comunicación oral con el computador

Se ofrece una panorámica sobre los conceptos y las técnicas implicadas en la comunicación oral entre el hombre y el ordenador. Hacer hablar al ordenador (síntesis de voz) es una meta más próxima a ser alcanzada que dotarle de la capacidad de entender el mensaje oral. De momento, hemos de conformarnos con éxitos parciales en ambos propósitos.

Climent Nadeu y José B. Mariño

SPEECH COMMUNICATION WITH COMPUTERS

Concepts and techniques involved in speech communication with computers are reviewed. Although the speech output is, at present, closer to the computer ability than the speech understanding, at the moment, only partial results for both purposes have been attained.

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 el 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 universidad 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 base de datos, reserva y venta de billetes de avión o ferrocarril, 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 trabajo) de ofrecer una conversación natural.

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 1. Ventajas e inconvenientes de la comunicación oral hombre-máquina.

En la tabla 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 *reconocer* las distintas palabras del mensaje oral e interpretar los contenidos sintácticos y semánticos; 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.

CODIFICACION

La codificación consiste en la conversión de la señal analógica en una sucesión de bits apta para ser almacenada, transmitida o procesada digitalmente. Dicha sucesión puede ser simplemente la que se tiene a la salida del convertidor A/D, es decir, la obtenida con PCM. Para preservar la calidad de voz telefónica con este tipo de codificación, se requieren, si la cuantificación es uniforme, 12 bits/muestra que, a una velocidad de muestreo —normalizada en telefonía— de 8 kHz, significan 96 kb/s.

Por otro lado, si consideramos el habla fonológicamente; es decir, la contemplamos como una sucesión en el tiempo de fonemas conectados o separados por pausas, 72 b/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 —menos de 40— en la mayoría de lenguas. 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 el fonema, ente abstracto, engloba bajo su denominación toda una variedad de realizaciones acústicas distintas (alófonos) que a su vez son modificadas por las características prosódicas de tono, duración e intensidad. 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 se encuentra.

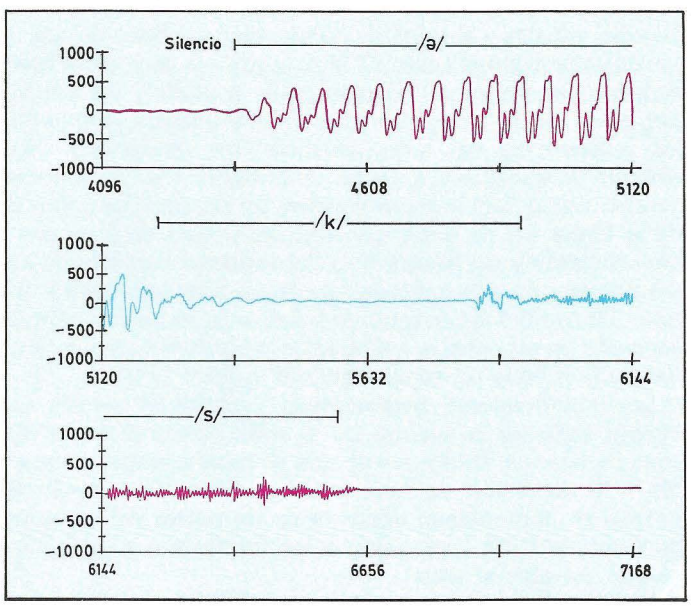

Figura 1. Señal de voz correspondiente a la primera sílaba de la palabra Expotrónica pronunciada en catalán. Sobre la figura se indica la extensión aproximada de cada fonema. La numeración en el eje temporal indica el número de muestras; la frecuencia de muestreo es de 8 kHz.

Esta 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 1, como 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 ancho de banda 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 [7]. En la primera, la finalidad primaria es mantener la

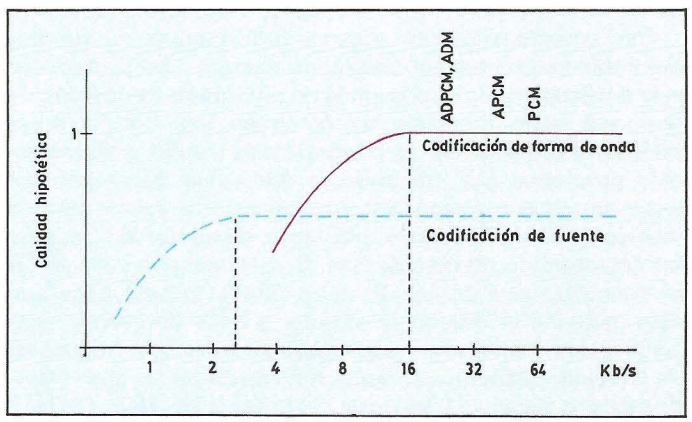

Figura 2. Relación calidad/comprensión de los dos tipos de codificación.

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 sofisticados. Los sistemas de la segunda clase consiguen eliminar redundancias basándose en un modelo de producción del habla y codificando la evolución temporal de sus parámetros significativos (*codificación de fuente*). En la figura 2 se muestra la diferencia de comportamiento de ambas clases de sistema por lo que refiere a la relación que establecen entre compresión de información y calidad del habla.

Codificación de forma de onda

La mayoría de codificadores de la foma de onda [4] buscan minimizar el error o valor cuadrático medio de la diferencia entre la señal original y su versión decodificada. Este error o ruido es el efecto resultante de la cuantificación (ver figura 3) y puede ser reducido, sin aumento del número de bits/s, a base de aprovechar ciertas características de la señal. Obviamente, ello equivale a reducir bits/s manteniendo la misma relación señal-ruido.

