Antecedentes y desarrollo de los sistemas actuales de reconocimiento automático del habla[1]

(Historical evolution in the current Automatic Speech Recognition System)

Varona Fernández, Amparo
Univ. del País Vasco
Dpto. de Electricidad y Electrónica
Apdo. 644
48080 Bilbao

BIBLID [1137-4411 (1997), 4; 321-346]

Se presenta un estudio de la evolución histórica de los sistemas de reconocimiento automático del habla continua. Estos sistemas se dividen fundamentalmente en dos etapas: la síntesis y el análisis. Un sintetizador es un instrumento capaz de reproducir la voz humana mientras que un analizador es un sistema capaz de reconocer la voz humana. Actualmente existen versiones muy desarrolladas de sistemas de síntesis, sin embargo el análisis continua siendo un problema abierto, muy importante en el campo de la inteligencia artificial. Por último, se realizan una serie de consideraciones sobre las perspectivas futuras en este área.

Palabras Clave: Reconocimiento del habla. Síntesis y análisis.

Artikulu honek, irakurleak mintzo jarraiaren ezagutza-sistema baten garapen historikoa, sistema osatzen duten atal guztiak, zenbait arazo espezifiko, hurbilketa formalak eta kasu bakoitzean proposatzen diren soluzioak ezagutu ahal izango ditu. Mintzo ezagutza automatikoaren sistema bi parte dauzka: sintesi eta analisi. Sintesiak mintzoa berregitzen du eta analisia mintzoa ezagutzen du. Gaur egun, sintesiak garapen handia dauka baina edozein hizlarirentzat baliogarria, hiztegi handiak eta gramatika naturalak erabiltzen dituen mintzo jarraiaren ezagutza, adimen artifizialaren arloan dagoen arazo garrantzitsuenetariko bat da. Azkenik, arlo honetako ikerketen etorkizun hurbila aztetuko da.

Giltz-Hitzak: Hizketaren ezagutza. Sintesia eta analisia.

On présente une étude de l'évolution historique des systèmes de reconnaissance automatique de la parole. Ces systèmes se divisent fondamentalement en deux étapes: la synthèse et l'analyse. Un synthétiseur est un instrument capable de reproduire la voix humaine alors qu'un analyseur est un système capable de reconnaître la voix humaine. Il existe actuellement des versions très développées de systèmes de synthèse; néanmoins, l'analyse continue toujours d'être un problème de très grande importance dans le domaine de l'intelligence artificielle. Pour terminer, on réalise une réalise une série de considérations sur les perspectives futures dans ce domaine.

Mots Cles: Reconnaissance de la langue. Synthèse et analyse.

1 Trabajo parcialmente financiado por el proyecto CICYT (TIC95-0884-C04-03)

1. INTRODUCCIÓN

El hombre desde siempre se ha sentido fascinado por su capacidad de hablar y escuchar. Sin esta habilidad para comunicar fluidamente ideas la sociedad actual, probablemente no habría podido desarrollarse. De aquí nace la obsesión por crear instrumentos capaces de producir y reconocer voz, imitándonos a nosotros mismos.

Un sistema de síntesis o sintetizador de voz es un instrumento capaz de producir una señal acústica que imita la voz humana. Mientras que un sistema de análisis o reconocimiento de voz puede ser definido como cualquier mecanismo, distinto del sistema auditivo humano, capaz de descodificar la señal acústica producida por el aparato fonador de un locutor, en una secuencia de unidades lingüísticas que contienen el mensaje que ese locutor desea comunicar. El nivel de desarrollo de cada uno de estos campos no es el mismo. Aunque se pueden encontrar en el mercado sintetizadores de muy altas prestaciones, los sistemas de reconocimiento son aún muy restrictivos: palabras aisladas, vocabulario muy limitado, dependencia del locutor, etc. Los sistemas de reconocimiento automático del habla para grandes vocabularios, discurso continuo e independientes del locutor, se están desarrollando aún con mayor o menor éxito en laboratorios de investigación de universidades o en departamentos de I+D de grandes empresas del mundo de la informática y de las telecomunicaciones (IBM, AT&T, Philips).

Tanto el proceso de síntesis como el de reconocimiento automático del habla se sitúan en un marco más general llamado *procesamiento de voz*, que incluye además problemas tan diversos como codificación de voz, traducción automática e identificación del locutor y de la lengua. Aunque los objetivos de cada una de estas ramas son completamente distintos, comparten conceptos, formulaciones y técnicas, lo cual da cuerpo al procesamiento de voz como disciplina autónoma englobada en el procesamiento de señales.

El objetivo básico del procesamiento de voz es la comunicación hombre-máquina. Esta idea abarca un amplio espectro de aplicaciones: acceso a sistemas de información automáticos, ayuda a minusválidos, traducción automática, transacciones bancarias automáticas, control oral de sistemas, etc.

El objetivo de este trabajo es desarrollar la evolución histórica sufrida tanto por los sistemas de síntesis, como de reconocimiento a lo largo de los dos últimos siglos. Tras una introducción que pone de relieve la importancia de ambos dispositivos dentro del problema general del procesamiento de señal, en la sección segunda se estudia el desarrollo histórico sufrido por los sistemas de síntesis. Esta sección se divide en dos sub-secciones. En la primera, se estudian los primeros dispositivos mecánicos y eléctricos y en la segunda los dispositivos eléctricos y electrónicos (computadores) que se han desarrollado a partir de la información dada por el espectrograma de la señal acústica. En la tercera sección se describe el desarrollo histórico de los sistemas de análisis o reconocimiento. Esta sección se divide de nuevo en dos sub-secciones. En la primera se hace un estudio de los primeros dispositivos eléctricos desarrollados, para pasar en la segunda a estudiar el gran avance de los sistemas de reconocimiento debido al desarrollo informático. Finalmente, el trabajo termina con una serie de consideraciones sobre las perspectivas presentes y futuras y una lista de referencias bibliográficas a las cuales se puede remitir el lector más interesado.

2. SÍNTESIS DE VOZ

Los primeros intentos de producción artificial de voz humana, se realizaron mediante dispositivos mecánicos. El siguiente paso consistió en la construcción de dispositivos eléctricos, para llegar en los últimos años a sistemas desarrollados gracias al creciente avance de la informática.

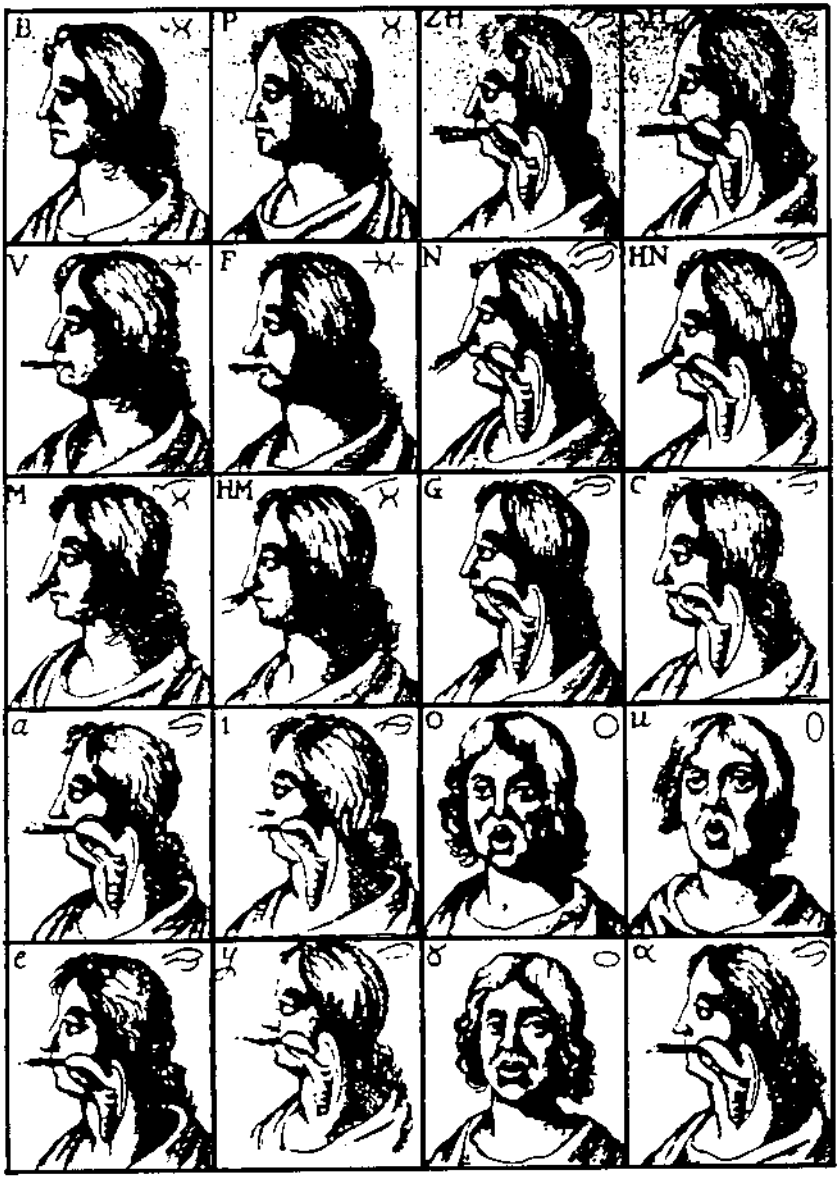

Figura 1. Diferentes posiciones del tracto vocal publicadas por B.J. Wilkins en 1668

