o alternativos { }. En los sistemas de síntesis por reglas hay una tendencia cada vez mayor a formularlas en el marco de programas interactivos y fácilmente asequibles al usuario lingüista sin experiencia en utilización de equipos informáticos, usándose recientemente lenguajes como el LISPLOG. Tal es la idea de sistemas como el SRS o el Delta [20] que permiten el desarrollo de las reglas de síntesis para cualquier lengua y la posibilidad de comprobar de inmediato el resultado.

El formalismo empleado corresponde esencialmente al descrito, usado corrientemente desde la introducción del modelo generativo en fonología por Chomsky y Halle en 1968. Aunque recientemente se asista a un cambio en la concepción de las representaciones fonológicas de la lengua, éste aún no ha cuajado en un formalismo apto para fines computacionales.

14.4.4 Sistemas de conversión de texto a voz

Un sistema de conversión de texto a voz, como hemos indicado, tiene como objetivo pasar de una representación discreta del habla a una onda sonora continua. A grandes rasgos, suele tomar una configuración similar a la de la figura 14.13 [17, 20]. El texto en ortografía habitual, entrado a través del teclado del ordenador, debe en primer lugar ser procesado por un componente que lo modifica en varios aspectos: introducción de marcas morfológicas especificando la estructura de la palabra para determinar después su pronunciación, introducción de acentos en el caso de palabras que no llevan acento ortográfico, codificación de los signos de puntuación que deben dar origen a las variaciones de los elementos suprasegmentales, etc. Una vez obtenido el texto modificado actúan las reglas de conversión de grafema (unidad de la lengua escrita) a fonema, preparadas por el lingüista formalizando hechos que conoce toda persona capaz de leer como, por ejemplo, que en español se pronuncia <*u*> en *agua* [*aɣwa*] pero no en *anguila* [*aŋgila*] y otros casos más complejos. Junto a estas reglas suele incluirse un diccionario de excepciones que registra las palabras que las reglas de conversión no pueden

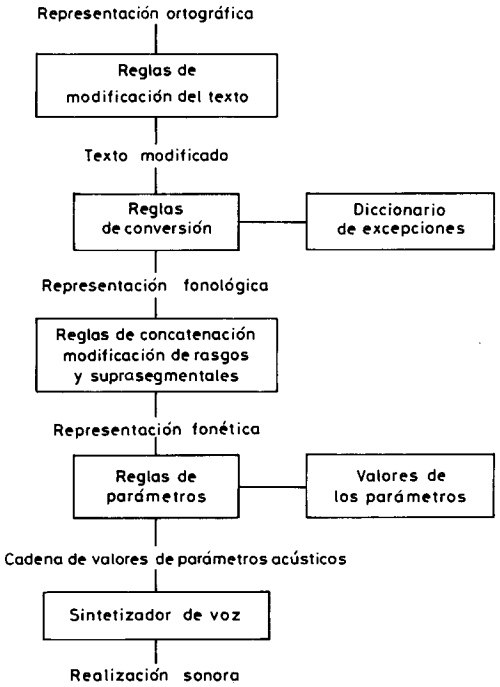

Figura 14.13 Esquema general de un sistema de conversión de texto a voz.

tratar correctamente, junto con casos como las siglas o las abreviaturas que suelen emplearse en la lengua escrita: por ejemplo, Sr. debe ser pronunciado en forma no abreviada *señor* [seɲor] [21].

Una vez realizado este proceso, se llega a una representación fonológica en la que cada elemento va asociado a una matriz de rasgos que lo especifican. Sobre esta cadena actúa un último módulo de reglas que determinan los rasgos que deben ser modificados en función del entorno —por ejemplo, un fonema con el rasgo sordo como /s/ adquiere el rasgo sonoro si va seguido de una consonante sonora, como en el caso de *mismo* que se pronuncia [mizmo]—, la concatenación de los elementos, y las modificaciones operadas por el acento y la entonación. Así puede generarse ya un fichero con los valores de los parámetros acústicos que controlan el sintetizador en función de los rasgos de cada uno de los elementos de la cadena fonética. Estas reglas determinan la posición de los formantes, el contorno de las transiciones y las variaciones de duración y de frecuencia fundamental o tono, datos que enviados al sintetizador configuran la realización sonora del enunciado [20].