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, los cuales modifican el

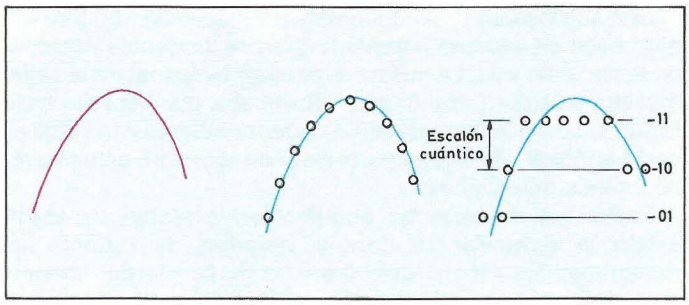

Figura 3. Señal analógica (a), discretización temporal (b) y discretización de amplitud (cuantificación) con 2 bits (4 niveles) (c).

escalón cuántico para hacerlo proporcional a la energía media de la señal en un intervalo corto de tiempo (PCM adaptativo: APCM). Esta estrategia consigue reducir el error en las zonas de baja energía de la señal, ya que el error es directamente proporcional al valor de l escalón cuántico.

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 cuantificación 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 implicando en la predicción 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/o los coeficientes del predictor codificados (información lateral).

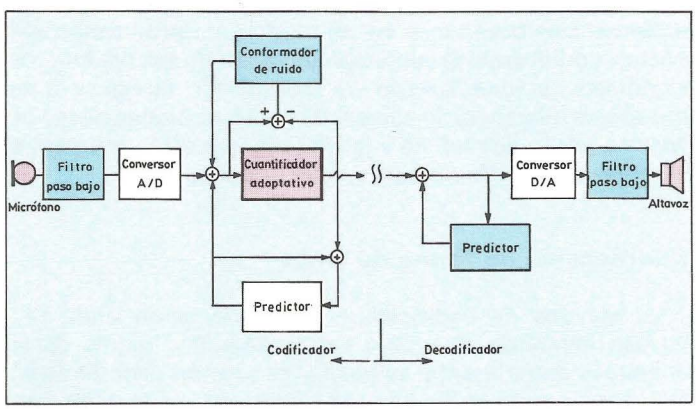

Figura 4. Esquema de un sistema completo de codificación y decodificación ADPCM. El filtro de entrada limita el espectro de la señal a la mitad de la frecuencia de muestreo a fin de evitar el solapamiento («aliasing») en la digitalización.

En la figura 4 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 [1].

Existen otras técnicas más complejas que no responden al esquema de la figura 4, tales como las que pasan la señal a un dominio transformado y allí la cuantifican (codificación en sub-bandas, ATC, etc.), o las que trabajan con vectores de muestras en lugar de muestras sueltas (cuantificación vectorial). Ninguna de dichas técnicas ha salido aún del laboratorio.

Codificación de fuente o paramétrica

Tal como se ha dicho, este tipo de codificación presupone un modelo paramétrico de generación de la señal. En el caso del habla, el modelo básico que se emplea es el de la figura 5;

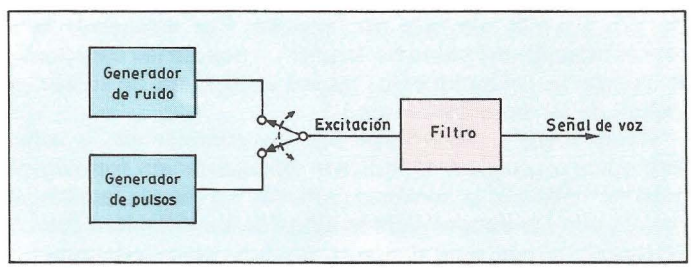

Figura 5. Modelo digital de producción de la voz.

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, (*voz sonora*) un ruido de banda ancha (*voz sorda*) o una combinación de ambos [3]. 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 /*a* o *e* sorda/ de la figura 1 es un fonema sonoro, por lo cual le corresponderá un espectro tal como el de la figura 6*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 6*b*, en el cual el detalle fino tiene un cariz errático o ruidoso [2].

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 aproxima la envolvente del espectro por medio de los coeficientes de predicción lineal (*modelo LPC*) [8], lo cual equivale a modelar el tracto vocal con un filtro cuya función de

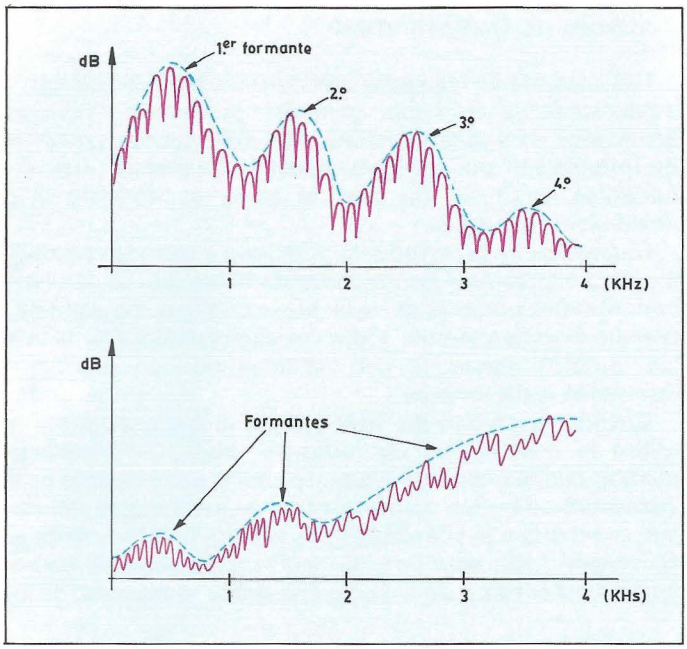

Figura 6. Espectros típicos de voz sonora (a) y sorda (b).

transferencia, que es la inversa de la del filtro de la figura 7, no presenta ceros, 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, en el caso de voz sonora, el valor del período del tono (se considera que la conformación de los pulsos es llevada a cabo también por el filtro).