La importancia histórica de estos dispositivos queda patente si se tiene en cuenta la incidencia que la comprensión de los mecanismos de producción de la voz humana ha tenido sobre el conjunto de técnicas de procesamiento de voz, y sobre el reconocimiento automático del habla en concreto. A continuación se va a llevar a cabo un recorrido histórico a través de los logros más relevantes.

Los primeros trabajos que se relatan relativos al estudio de la voz datan de 1668, cuando B.J. Wilkins publicó un libro en el que mostraba las posiciones del tracto vocal para diferentes caracteres alfabéticos (Figura 1). Propuso un alfabeto fonético, donde los símbolos representaban las posiciones de la boca al pronunciar los distintos sonidos. Su alfabeto consistía de 8 vocales y 26 consonantes, que representaban fundamentalmente a los sonidos ingleses según Poulton [1983].

Pero lo que es en sí la historia del procesamiento de voz comienza con los primeros dispositivos mecánicos capaces de sintetizar voz humana.

2.1. Primeros dispositivos

Uno de los primeros trabajos documentados fue presentado por el fisiólogo alemán C.G. Kratzenstein en 1779 cuando la *Academia Imperial de San Petersburgo* ofreció un premio para aquél trabajo que diera una explicación a las diferencias fisiológicas entre los cinco sonidos vocálicos y consiguiera un aparato de demostración para reproducir dichos sonidos, según Flanagan [1972]. El aparato consistía en cinco tubos con diferentes formas (Figura 2). Los tubos para los sonidos 'A', 'E', 'O' y 'U' estaban equipados con una lengüeta, mientras que por el tubo para la 'I' se soplaba directamente. Les dio estas extrañas formas para intentar crear las mismas resonancias que las que se producían en el tracto vocal cuando un locutor humano pronunciaba esos sonidos.

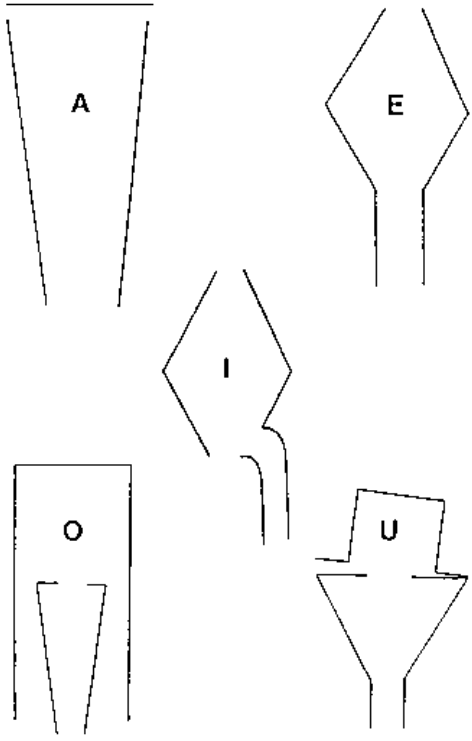

Figura 2 . Tubos construidos por C.G. Kratzenstein en 1779 para reproducir los sonidos vocálicos

Aproximadamente por esa misma época (1769), W.R. Kempelen, ingeniero y arquitecto húngaro, bien situado en el gobierno de su monarquía, ya había comenzado a trabajar en su *máquina parlante* (*speaking machine*). Se trataba de un dispositivo mucho más sofisticado, que era capaz de producir también sonidos consonánticos. Esta tarea le ocupó más de 20 años y sus resultados fueron publicados en un extenso volumen de 456 páginas, en 1791.

Los científicos de la época no tomaron en serio la máquina parlante de Kempelen, a pesar del gran avance que implicó. El motivo fue la decepción provocada por un trabajo anterior suyo: la máquina del *juego del ajedrez*, que se exhibió en prácticamente toda Europa. Dibujó un tablero de ajedrez sobre una mesa, detrás de la cual estaba sentada la figura de un turco. Se suponía que la máquina era lo suficientemente *inteligente* como para decidir las jugadas a desarrollar ante los movimientos de otro jugador. Como el propio Kempelem admitiera posteriormente, el principal componente de esta máquina era un hombre sin piernas, oculto debajo de la mesa. Su nombre era Worousky y había sido un antiguo comandante del régimen polaco y un experto jugador de ajedrez.

Sin embargo, su máquina parlante (Figura 3) fue el resultado de una gran cantidad de pruebas y errores, durante las cuales en al menos tres ocasiones tiró completamente el diseño y comenzó de nuevo. Se dio cuenta que la mejor fuente de sonido para imitar las cuerdas vocales era el zumbido que una lengüeta provocaba debido al paso de aire a su través. El aire se suministraba a través de un fuelle y se recogía en una cámara de aire comprimido, según Casacuberta [1987].

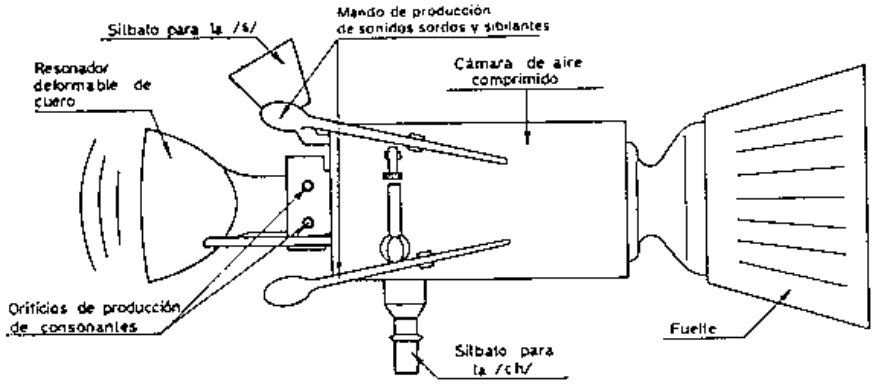

Figura 3. "Máquina parlante" de W. Kempelen (siglo XVIII)

Posteriormente, él mismo modificó el forro de la lengüeta por una piel blanda, para que los sonidos vocálicos no se produjeran con demasiada brusquedad. Las vocales se producían gracias a una cámara en forma de campana unida a la lengüeta. Las propiedades resonantes del timbre se alteraban colocando la mano izquierda sobre la abertura de salida. Dependiendo de la posición exacta de la mano, se producían los diferentes sonidos vocálicos. Dos pequeños orificios localizados más allá de la lengüeta pero antes de la campana, normalmente tapados con los dedos de la mano derecha, se abrían para producir los sonidos *nasales* 'm' y 'n'. El sonido 'l' se producía al dividir la corriente de aire en el timbre con el pul-

gar, de la misma forma que la lengua divide la corriente de aire en la boca para este sonido. Las *oclusivas* 'p' 'b' 't' 'd' 'k' 'g' se producían cerrando todos los orificios y ejerciendo presión para obstruir la cámara de aire comprimido y entonces se quitaba la mano de repente. Un fuelle auxiliar daba energía extra a esas oclusivas. Se utilizaron diferentes resonadores para producir los sonidos 's' y 'sh'. El aire se suministraba a las cámaras resonantes en diferentes niveles. Finalmente, la 'f', 'v', 'h' y el sonido germano 'ch' se podían generar permitiendo que el aire de la cámara de aire comprimido se escapara suavemente, o bien con alta presión para la 'f' o bien con baja presión para la 'h'.