Los sistemas de conversión de texto a voz pueden variar en cuanto al tipo de unidad usada —aunque en general se utiliza el fonema o el alófono—, la extensión del diccionario de excepciones, la formulación de las reglas de conversión, el modelo de parámetros adoptado o la forma de generarlos para el sintetizador. Como hemos visto, utilizando unidades como el difonema o la semisílaba, las transiciones entre los elementos forman ya parte de las unidades almacenadas, mientras que trabajando con fonemas es preciso calcularlas e interpolarlas por regla. Ciertas estrategias de síntesis utilizan en cambio un modelo articulatorio, especificando los parámetros en términos de configuraciones del tracto vocal definidas a partir de la posición de la lengua, labios, velo del paladar, etc. [22].

Es indudable que la síntesis constituye un campo relativamente avanzado pero, como ha escrito recientemente J. Allen, «no hay actualmente sistemas de síntesis que empiecen a acercarse al nivel de variabilidad fonética que se observa en el habla natural»; éste va a ser, probablemente, uno de los retos de la investigación en los próximos años.

14.5 BIBLIOGRAFIA

[1] F. Fallside, W.A. Woods (Eds.), *Computer Speech Processing*, Prentice-Hall, 1985.

[2] L.R. Rabiner, R. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[3] N. Jayant, P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, 1984.

[4] S.A. Schane, *Generative Phonology*, Prentice-Hall, 1973. Traducción española: *Introducción a la fonología generativa*, Labor, 1979.

[5] A. Cutler, R.D. Ladd (Eds.), *Prosody: Models and Measurements*, Springer-Verlag, 1983.

[6] G.J. Borden, K.S. Harris, *Speech Science Primer. Physiology. Acoustics and Perception of Speech*, Williams & Wilkins, 1980.

[7] A. Quilis, *Fonética acústica de la lengua española*, Gredos, 1981.

[8] J. Flanagan y otros, «Speech Coding», *IEEE Trans. on Communications*, vol. COM-27, pp. 710-737, Abril 1979.

[9] J. Makhoul, «Linear Prediction. A Tutorial Review», *Proc. IEEE*, vol. 63, pp. 561-580, Abril 1975.

[10] Federal Standard, «Telecommunications: Analogic to Digital Conversion of Voice by 2400 bits/second Linear Predictive Coding», FED-STD-1015.

[11] I.H. Witten, *Principles of Computer Speech*, Academic Press, 1982.

[12] D. Klatt, «Software for a Cascade Parallel Formant Synthesizer», *J. Acoust. Soc. Am.*, vol. 67, pp. 971-995, Marzo 1980.

[13] D. O'Shaughnessy, «Automatic Speech Synthesis», *IEEE Comm. Magazine*, vol. 21, n. 9, pp. 26-34, Dic. 1983.

[14] V. Zue, «Computer Voice Response and Speech Synthesis», *Trends and Perspectives in Signal Processing*, vol. 2, n. 4, pp. 7-9, Oct. 1982.

[15] P. Mermelstein, «Voice Message Systems», *IEEE Comm. Magazine*, vol. 21, n. 9, pp. 8-10, Dic. 1983.

[16] L.R. Rabiner, R.W. Schafer, «Digital Techniques for Computer Voice Response: Implementations and Applications» *Proc. IEEE*, vol. 64, n.º 4, pp. 416-433, Abril 1976.

[17] J. Allen, *From Text to Speech: the MITalk System*. Cambridge University Press, 1985.

[18] J.L. Flanagan, L.R. Rabiner (Eds.), *Speech Synthesis*, Dowden, 1973.

[19] S.D. Isard, D.A. Miller, «Diphone Synthesis Techniques», *Int. Conf. on Speech I/O Techniques and Applications*, IEE Conf. Publ. 258, pp. 77-82, 1986.

[20] S.R. Hertz, J. Kadin y K.J. Karplus, «The Delta Rule Development System for Speech Synthesis from Text», *Proc. IEEE*, vol. 73, pp. 1589-1601, 1985.

[21] M. Dilts, «Text to Speech», en G. Barlow (Ed.) *Electronic Speech Synthesis. Techniques, Technology and Applications*, Granada, pp. 94-113, 1984.

[22] M.P. Haggard, «Experience and Perspectives in Articulatory Synthesis», en B. Lindblom, and S. Ohman (Eds.), *Frontiers of Speech Communication Research*, Academic Press, pp. 259-274.