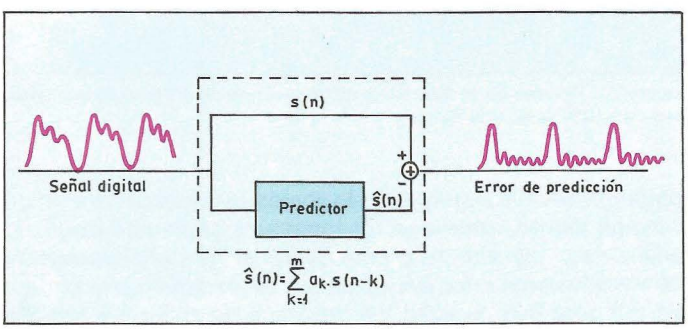

Figura 7. Filtro de predicción lineal. s(n) es la muestra de la señal en el instante *n*.

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 la 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 [3].

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 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 b/s, que aún puede ser reducido hasta 800 b/s si se cuantifican vectorialmente los parámetros. El máximo resulta ser, con los valores anteriores, de 6000 b/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.

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 [5]. 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 [1], [9]. Un inconveniente de este tipo de codificación reside en el análisis de los formantes. Dicho análisis es complejo, porque los formantes presentan un amplio margen de variación en frecuencia, y frecuentemente tiene lugar la confusión entre dos formantes cercanos.

SINTESIS

Lógicamente, la forma general de producir habla con un ordenador requiere el almacenamiento en forma codificada de la señal acústica correspondiente a unidades lingüísticas elementales (fonemas, sílabas, palabras, etc.) y la concatenación posterior de segmentos de voz formados a partir de ellas [5], [10].

Los sistemas de síntesis del habla se diferencian, pues, por el tamaño de las unidades elementales y el método de codificación utilizados. Cuanto más grandes son las unidades, mejor es la calidad del habla, pero mayor es la memoria precisada, puesto que el número de unidades crece exponencialmente con su tamaño. De la misma forma, por lo que se refiere al tipo de codificación, los métodos eficientes en cuanto a ahorro de memoria (codificación de fuente) producen una cierta degradación del habla. Así pues, maximización de la calidad del habla y minimización de la memoria requerida demandan un compromiso que dependerá de la aplicación concreta. Otros factores a minimizar además del espacio de memoria, son la complejidad algorítmica, la velocidad de cálculo y la dificultad de creación del vocabulario de unidades elementales.

Los sistemas prácticos de producción de habla con ordenador se pueden dividir según el tipo de aplicación en dos clases. Por un lado, podemos mencionar los sistemas de *respuesta oral por ordenador*, también llamados más explícitamente sistemas de almacenamiento y reproducción de mensajes orales, los cuales requieren tanto análisis como síntesis y utilizan primordialmente técnicas de procesado de señal. Por otro lado, cabe citar los más propiamente denominados *sintetizadores de habla*, cuya aplicación más sobresaliente es la conversión de lenguaje escrito a oral; estos sistemas utilizan unidades pequeñas de voz y requieren un procesado lingüístico extenso.

Respuesta oral por ordenador

La forma más familiar de producir respuesta oral automáti-

ca con técnicas analógicas es la utilización de 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 los distintos segmentos de voz almacenada. En cambio, la grabación en forma digital con codificación de forma de onda permite un acceso rápido a cualquier segmento de voz, a un relativamente bajo coste y sin deterioro apreciable en su calidad.

El sistema más simple de producción del habla graba enteras las expresiones orales a reproducir. Cuando el coste en memoria requerida resulta excesivo se hace indispensable la concatenación de pequeñas partes grabadas separadamente. Parece razonable, entonces, la participación en palabras (o frases cortas) y la posterior yuxtaposición de ellas. Esta es, actualmente, la forma más utilizada de síntesis del habla, pues aunque el número de palabras fonéticamente distintas sea muy elevado, muchas de las aplicaciones emplean un vocabulario reducido (100 ó 200 palabras suelen bastar).

Sin embargo, las frases formadas concatenando varias palabras grabadas por separado o en un contexto diferente se perciben con falta de naturalidad. La inserción de una palabra en una oración, cuando el ritmo y la entonación no son los apropiados, hace disminuir su probabilidad de ser entendida por el oyente, mientras que, si son los apropiados para conseguir una expresión natural, la aumenta (experimento de Stowe y Hampton, 1961) [5]. 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.

Dificultades de la concatenación

La primera dificultad a vencer, cuando se concatenan palabras u otras unidades de habla, es la coarticulación existente en el habla natural, es decir, la influencia que cada sonido ejerce sobre los adyacentes debido a las apreciables constantes de tiempo (decenas o centenas de milisegundos) de los transitorios producidos por el movimiento de los órganos configuradores del tracto vocal. La segunda dificultad es la de tipo prosódico; en este caso, no *únicamente deben considerarse* los sonidos adyacentes, sino que los efectos se extienden sobre segmentos mucho más largos que las palabras [5].