Esta máquina producía sonidos que eran comunes a todas las lenguas europeas. El inventor aseguraba que cualquier persona podía lograr en tres semanas realizar síntesis de voz, realmente sorprendente, en las lenguas francesa, italiana y latina. El alemán era más difícil por la prevalencia de los sonidos consonánticos. Usando la descripción dada por Kempelen, C. Wheatstone, físico ingles nacido el 6 de febrero de 1802 en Gloucester y muerto el 19 de octubre de 1875 en París, construyó una versión mejorada de la máquina parlante (Figura 4).

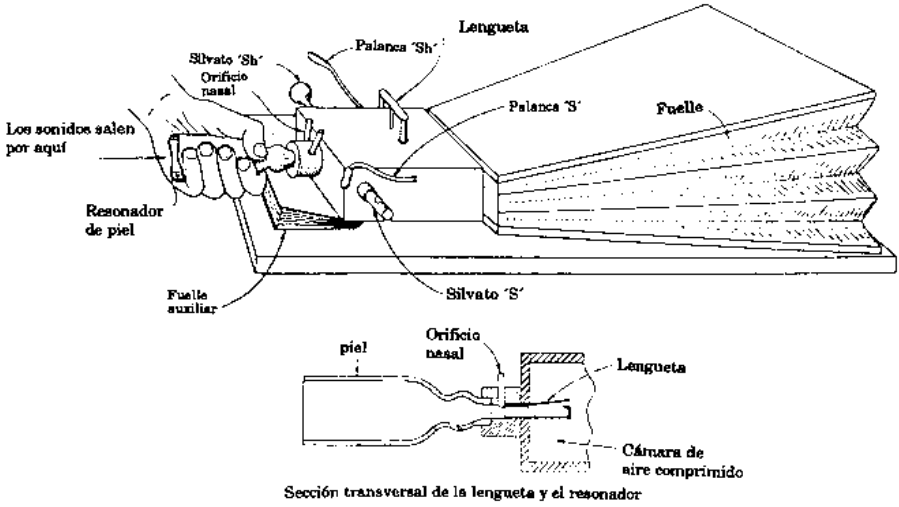

Figura 4. Máquina parlante de G. Wheatstone presentada en 1835

Wheatstone (Figura 5) hizo un demostración en la reunión de la *Asociación británica para el progreso de las ciencias* en Dublín en 1850. La principal diferencia consistía en la sustitución del timbre de hierro por un resonador de piel.

A. G. Bell (Figura 5), físico norteamericano inventor del teléfono nacido el 3 de marzo de 1847 en Edimburgo y muerto el 2 de agosto de 1922 en Nueva Escocia, siendo un niño en Edimburgo, tuvo la oportunidad de ver la construcción hecha por Wheatstone de la máquina parlante. Está le impresiono bastante y con el apoyo de su hermano Melville y de su padre A. M. Bell, construyó la máquina parlante sucesora del dispositivo de Wheatstone. Se propusieron la construcción de un aparato lo más parecido posible al aparato fonador humano usando materiales como goma, algodón, alambre, etc. El dispositivo conseguía sonidos vocálicos y nasales muy satisfactorios e incluso frases.

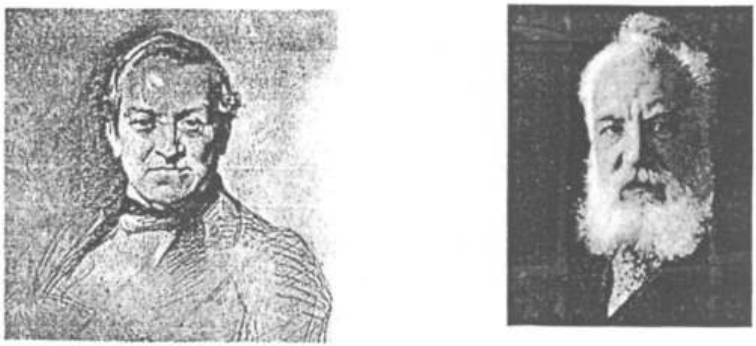

Figura 5. Retratos de G. Wheastone a la izquierda y A.G. Bell a la derecha

Nuevos dispositivos mecánicos destinados a la modelización y síntesis continuaron desarrollándose a lo largo del siglo XIX y principios del XX. En 1846 J. Faber construyó una máquina llamada *Euphonia*. Este instrumento representó un avance significativo sobre la máquina de Kempelen porque era posible variar el tono fundamental, lo cual permitía producir voz normal, en susurro, entonar las preguntas, etc. Otra máquina que incluía esta aproximación, propuesta por H.L.F. Helmoltz en 1875, usaba diapasones para producir las vocales artificialmente.

La invención del gramófono en el último cuarto del siglo XIX abrió las puertas a nuevas posibilidades para la investigación de la voz humana.

El primer sistema desarrollado fue el fonógrafo. Este aparato consistía en un cilindro con ranuras en forma de hélice, recubierto con una delgada hoja de estaño solidario de una barra roscada; se desplazaba delante de una bobina acústica cerrada por un lado por un diafragma de pergamino, extendido sobre los bornes de una pequeña caja cilíndrica. Un estilete redondeado, pegado sobre el diafragma presionaba más o menos sobre la hoja de estaño produciendo así protuberancias y huecos debido a la diferente presión ejercida por las vibraciones del aire provocadas por la voz, según Gendre [1990].

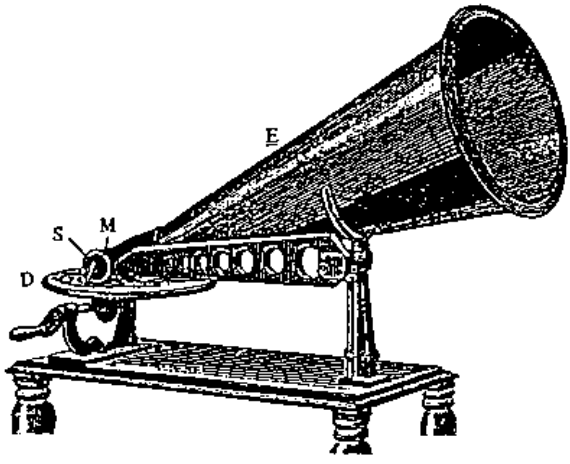

Figura 6. Gramófono de Berliner (1888). Comienzos del registro sonoro

Para escuchar lo grabado, se colocaba el estilete de nuevo al comienzo del cilindro, y la superficie rugosa provocaba variaciones de presión en la caja que eran amplificadas por la bocina. Poco después los cilindros se sustituyeron por discos planos, lo que dio lugar al gramófono (Figura 6).

W.H. Preece y A. Stroh en 1879 examinaron bajo microscopio las estrías producidas por el gramófono para intentar descubrir la naturaleza física de los sonidos. Con esas observaciones no conseguían ningún resultado de provecho y decidieron seguir la aproximación inversa. Construyeron un sintetizador mecánico que generaba un tono complejo gracias a la suma de un tono puro y un número variable de armónicos. Lo construyeron con un conjunto de ruedas engranadas que rotaban a diferentes velocidades. La huella generada por el sintetizador se comparaba con las huellas que producía el gramófono. Esta idea de usar la síntesis como ayuda al análisis de la señal, ha demostrado ser una importante aproximación al problema, según Poulton [1986].

C. Paget en 1923 descubrió que había dos componentes frecuenciales en todos los sonidos vocálicos e hizo una tabla con ellas, por observación de su propia voz. En la actualidad se acepta que hay otras componentes o *formantes* además de los observados por Paget. Construyó un sistema formado por un conjunto de resonadores acústicos en forma parecida a los tubos de Kratzenstein. Estos resonadores estaban hechos de arcilla y goma e individualmente podían producir todos los sonidos vocálicos y consonánticos.

Por esta época comenzó el desarrollo de dispositivos eléctricos para realizar el proceso de síntesis. El primer dispositivo completamente eléctrico fue desarrollado por J.Q. Stewart en 1922 (Figura 7).

Este sintetizador consistía en un interruptor para simular las cuerdas vocales y una serie de circuitos para simular las resonancias de las cuerdas vocales. Otro interruptor cortaba o permitía el paso de la corriente, de la misma manera que lo hacen las cuerdas vocales al provocar una pulsación en la corriente de aire. Stewart, al igual que Paget, trabajo en la teoría de la existencia de los dos formantes y para ello diseñó dos circuitos resonantes separados, compuestos de un número variable de bobinas y condensadores. Un dispositivo de resistores variables controlaban las amortiguaciones y las intensidades relativas de los dos resonadores. Este sintetizador producía unos resultados bastante aproximados a los reales para un cierto número de sonidos.