La codificación de forma de onda no permite subsanar estas dificultades, ya que la coarticulación no puede ser simulada por simple interpolación de la forma temporal de la señal y, aunque la entonación puede variarse quitando o añadiendo muestras en cada período, los cambios consiguientes en el espectro originan una perceptible degradación de la calidad del habla.

Sin embargo, tanto la coarticulación como las características prosódicas pueden ser incorporadas en la concatenación de segmentos del habla si el almacenamiento se lleva a cabo con la codificación de fuente. La razón está en la flexibilidad de los sintetizadores paramétricos que, al efectuar la separación de excitación y filtro, permiten modificar las características prosódicas (asociadas a la excitación), o realizar la interpolación o alisado de los parámetros del filtro (para simular el paso gradual de una posición a otra del tracto vocal), sin alterar las demás características.

Sistemas de texto limitado

Los sistemas de respuesta oral que yuxtaponen palabras u otras unidades de habla 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.

Generalmente la reproducción se lleva a cabo sin producir ninguna alteración a las unidades de habla que se concatenan. Aunque, como ya se ha dicho, la codificación de fuente permite modificar con facilidad las características de la voz, los intentos realizados con palabras no han producido resultados satisfactorios.

Cuando la calidad del habla es un factor importante se utiliza la codificación de forma de onda. Sin embargo, cuando puede permitirse cierta tolerancia en la calidad de la voz sintetizada o los requerimientos de memoria lo aconsejan, se recurre a la codificación de fuente, habitualmente en su versión LPC; este tipo de codificación requiere que el análisis automático de la voz para estimar los valores de los

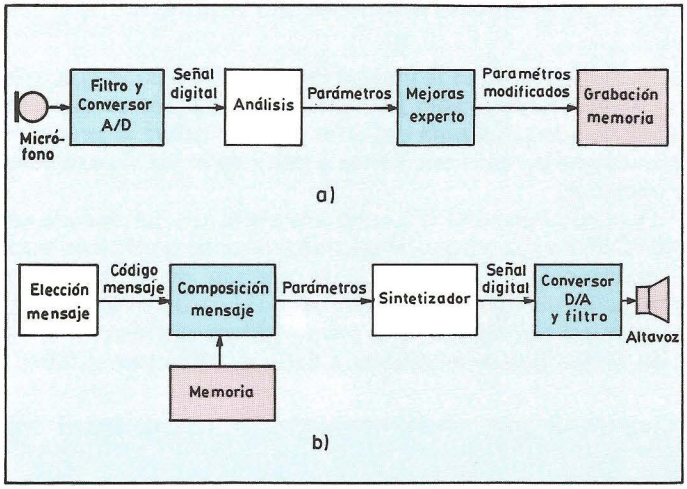

Figura 8. Proceso de análisis (a) y síntesis (b) en un sistema de respuesta oral con codificación de fuente.

parámetros vaya seguido por la acción de un experto a fin de corregir dichos valores en los intervalos de tiempo donde el análisis sea incorrecto o para codificar más eficientemente contenidos concretos sin introducir degradación perceptual; en la figura 8 se ilustran los procesos de análisis y síntesis implicados en este tipo de sistemas.

Síntesis por regla

En aplicaciones más generales en las que el texto es ilimitado (conversión de texto a voz) es necesario recurrir a unidades más pequeñas que las palabras, tales como fonemas o partes de sílabas. Ahora bien, la simple concatenación sin modificaciones de dichas unidades produce un habla inaceptable, puesto que la reducción de su tamaño conlleva una agudización notable de las dificultades ya citadas de la concatenación. Esta exigencia de adaptación al contexto de cada segmento codificado obliga a una flexibilidad en la representación de la voz que sólo puede obtenerse con codificación de fuente. Las unidades almacenadas en memoria se suelen generar artificialmente, es decir, no son el resultado de la codificación de habla natural.

El *fonema* es la unidad lingüística más pequeña. Si se utiliza el fonema como unidad básica de síntesis, los requerimientos de memoria son extremadamente reducidos (72 b/s, como ya se ha dicho). Sin embargo, la síntesis con fonemas exige la incorporación de un complejo conjunto de reglas que controlan la adaptación de cada fonema al contexto [1]. Si en lugar de almacenar un solo alófono por fonema se almacenan varios, se reduce el problema de la coarticulación; sin embargo, la elección de los alófonos apropiados a la expresión vocal a sintetizar resulta, entonces, más complicada.

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. Puesto que la parte central de muchos fonemas tiene un comportamiento estacionario, la transición entre difonemas es relativamente fácil de llevar a cabo. Por esto, el procedimiento de síntesis se concentra más bien en los aspectos prosódicos; por ejemplo, el ajuste de la duración temporal presenta una solución más difícil que en el caso de síntesis con fonemas.

Las semisílabas, tal como su nombre indica, abarcan la primera o la segunda mitad de una sílaba. Aventajan a los difonemas en la consecución de sílabas que empiezan o acaban con agrupaciones de consonantes difíciles de producir por concatenación de difonemas. Respecto de las sílabas, poseen la ventaja de ser mucho menos numerosas [5].

Para conseguir un buen resultado, la síntesis por regla requiere que la estructura sintáctica de cada oración esté especificada y que el vocabulario sea conocido. Parece que la pobre calidad del habla sintetizador, es causada en mayor medida por la carencia de un conocimiento suficientemente completo del conjunto de reglas, que por la falta de adecuación a la realidad del modelo de producción de la voz usado por los sintetizadores.

Conversión de texto a voz

En la figura 9 se ilustra un sistema general de conversión de texto a voz [11]. El primer paso requerido en tales sistemas y que introduce un importante factor de complejidad, es la necesidad de pasar del texto normal a una representación fonética (transcripción fonética) y prosódica del mismo, a la cual se han de aplicar las reglas de concatenación para determinar los parámetros que gobiernan el sintetizador.

Los sistemas desarrollados se basan en una combinación de síntesis por regla y de búsqueda en un diccionario de excepciones. La inclusión de dicho diccionario de palabras es debida a la existencia de palabras de frecuente aparición que son tratadas incorrectamente por las reglas. El tamaño, tanto del conjunto de reglas como del diccionario necesarios para conseguir un determinado tanto por ciento de palabras sin error, depende del idioma. En idiomas en los que la escritura sigue más de cerca a la fonética, el tamaño es más reducido. Así ocurre —por ejemplo— con el castellano frente al inglés.

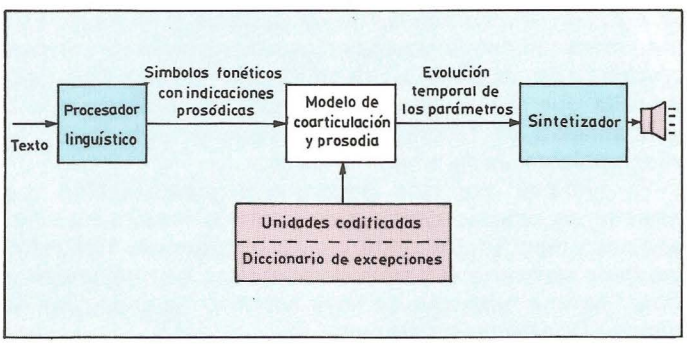

Figura 9. Sistema de conversión de texto a habla.

La unidad de habla que se utiliza es el fonema o alófono; unidades más grandes como la semisílaba o el difonema solo son consideradas, por ahora, en el marco de la investigación.

RECONOCIMIENTO

El interés final de la investigación en reconocimiento del habla hay que situarlo en el diseño de sistemas capaces de interpretar el mensaje oral tal como es producido por un interlocutor humano cuando se comunica con sus semejantes (*habla continua*).

Por el momento, sin embargo, únicamente las técnicas de reconocimiento de *palabras aisladas*, es decir, con un breve intervalo (200 ms) de silencio entre ellas, han llegado a su maduración. Buena parte de dichas técnicas son compartidas por los sistemas, actualmente en fase de experimentación, de reconocimiento del *habla conectada*, o sea, la que requiere una pronunciación cuidadosa de cada palabra.

Pero el tipo de habla no es el único aspecto a considerar en el reconocimiento. Son también factores determinantes de todo sistema, la dependencia o independencia respecto del hablante o locutor y el lenguaje utilizado (léxico, sintaxis, etc.). En cuanto al vocabulario, es importante el número de palabras y su parecido.

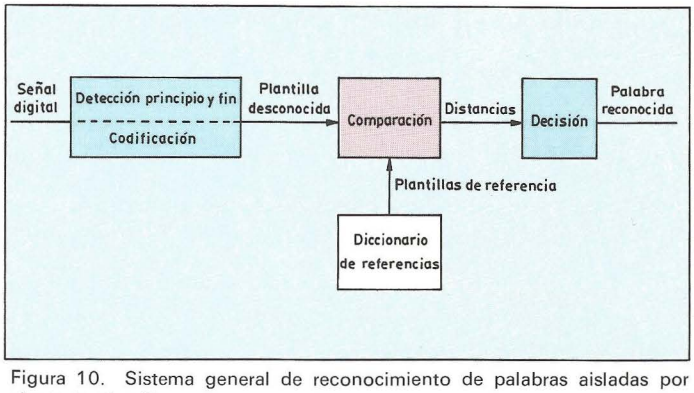

Figura 10. Sistema general de reconocimiento de palabras aisladas por ajuste de plantillas.

Reconocimiento de palabras aisladas por ajuste de plantillas

La figura 10 muestra un diagrama de bloques de un sistema de reconocimiento automático de palabras por ajuste de plantillas. La mayoría de sistemas se basan en este esquema, a pesar de que el contenido de cada bloque pueda ser muy distinto de un sistema a otro [13].

Básicamente, el procedimiento consiste en comparar la plantilla representativa de la palabra sujeta a reconocimiento con cada una de las plantillas de referencia, que pueden ser una o más por cada palabra del vocabulario. La comparación se lleva a cabo evaluando una medida de distancia entre plantillas que dependerá del tipo de codificación que se utilice.

Cada plantilla consta de un conjunto de vectores de parámetros codificados. Dichos vectores, uno por cada segmento de 10-30 ms de la señal, representan, en forma discretizada, la variación temporal del espectro de la señal dentro del intervalo temporal correspondiente a la palabra. La estimación espectral, en los sistemas más desarrollados, se basa, o bien en un análisis de predicción lineal o bien en un banco de filtros a cuya salida se mide la energía media —por segmento de la señal— de cada banda frecuencial. Lógicamente, junto con la codificación es preciso delimitar con cierto cuidado el intervalo temporal en el que la señal transporta la palabra, separando el habla del silencio o de ruidos ambientales. Muchos métodos usan para tal fin, como característica predominante, la medida de la evolución de la energía. La complejidad requerida depende de la relación señal-ruido; por ejemplo, en la señal de la figura 1, el silencio inicial es fácilmente discriminable de la voz porque la grabación fue realizada en un ambiente silencioso.