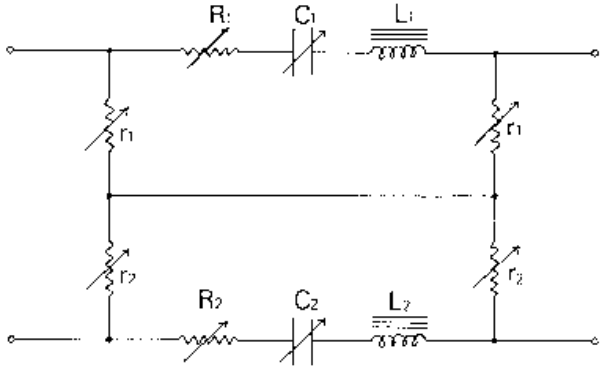

Figura 7. Sintetizador eléctrico de Stewart (1922).

El *Voder* fue uno de los primeros sintetizadores totalmente eléctricos de discurso continuo controlado desde un teclado por un experto. El origen del nombre viene de las palabras inglesas *demostrador de voz (VOice DEmostratoR*). Este sistema fue exhibido en la *feria mundial* de New York, 1939 (Figura 8) y en San Francisco en 1940. Para poder manejar este equipo era necesario bastante entrenamiento, del orden de un año. El operador manipulaba 14 *llaves* con los dedos, para controlar la estructura resonante del tracto vocal y un pedal de pie derecho que permitía lograr un tono variable.

Una vez que se conocía su funcionamiento, se podía producir discurso continuo inteligible con facilidad. El Voder tenía dos fuentes de sonido: un oscilador que generaba un zumbido periódico, análogo al interruptor de Stewart para los sonidos sonoros, y un ruido aleatorio para los sonidos sordos. La sección de resonadores era más complicada que la de Stewart y contenía 10 filtros pasa-banda que abarcaban todo el rango de frecuencias de la señal. El control de ganancia de cada uno de los 10 filtros podía ser ajustado individualmente mediante llaves. Los sonidos *oclusivos* se podían producir por tres llaves extras que generaban pulsos transitorios.

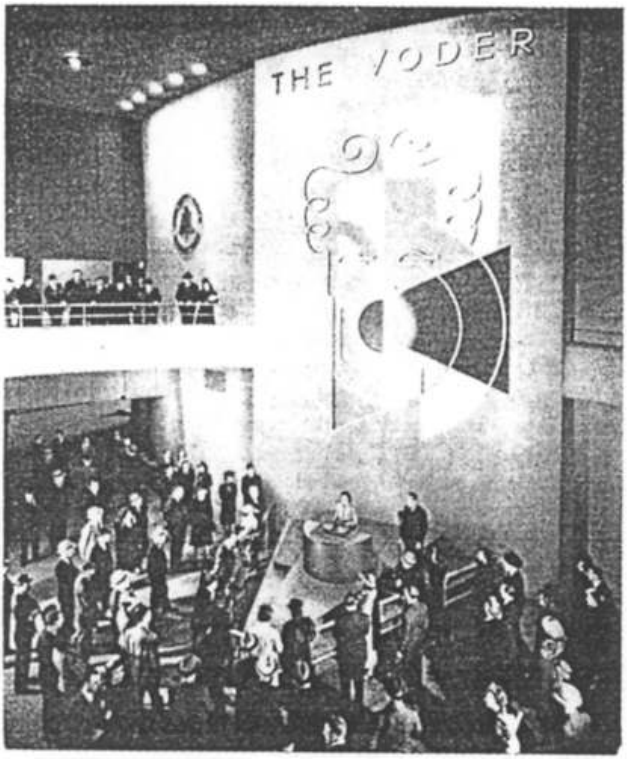

Figura 8. Demostración del Voder en la feria mundial de Nueva York, en 1939

Contemporaneamente con el Voder apareció el *Vocoder*,. Se trata de un compresor de banda ancha para la telefonía. Fue proyectado en 1939 y uno de los modelos más recientes corresponde a los *Laboratorios Siemens* de Munich. Este instrumento parte de un análisis del habla real, por ello, más que una creación a partir de cero es una reconstrucción de algo ya dado.

La señal de voz se codifica de forma que pueda ser transmitida eficientemente y descodificada posteriormente al otro lado de la línea telefónica. Un banco de filtros separa las diferentes bandas de frecuencia de la señal. Para poder caracterizar a la señal acústica original se calculan una serie de parámetros. Se detecta si hay sonido sordo o sonoro (las cuerdas vocales sólo vibran para los sonidos sonoros, por ejemplo para las vocales). Se mide también el tono fundamental de vibración de las cuerdas vocales, la amplitud de la señal en cada banda, etc. Estos valores son multiplexados y transmitidos a través de la línea telefónica. El aparato receptor realiza la operación inversa mediante un demultiplexor y se reconstruye la señal. Es mejor transmitir los parámetros en vez de señal completa, ya que éstos varían más lentamente y sólo es necesario transmitirlos cada 20 ms. Pero el equipo necesario era caro para la época y por eso el Vocoder sólo se usó para aplicaciones muy especializadas.

El Vocoder fue importante ya que su sección de recepción es análoga a la que poseen muchos sintetizadores modernos y la sección de transmisión es similar a la etapa de análisis espectral de la mayoría de los reconocedores actuales.

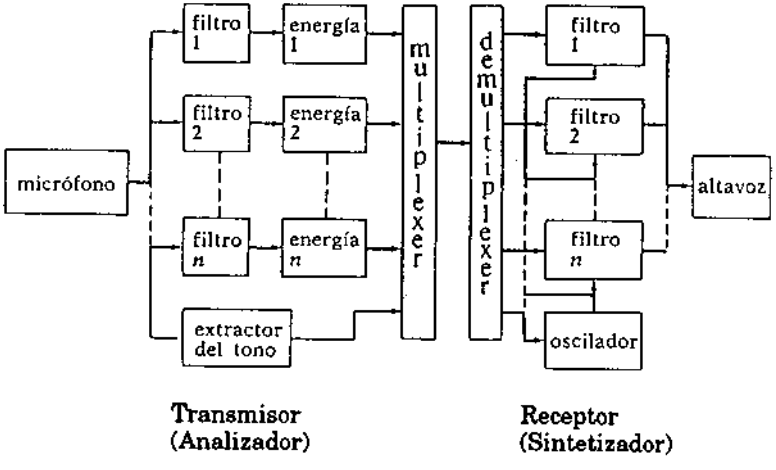

Figura 9. Diagrama del Vocoder

2.2 . Síntesis a partir del espectrograma

Desde que el matemático S. B. J, Fourier, nacido el 21 de marzo de 1768 en Auxerre y muerto el 16 de mayo de 1830 en París, diera a conocer su famoso teorema en 1822, según el cual toda onda compleja se descompone en elementos sinusoidales simples, se sabía que podía llevarse a cabo el análisis de cualquier sonido, aunque con grandes complicaciones y empleando bastante tiempo en ello.

A principios del siglo XX, ya se había solucionado la complejidad de la descomposición por medios instrumentales. B. Malmberg da noticia de las primeras descomposiciones llevadas a cabo por científicos alemanes de la casa *Siemens* antes de la primera guerra mundial, según Martínez [1986]. Este *analizador* o *espectrómetro* ya separaba los distintos armónicos de la onda compleja; produciendo resultados como los que aparecen en la Figura 10.

Durante la segunda guerra mundial fueron los Estados Unidos los que avanzaron en el perfeccionamiento del espectrógrafo; su nombre comercial era *Sona-graph* y fue desarrolla-

do por los *Laboratorios Bell* en los años cuarenta La primera descripción del aparato y las primeras muestras de los *espectrogramas* o *sonogramas* aparecen en 1947, en el libro de R. K. Potter, G. A. Kopp y H. G. Green, titulado *Visible Speech*. Este sugestivo título daba a entender la primera intención con la que había sido creado: volver visible el habla para que los sordos pudieran leerla.