Puesto que una palabra pronunciada en dos ocasiones por una misma persona o por dos personas diferentes presenta duraciones temporales distintas en cada una de sus partes, se hace necesario algún tipo de alineamiento entre las palabras que se comparan. Existen diversas posibilidades de realizar dicho alineamiento, pero las técnicas basadas en la programación dinámica han demostrado ser sumamente útiles en una amplia gama de sistemas de reconocimiento del habla [14]. En ellas, el alineameinto se consigue, tal como se ilustra en la figura 11, ensanchando la plantilla en

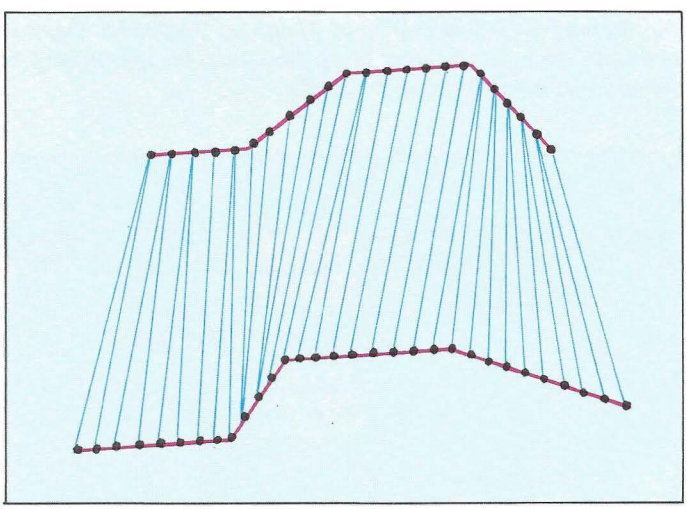

Figura 11. Ejemplo de alineamiento temporal óptimo (con restricciones) de plantillas unidimensionales. No se permite que a un punto le correspondan más de dos en la otra plantilla.

ciertos puntos y comprimiéndola en otros hasta encontrar una solución óptima en el sentido de minimización de la distancia total entre ambas palabras, calculada como suma normalizada de las distancias entre los vectores de parámetros que están relacionados por la función de alineamiento.

Los sistemas de reconocimiento independientes del locutor se distinguen de los que están adaptados a un solo hablante por la forma de construirse el diccionario de referencia [1]. En estos últimos, el locutor pronuncia una o varias veces cada palabra durante la fase de adiestramiento y con ellas se forman una o más plantillas por palabra (típicamente dos o tres). En los sistemas independientes del locutor el adiestramiento se realiza sobre una porción estadísticamente significativa del conjunto de los usuarios. Puesto que en ellos se usan muchas más plantillas por palabra, el tamaño del vocabulario es considerablemente menor.

Finalmente debemos considerar la estrategia de decisión, la cual puede ser tan simple como la elección de la palabra correspondiente a la plantilla más próxima, o más complicada, comparando —por ejemplo— promedios de distancias a varias plantillas de la misma palabra (procedimiento KNN). Como se expondrá más adelante, lo que a veces se requiere no es la elección de una sola palabra, sino el suministro de un conjunto de etiquetas del diccionario ordenadas de acuerdo con la distancia calculada, con el fin de pasar a un nivel superior de reconocimiento, tal como un analizador sintáctico.

En sistemas dependientes del locutor se han logrado tasas de reconocimiento superiores al 99 %, por lo menos cuando el vocabulario no supera unos pocos cientos de palabras, y éstas no son excesivamente similares. En el caso de independencia del locutor la tasa es un poco inferior, pero con vocabularios reducidos se ha superado el 98 %. Cuando las condiciones acústicas de la voz o del ambiente son distintas a las existentes en el adiestramiento, la tasa de reconocimiento es aún inferior.

Reconocimiento fonético de palabras aisladas

En aplicaciones con grandes vocabularios, el sistema de reconocimiento por ajuste de plantillas no resulta práctico por razones de memoria y tiempo de proceso requeridos. En esta situación adquiere especial interés la técnica de reconocimiento basada en una representación fonética de las palabras.

Las palabras del diccionario son almacenadas en forma de secuencias de símbolos fonéticos. La estrategia de reconocimiento posee dos o más niveles. En el primero, a partir de la señal y de un conjunto de referencias, se transforma la palabra desconocida en una secuencia de símbolos fonéticos. En el segundo, dicha secuencia es comparada con las del diccionario de palabras. La comparación se lleva a cabo penalizando convenientemente las diferencias a nivel de símbolo (ausencia, cambio o presencia errónea de un símbolo).

Reconocimiento y comprensión de habla continua

En la conversación natural, la señal acústica correspondiente a una palabra —u otra unidad lingüística— concreta presenta muy diferentes realizaciones, en función de la prosodia de la frase, el cuidado que se haya puesto en la pronunciación, las palabras o fonemas adyacentes, el locutor, etc. Además, las pausas entre palabras son mínimas, y, en ocasiones, prácticamente inexistentes. Esta situación provoca que el rendimiento de los sistemas de reconocimiento, aun de los más sofisticados, baje a cotas del 80 % de aciertos en el reconocimiento de palabras o fonemas. Esta tasa prácticamente incapacita a un reconocedor de palabras aisladas para abordar la identificación de una frase, por sencilla que sea. A título de ejemplo, considérese que la probabilidad de reconocer correctamente una frase con cinco palabras sería inferior al 33 %.

Es evidente, por otro lado, que la comunicación oral supone un proceso cognoscitivo, lo que implica que está asociada inescrutablemente con la inteligencia. Por tanto, no debe esperarse el diseño de máquinas que entiendan el habla natural hasta que se haya adquirido la capacidad de simular la inteligencia humana.

El reconocimiento del habla continua precisa explotar la

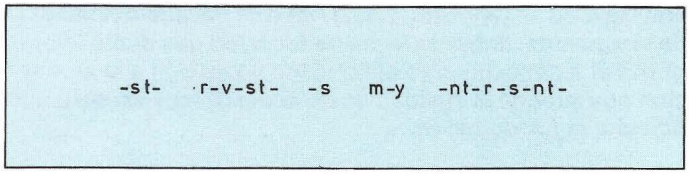

Figura 12. Ejemplo de redundancia en el lenguaje: la omisión de todas las vocales no impide el reconocimiento de la frase.

redundancia del lenguaje para paliar los errores que, inevitablemente, se van a producir al procesar la señal acústica. La presencia de esta redundancia puede revelarse mediante el ejemplo de la figura 12; aunque se han omitido todas las vocales de la frase, ésta es perfectamente identificable.

Esta redundancia del lenguaje es fruto del conocimiento previo común a los interlocutores humanos relativos al léxico, la sintaxis y la semántica. Consistentemente con ello, los sistemas de reconocimiento de habla continua, actualmente en experimentación, incorporan, además de un *procesador acústico,* un *procesador lingüístico*, los cuales interactúan (o no) entre sí. En la figura 13 se ilustra esta arquitectura.

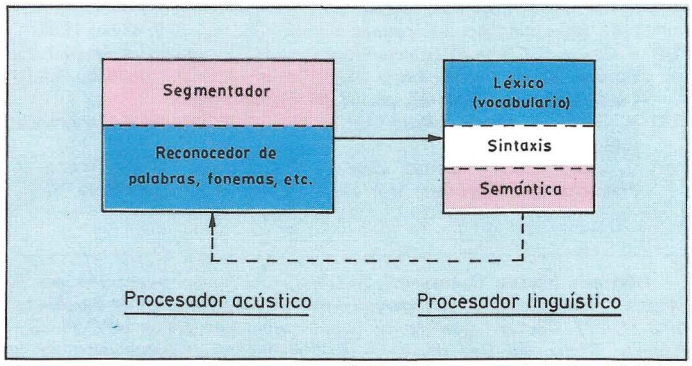

Figura 13. Esquema básico de un sistema de reconocimiento del habla continua.

El procesador acústico proporciona un conjunto de hipótesis sobre los componentes acústicos (palabras, fonemas, etc.) de la frase, con una etiqueta indicativa del grado de acuerdo entre la señal acústica y las plantillas de referencia de que dispone el procesador correspondiente al componente.

El procesador lingüístico recoge las hipótesis del procesador acústico y genera hipótesis de un nivel superior, también con su correspondiente etiqueta. Si el procesador acústico proporciona información sobre fonemas o partes de sílabas, el primer nivel del procesador lingüístico será el léxico. El nivel sintáctico confeccionará hipótesis sobre la frase a reconocer, que el reconocedor semántico finalmente ha de aceptar o rechazar. Evidentemente, los distintos niveles no han de ser necesariamente compartimentos estancos, ya que el nivel superior puede orientar al inferior en la confección de hipótesis. Esta estructura genérica ha dado lugar a múltiples versiones de máquinas para reconocer el habla continua que se diferencian en la información fonológica procesada (palabras, fonemas, difonemas, semisílabas, etc.), en el procedimiento de reconocimiento, en la forma de incorporar el conocimiento sintáctico-semántico y en la interrelación entre los distintos niveles de hipotesización.

Un ejemplo sencillo de reconocedor de habla continua lo proporciona una máquina que ha de identificar una frase entre un conjunto dado, cuyo procesador acústico proporciona hipótesis sobre palabras. La información lingüística se incorpora mediante una lista de las frases admitidas; el procesador acústico proporciona varios candidatos (de entre el vocabulario que reconoce) para la primera palabra de la frase, con su correspondiente etiqueta de similitud, y lo mismo hace para el resto de las palabras de la frase. El procesador lingüístico toma como primera hipótesis, la frase constituida por las palabras candidatas para cada posición con mayor similitud; si la frase es rechazada (no pertenece al conjunto a reconocer), genera como hipótesis la frase que, partiendo de la anterior y sustituyendo una palabra, tiene mayor similitud; y así sucesivamente. Esta máquina (muy rudimentaria en su concepción, aunque puede llegar a ser muy complicada en su realización) ha sido aplicada con éxito en la búsqueda de información en un directorio telefónico. Tomando como vocabulario las 26 letras del abecedario, cuya probabilidad de ser reconocidas correctamente era el 79 %, Rosenberg y Schmidt obtuvieron una tasa de 92 % de aciertos en la solicitud de información telefónica a un directorio de la Bell con 18.000 teléfonos [13].