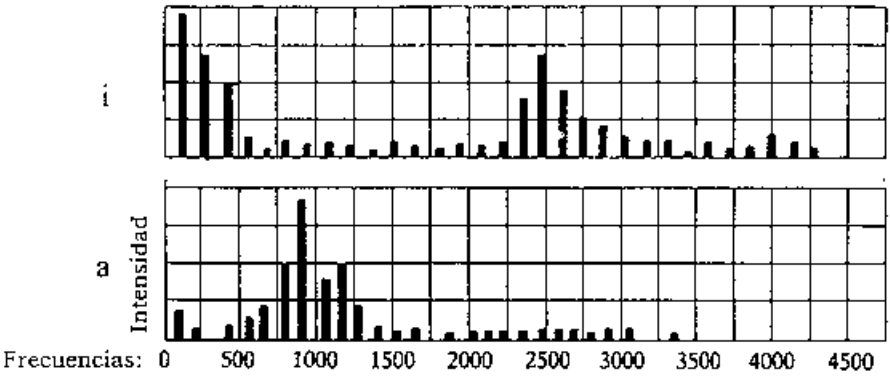

Figura 10. Dos espectros de vocales

El espectrógrafo de sonidos produce un dibujo de la distribución de energía en el dominio del tiempo y de la frecuencia llamado espectrograma. Pero un dibujo de tales características requiere tres dimensiones y el papel sólo tiene dos por lo que la energía se muestra en el grado de ennegrecimiento sobre el papel. El tiempo se representa en el eje horizontal y la frecuencia en el eje vertical. Por tanto en periodos de silencio no se dibuja nada sobre el papel mientras que los sonidos de mediana intensidad aparecen con tonos de grises (Figura 11).

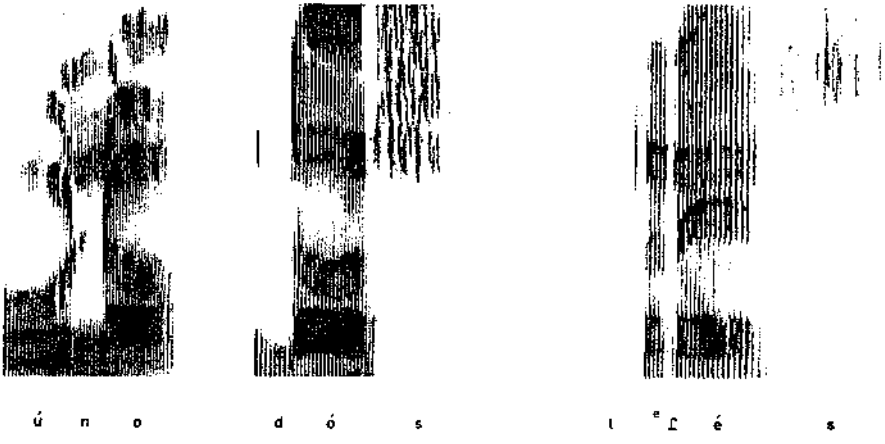

Figura 11. Espectrograma obtenido para las pronunciaciones de varios dígitos castellanos.

Inmediatamente después, a lo largo de los años cincuenta, los fonéticos y fonólogos comenzaron a estudiar gradualmente la producción del aparato fonador humano y a sistematizar los resultados a partir de los cuales se fundamentaron las teorías fonológicas, tal y como hizo R. Jakobson, con sus colaboradores G. Fant y M. Halle.

En este contexto surge una nueva aproximación a la síntesis de voz hecha por H.K.Dunn en 1950. Este dispositivo eléctrico (Figura 12) modelaba el tracto vocal y se lograba una gran mejora de los resultados con respecto a los proporcionados por el Voder.

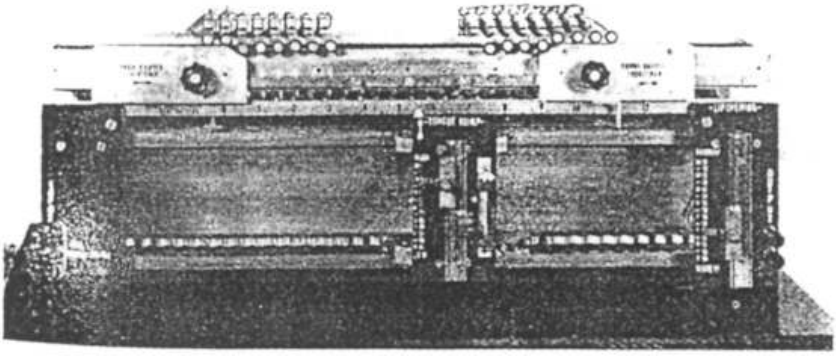

Figura 12. Sistema construido por H.K. Dunn en 1950 para modelar el tracto vocal

El sistema que construyó estaba basado en una fuente de energía eléctrica que simulaba las cuerdas vocales y un modelo de lineas de transmisión (una escala de bobinas y condensadores) que representaban al tracto vocal. Usaba filtros pasa-baja que proporcionaban el retardo experimentado por la onda sonora a través de la cavidad bucal.

Dunn obtuvo medidas del tracto vocal a través de fotografías obtenidas mediante rayos X, y calculó las frecuencias de resonancia aproximando la forma del tracto vocal a cilindros. Los resultados que obtuvo coincidían con las medidas experimentales del espectrograma de la señal, al menos en los tres primeros formantes.

Con este sintetizador el tono del sonido sólo podía cambiarse mediante ajustes manuales. Sin embargo, G. Rosen introdujo en 1958 el primer sintetizador *controlable dinámicamente*. Estaba basado en la misma idea que el de Dunn, usando condensadores y bobinas. Los condensadores eran circuitos de válvulas que usaban el *principio de Miller* para variar su capacidad efectiva. El sintetizador era controlado por un dispositivo con retardo que seleccionaba una secuencia de configuraciones del tracto vocal. Las transiciones entre una configuración y otra eran suavizadas electrónicamente.

Puesto que el espectograma de una señal acústica es adecuado y contempla todas las características acústicas importantes de la voz, es obvio que siguiendo sus pautas se debía poder reconstruir de nuevo dicha voz. Los sintetizadores basados en el estudio de formantes también podían ser controlados dinámicamente. El primer sintetizador que contenía tres filtros de formantes variables y uno fijo conectado en serie fue descrito por J. Anthony y W.

Lawrence en 1962. Este sintetizador era controlado por un dispositivo electromecánico que leía el esquema de formantes dibujado como un grafo con tinta conductora.

De igual forma los *Laboratorios Haskins* de Nueva York llevaron a cabo la síntesis del lenguaje de una manera completa a partir del espectrograma en el llamado *reproductor de patrones* (*Pattern Playback*) de Hanskins (Figura 13).

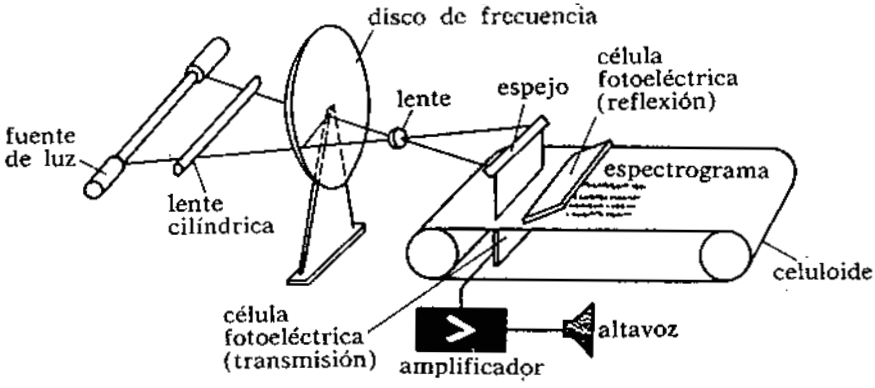

Figura 13. Esquema del Pattern Playback

El procedimiento que se emplea en el *Playback* consiste en dibujar sobre una banda de celuloide transparente un espectrograma inspirado en uno real o inventado. Puesto que el conocimiento del análisis espectrográfico permite conocer los elementos y partes principales de cada sonido, dibujándolas se podía conseguir que el aparato las leyera y pronunciase (Figura 14).

La banda de celuloide pasa a una velocidad conveniente delante de un sistema de células fotoeléctricas y de vibradores, que en cada momento mezcla las diferentes componentes con las amplitudes proporcionales al dibujo de los espectros instantáneos que aparecen sobre la banda en ese lugar. Y reconstruye las fluctuaciones de los objetos más o menos esquematizados a lo largo del tiempo, los amplifica y los pronuncia por un altavoz o sobre un magnetófono, según Martínez [1986]

Otro sintetizador del habla dinámicamente controlable que parte del espectograma, pero que no necesita dibujarlo es el *pronunciador de parámetros artificiales* (*Parametric Artificial Talker* (PAT))(Figura 15). Es un análogo acústico del aparato fonador humano, en el que unos circuitos eléctricos resuenan cuando un estímulo similar a la vibración laríngea los pone en movimiento, de manera parecida a lo que ocurre en las cavidades de resonancia del aparato fonador humano.

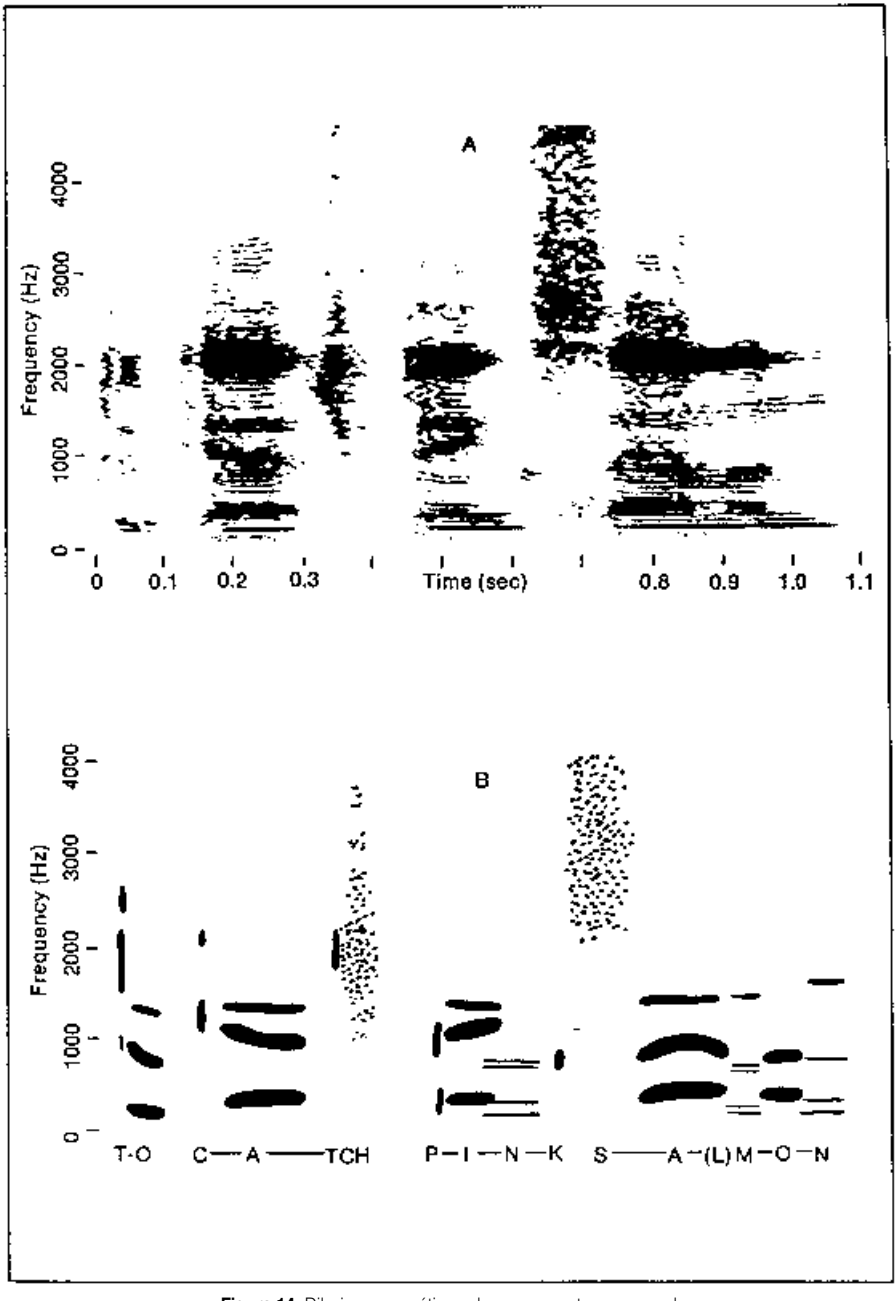

Figura 14. Dibujo esquemático sobre un espectrograma real

Para imitar acústicamente un sonido son necesarios dos fuentes de energía: una de sonido periódico, generador de la vibración glotálica en imitación de la laringe y otra de sonido aperiódico o ruido.

Estas dos fuentes se controlaban a través de 8 parámetros variables, que dan lugar a los diferentes sonidos consonánticos: amplitud laríngea, frecuencia laríngea o tono, primer formante de los sonidos con características vocálicas, segundo formante, tercer formante, frecuencia del ruido, amplitud del ruido y ruido que se manifiesta a través de los formantes.

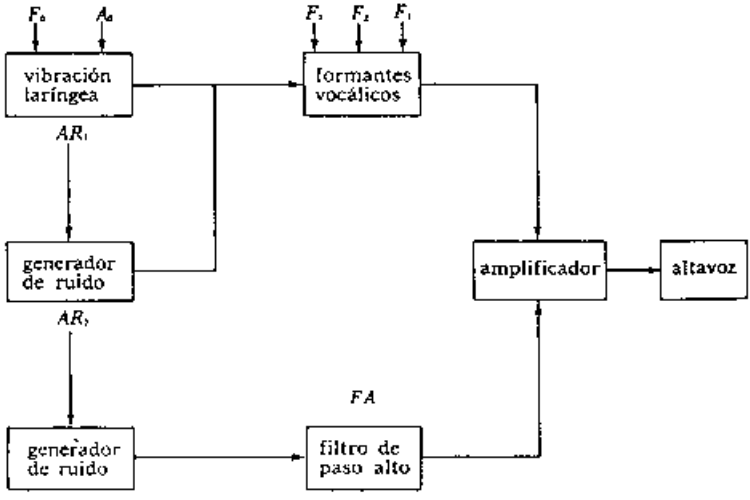

Figura 15. Esquema del PAT desarrollado en Edimburgo

La información necesaria para cada uno de estos ocho parámetros se conseguía a través del espectrograma de dos maneras principalmente. La primera consiste en restituir con una tinta conductora los parámetros, en un papel especialmente calibrado sobre el mismo espectrograma. La segunda se realiza utilizando manualmente los mandos y proporcionando aquellos datos que resulte interesante comprobar.

El *OVE II* es un aparato similar, puesto a punto por G. Fant y otros en 1956 en los *laboratorios de transmisión del habla* de Estocolmo. El posterior OVE III, que también usaba la información proporcionada por los espectrogramas era controlado por un computador digital CDC 1700. El computador se encargaba de variar los parámetros a través de convertidores *digital/analógicos*. Se guardó en la memoria del computador una *librería* que contenía todos los parámetros para las vocales y consonantes suecas.

Debido al grado de efectividad que se logró con los computadores digitales, se hizo evidente que debían llevar a cabo el proceso de síntesis completo y no sólo tareas de control.

Por esta época hubo también importantes avances en programación como el algoritmo de la *transformada rápida de Fourier*, según Brigham [1974] que permite hacer el espectrograma mucho más rápido. Para sintetizar una única palabra es necesario una gran cantidad de cálculos. Si se pretende que el sistema funcione en tiempo real, es decir al ritmo normal del lenguaje hablado, hay un tiempo limitado para hacer las operaciones. Hay dos soluciones, o bien hacer el computador más rápido o hacer programas más eficientes. Ambas aproximaciones han sido ampliamente estudiadas.

Actualmente en síntesis se emplean exhaustivamente las técnicas digitales y hay muchos sistemas que consisten en uno o más circuitos integrados que contienen cientos de transistores.

Como resumen podemos decir que la síntesis está bastante bien establecida si lo medimos en términos de la cantidad de productos que hay en el mercado. Los sintetizadores están disponibles como accesorios en las más populares marcas de computadores, y también como productos en sí mismos. El sonido de los primeros sintetizadores era muy artificial, pero la calidad ha aumentado notablemente en los últimos años. La mayoría de los sintetizadores actuales tratan de copiar los patrones de entonación de la voz humana, por lo que la voz ya no es monótona. Los vocabularios varían desde 24 palabras para los más simples hasta los que tienen cientos de palabras o más. Los sintetizadores fonéticos que generan palabras a partir de una cadena de fonemas (la unidad más básica) tienen en principio un vocabulario ilimitado. No obstante, este método implica una excesiva simplificación de la teoría fonética y el sonido que se produce no es enteramente natural en algunas ocasiones debido a las transiciones entre unos fonemas y otros.

3. ANÁLISIS DE VOZ

El análisis o reconocimiento de una señal de voz es una tarea que presenta mayor dificultad que el proceso de síntesis explicado anteriormente. Las principales características y dificultades que presenta una señal de voz a ser analizada son su continuidad (no existen pausas entre sílabas, palabras, etc.), la redundancia natural al expresarnos normalmente y la gran variabilidad entre distintos locutores. Las razones para esa variabilidad son: anatómicas (longitud del tracto vocal, forma de la cavidad nasal, sexo, edad, etc.), dialectales, hábitos de pronunciación, personalidad del locutor, estilo. Pero esta variabilidad no sólo se observa entre pronunciaciones de distintos locutores, sino que para un locutor resulta materialmente imposible pronunciar una misma palabra dos veces igual. Y es que un mismo locutor puede hablar en un tono bajo o alto, susurrando o gritando, relajado o tenso.