La utilización de la teoría formal del lenguaje, introducida por Chomsky, como vehículo de la información sintáctico-semántica, ha permitido obtener resultados interesantes con máquinas experimentales diseñadas para aplicaciones concretas, tales como la venta telefónica de billetes de avión [13], el juego del ajedrez (DRAGON, HEARSAY) [6], el suministro de información sobre análisis químicos (SPEECHLIS), etc. [15].

Todos los sistemas de reconocimiento del habla continua mejoran apreciablemente sus prestaciones cuando el locutor pone especial cuidado en la pronunciación de las palabras, procurando marcar convenientemente cada una de ellas («*habla conectada*»).

PANORAMA ACTUAL E INMEDIATO

Cuando una tecnología sale del laboratorio y encuentra su lugar en el mercado, la demanda genera nuevas necesidades que impulsan a los equipos de investigación a realizar progresos con rapidez. Pues bien, parece haber llegado ya el momento, para los productos que tratan con el habla, de su primer gran impacto en el mercado, por lo que es previsible asistir a desarrollos notables en los próximos años.

En efecto, desde que hace un lustro se comercializó el primer chip sintetizador LPC, ha aparecido un apreciable número de chips sintetizadores haciendo uso de esa y otras técnicas. Ello, unido al desarrollo de los chips especializados en procesado de señal y el rápido abaratamiento de las memorias, está permitiendo la salida al mercado de sistemas de conversión de texto a voz a precios que harán aumentar notablemente el índice de ventas. Es en este área donde se pueden esperar avances más rápidos; y no sólamente serán debidos a un mejor conocimiento del dinamismo de la coarticulación, sino que las nuevas técnicas de codificación también tendrán algo que aportar.

En cuanto al reconocimiento, los sistemas actualmente en venta que ofrecen prestaciones aceptables, tratan con palabras aisladas y son dependientes del locutor [16]. Se puede esperar, no obstante, que muy pronto aparezcan sistemas de habla conectada, independientes del locutor, con vocabularios reducidos y una tasa de reconocimiento superior a 97 %. Asimismo, el dictado automático puede ser

una realidad en japonés a finales de esta década o principios de la siguiente, debido a la mayor facilidad que dicho idioma ofrece al reconocimiento automático y también a la motivación que supone la realidad de un alfabeto muy extenso que dificulta la transcripción. ●

REFERENCIAS

[1] Número 144 de Mundo Electrónico, dedicado especialmente a Reconocimiento y Síntesis de voz, Octubre 1984.
[2] E. Martínez Celdrán, «Fonética», Ed. Teide, 1984.
[3] L. Rabiner, R. Schafer, «Digital Processing of Speech Signals», Prentice-Hall, 1978.
[4] N. Jayant, P. Noll, «Digital Coding of Waveforms», Prentice-Hall, 1984.
[5] I. Witten, «Principles of Computer Speech», Academic Press, 1982.
[6] W. Lea, Ed., «Trends in Speech Recognition», Englewood Cliffs, NJ: Prentice-Hall, 1980.
[7] J. Flanagan et al., «Speech Coding», IEEE Trans. on Communications, vol. COM-27, pp. 710-737, Abril 1979.
[8] J. Makhoul, «Linear Prediction: A Tutorial Review», Proc. IEEE, vol. 63, pp. 561-580, Abril 1975.
[9] D. Klatt, «Software for a Cascade/Parallel Formant Synthesizer», J. Acoust. Soc. Am., vol. 67, pp. 971-995, Marzo 1980.
[10] D. O'Shaughnessy, «Automatic Speech Synthesis», IEEE Comm. Magazine, Vol. 21, n.º 9, pp. 26-34, Diciembre 1983.
[11] J. Allen, «Synthesis of Speech from Unrestricted Text», Proc. IEEE, vol. 64, pp. 433-442, Abril 1976.
[12] V. Zue, «Computer Voice Response and Speech Synthesis», Trends and Perspectives in Signal Processing, vol. 2, n.º 4, pp. 7-9, Octubre 1982.
[13] L. Rabiner, S. Levinson, «Isolated and Connected Word Recognition», IEEE Trans. on Communications, vol. 29, pp. 621-659, Mayo 1981.
[14] H. Sakoe, S. Chiba, «Dynamic Programming Algorithm Optimization for Spoken Word Recognition», IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, pp. 43-49, Febrero 1978.
[15] N. Dixon, T. Martin, «Automatic Speech and Speaker Recognition», IEEE Press, 1979.
[16] G. Doddington, T. Schalk «Speech Recognition: Turning Theory to Practice», IEEE Spectrum, Vol. 18, n.º 9, pp. 26-32 septiembre 1981.

Climent Nadeu Camprubí. *Ingeniero de Telecomunicación por la Escuela Técnica Superior de Ingenieros de Telecomunicación de Barcelona, UPC (1977). Doctor Ingeniero de Telecomunicación por la UPC (1982). Profesor Titular del Departamento de Tratamiento y Transmisión de la Información de la ETSITB. Interesado en análisis espectral y tratamiento del habla. Premio J. Bertrán Marqués, 1984.*

José B. Mariño Acebal. *Ingeniero de Telecomunicación por la Escuela Técnica Superior de Ingenieros de Telecomunicación de Madrid (1972) y Doctor Ingeniero de Telecomunicación por la E.T.S.I.T. de Barcelona (1975). Catedrático del Departamento de Tratamiento y Transmisión de la Información de la E.T.S.I.T. de Barcelona. Interesado en el tratamiento de señal y sus aplicaciones a radar y voz.*