El desarrollo de muchos aspectos de los sistemas de reconocimiento automático del habla son total o parcialmente dependiente de la lengua y de la aplicación. En el estado actual de la tecnología en este campo, no existe una solución general al problema. Sin embargo, existen reconocedores de palabras aisladas y de discurso continuo para pequeños vocabularios y tareas muy concretas que permiten abordar el desarrollo de numerosos productos de interés: control de *robots* industriales o de electrodomésticos mediante voz, dictado automático, ayuda a discapacitados (por ejemplo, a ciegos), acceso y navegación por bases de datos, operaciones comerciales (venta de billetes en una terminal de aeropuerto), sistemas de seguridad (controles de entrada mediante voz), operaciones telefónicas automáticas, etc. Los sistemas *multimedia*, que permiten la transmisión y el tratamiento simultáneo de voz, imagen y datos en tiempo real, presentan un amplio abanico de posibilidades en nuestro entorno industrial: trabajo en casa, disminución de costes de transporte y almacenamiento de datos, tele-reuniones, acceso rápido e integrado a bloques selectivos de información, etc.

Al contrario que en el desarrollo de síntesis en la que primero hubo un gran desarrollo de sistemas mecánicos y eléctricos, en análisis los principales avances se han logrado con el desarrollo de los computadores en los años setenta. No obstante, se construyeron algunos sistemas eléctricos a principio del siglo XX.

3.1. Primeros dispositivos

Si un sistema de síntesis puede ser considerado como una boca artificial, un sistema de análisis o reconocimiento puede ser considerado como un oído artificial. En definitiva se trata de utilizar la voz para comunicarse con las máquinas en vez de usar teclados, mandos, etc. El sistema debe aceptar como entrada las señales acústicas y actuar en consecuencia. En un primer nivel la respuesta podría ser la simple transcripción de la señal acústica, para pasar en niveles superiores a un sistema que entienda el mensaje y realice tareas concretas en respuesta a frases pronunciadas de forma natural.

J.B. Flowers en 1916 fue quizá el primero en diseñar una máquina para transcribir voz. Tenía conocimiento sobre cómo se realizaba la transmisión de mensajes por cable entre submarinos. Éstos eran transmitidos en un código alfabético especial y se registraban como una linea ondulada sobre papel para ser posteriormente descifrados. Con un poco de experiencia esta señal registrada podía ser leída como escritura ordinaria. Este conocimiento permitió a Flowers proponer una máquina que podía convertir los sonidos de las letras en ondas de la misma naturaleza que las que se transmitían entre los submarinos, a las que llamó *alfabeto fonográfico (phonographic alpahabet)*. Aunque la moderna teoría de formantes aún no se había desarrollado, se dio cuenta de la existencia de tonos a diferentes frecuencias en la señal acústica (100, 200 y 1000 Hz para el caso de un hombre pronunciando la palabra inglesa *go*). La máquina propuesta por Flowers es interesante ya que en su forma básica no usaba ningún componente electrónico. Consistía de dos electroimanes y condensadores ajustados para cubrir todo el rango de frecuencias. El dispositivo puede verse en la Figura 16.

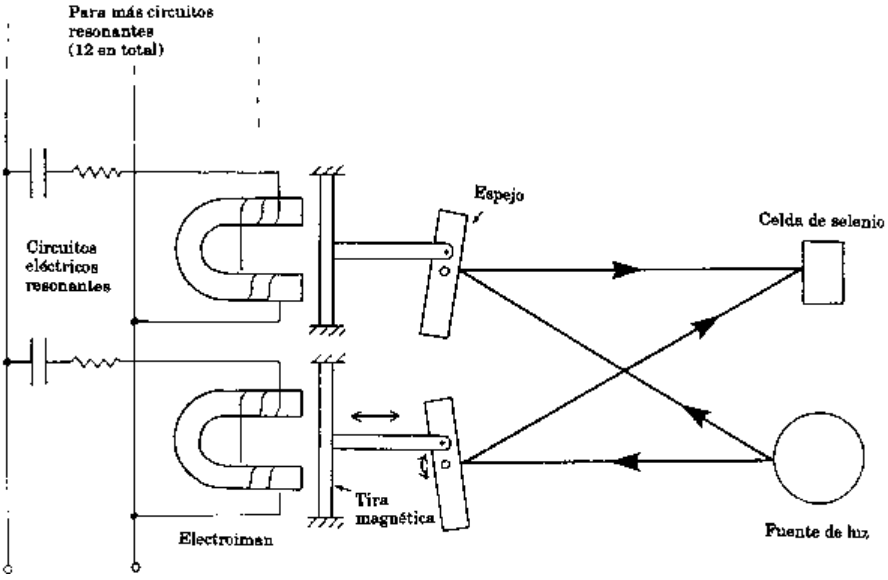

Figura 16. Diseño de la primera máquina que permitía una rudimentaria transcripción a partir de voz en 1916.

Como ya se ha puesto de manifiesto anteriormente el reconocimiento de voz es mas difícil *a priori* que la síntesis, hasta el punto de que en 1930 las autoridades de patentes alemanas desestimaron otorgar una patente a T. Nemes por su *transcriptor fonético*. En su científica opinión un transcriptor fonético a partir de voz era imposible *por principio*. Lo novedoso de la propuesta de Nemes era que su dispositivo era óptico. Aunque probó que su técnica era admisible, se le previno de que no siguiera desarrollando la idea a causa de la guerra, según Poulton [1983].

Pero lo que se puede llamar en sí la historia del reconocimiento del habla es mucho más reciente y comienza en los años cuarenta cuando se desarrolló un dispositivo capaz de visualizar la señal acústica sobre papel y que naturalmente fue el *espectrógrafo*. Como ya se ha mencionado anteriormente se trata de un dispositivo que permite la obtención de un registro de la energía contenida en las diversas bandas de frecuencia de una palabra o frase en función del tiempo, llamado *espectrograma*. A partir de este momento comenzó a vislumbrarse la posibilidad de realizar sistemas para el reconocimiento automático del habla. Es en 1952 cuando K.H. Davis, R. Biddulph, S. Balashek, de los *Laboratorios Bell* construyen el primer dispositivo de reconocimiento, capaz de discriminar con cierta precisión los diez dígitos ingleses pronunciados de forma aislada por un único locutor (Figura 17).

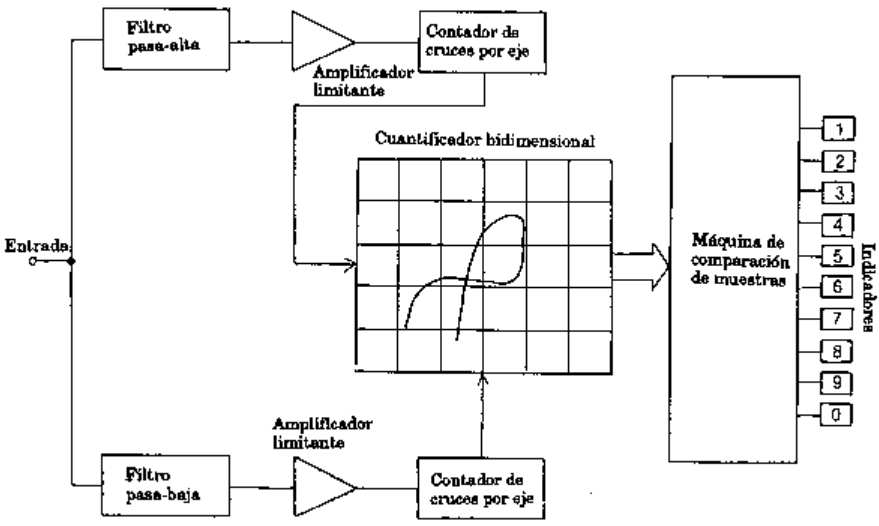

Figura 17. Reconocedor de dígitos desarrollado por Davis, Biddulph y Balashek en 1952

El dispositivo era totalmente electrónico y para la comparación hacía uso de técnicas de intercorrelación entre los parámetros (posiciones de los dos primeros formantes) del dígito pronunciado y los patrones correspondientes a cada una de las diez posibles pronunciaciones esperadas, según Davis [1952].

La señal de voz era separada por dos filtros dentro de dos bandas de frecuencia. Un filtro *pasa-alta* permitía el paso de las componentes frecuenciales por encima de los 900Hz y un fil-

tro *pasa-baja* permitía el paso de las que eran inferiores. La secuencia principal en cada banda se localizaba a partir de la cuenta del número de *cruces por cero*, que simplemente determina la frecuencia con la que la señal cambia su polaridad. En resumen, esto quiere decir que la señal completa se rompía en dos señales que representaban la frecuencia de los dos primeros formantes y estas variaban de forma relativamente suave, durante el curso de la pronunciación. Entonces estas señales eran aplicadas a los canales X e Y de un osciloscopio.

Cuando se pronunciaba un dígito a través del micrófono, se dibujaba una curva de dos dimensiones en la pantalla del osciloscopio. Estas curvas eran bastante diferentes para cada uno de los 10 dígitos ingleses. Para comparar las curvas automáticamente era necesario cuantificarlas de alguna manera. Esto se hizo dividiendo el dibujo de dos dimensiones dentro de cuadros y midiendo el trozo de curva que ocupaba cada cuadro. Había 28 cuadros para representar las diferentes combinaciones de los formantes uno y dos. Estos cuadros estaban definidos por un circuito de válvulas que activaba uno de los 28 relés en función de la forma del dibujo. Cada relé correspondía a uno de los cuadros en los que se había dividido la pantalla del osciloscopio. En función de los relés activados se encendía uno de los indicadores correspondiente a un número reconocido.

Las medidas de los patrones se hicieron con 100 repeticiones de cada uno de los 10 dígitos y se obtenían tasas de reconocimientos comprendidos entre un 97% y 99%.

Otro trabajo paralelo e independiente se desarrolló en los *Laboratorios RCA* en 1956 según Rabiner [1993]. H.F. Olson y H. Bellar desarrollaron la *máquina de escribir fonética*. La señal obtenida del micrófono pasaba a través de un amplificador y un compresor antes de ser aplicada a un banco de 8 filtros. El propósito de la compresión era ajustar el valor medio de la señal, para que fuera lo más parecida entre locutores que hablaban tanto en tono alto como bajo. Los filtros se ajustaban para calcular la envolvente. Esta información se pasaba a un banco de relés mediante un interruptor que actuaba cada 40 ms desde el momento en que la pronunciación comenzaba. Se obtenía un dibujo característico para los distintos sonidos, dado por las diferentes configuraciones de relés.

Olson y Bellar se preguntaron en un primer momento que unidades léxicas sería mejor reconocer: fonemas, sílabas o palabras. En un principio probaron con los tres conjuntos, pero las transcripciones obtenidas con los fonemas eran poco ajustadas a la realidad, con sílabas mejoraban y con palabras completas la transcripción era perfecta, si la palabra era reconocida correctamente. Pero para vocabularios razonables había 36 palabras, 1.000 sílabas y 10.000 palabras. Por esta razón si se usaban las palabras, la cantidad de *memoria* necesaria para guardar sus características debía ser muy grande (problema que sigue siendo fundamental en nuestros dias). Ésto les llevó a desarrollar un sistema de reconocimiento basado en sílabas. Los datos proporcionados por los relés (memoria espectral) eran descodificados en sílabas y después en letras individuales para actuar sobre las teclas de una máquina de escribir. El sistema fue probado con frases constituidas por 10 sílabas en varias permutaciones diferentes. Si la entonación de la frase era cuidada, resultaba un porcentaje de acierto de un 98%. Continuaron sus experimentos y llegaron a analizar palabras aisladas traduciéndolas a diferentes idiomas. El sistema reconocía palabras inglesas y francesas y las podía traducir a ingles, francés, alemán y castellano.

Otros trabajos de la época también basados en dispositivos analógicos, como el de J. Wiren y H.L. Stubbs en 1956 obtenían información sobre el contenido espectral de las seña-

les y usaban como criterio de clasificación la frecuencia de resonancia de las vocales. Usaban un *árbol binario* de selección, de modo que en cada hoja se hacía una separación. Por ejemplo, si en una hoja del árbol se separaban los fonemas sordos de los sonoros, en la siguiente inferior en la que todas las unidades eran sordas, se separaban en *fricativas* o no, etc. Los circuitos necesarios para hacer estas separaciones incluían una gran cantidad de válvulas, porque el transistor era un elemento demasiado nuevo para esa época.

3.2. Era informática

Los computadores digitales causaron un fuerte impacto en los sistemas de reconocimiento, al igual que en los de síntesis. En 1960 P. Deves y M. V. Mathews construyeron un reconocedor de dígitos basándose en un computador IBM 704. Introdujeron el concepto de *normalización temporal no lineal (Dynamic Time Warping)* que permite comparar los parámetros de palabras iguales pronunciadas a distinta velocidad. Esta normalización no era posible con las técnicas analógicas anteriores e implicó un fuerte aumento en las tasas de reconocimiento. En una prueba realizada con cinco locutores masculinos, el sistema dio un 6% de error cuando se realizaba la normalización y un 12% sin ella. El esquema básico de estos primeros sistemas de reconocimiento, según Peinado [1994], se puede ver en la Figura 18. A partir de estas fechas se realizaron una gran cantidad de trabajos, principalmente de reconocimiento de palabras aisladas y monolocutor.

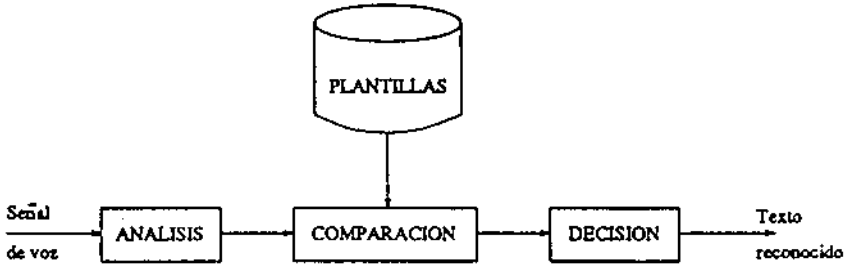

Figura 18. Esquema básico de un sistema de reconocimiento

Además de los *Laboratorios RCA y Laboratorios Bell* también comenzaron su andadura la *corporación eléctrica nipona (Nippon Electric Corporation (NEC))* y la *Universidad Carnegie Mellow* que continúan hasta nuestros dias. Para el año 1970 ya se habían logrado una gran cantidad de progresos en los niveles básicos del reconocimiento automático del habla. Por ejemplo, D.R. Reddy en 1967 describió un sistema que consistía en dos ordenadores interconectados (un IBM 7090 y un PDP1). Este sistema identificaba los fonemas vocálicos y consonánticos. Pero el principal problema era cómo aplicar los conocimientos de niveles lingüísticos superiores: fonéticos, sintácticos, semánticos y pragmáticos para ayudar a la comprensión del mensaje pronunciado.

G. Fant en 1970 sugirió un modelo en el cual la tarea de reconocimiento se dividía en una secuencia de 5 pasos: extracción de parámetros, detección de segmentos, transcripción fonética, identificación de palabras e interpretación semántica. Por el contrario el sistema HEARSY (de las palabras inglesas HEAR=oir y SAY=decir) propuesto por R. Reddy estaba ba-

sado en un modelo de reconocimiento en paralelo que incluía tres reconocedores por separado acústico, sintáctico y semántico. El HEARSY I es una continuación del anterior y también contenía los tres reconocedores independientes. La frase era preprocesada para extraer sus características y el resultado se pasaba a los tres módulos. Cada reconocedor hacia sus propias hipótesis, basándose en sus diferentes niveles de conocimiento y al final un sistema *mediador* que tenía toda la información controlaba el sistema y daba los resultados.

Hubo una extrapolación optimista por parte de investigadores y organismos financieros de llegar en poco tiempo a sistemas capaces de reconocer de forma precisa frases casi-cualesquiera, pronunciadas por un locutor cualquiera de forma continua. Con este objetivo más o menos en mente, se lanzaron grandes proyectos de investigación en los que se pretendía llegar a las menores restricciones gramaticales posibles de las frases a reconocer, así como del léxico utilizado. Son varios los países en los que se comenzó a trabajar en proyectos de esta índole (Japón, Francia, etc.), pero es en Estados Unidos donde se lanza en 1971, el mayor proyecto conocido en la reciente historia del reconocimiento automático del habla. Se trata del *Advanced Research Projects Agency -Departamento de Defensa- Speech Understanding Research (ARPA-SUR)* con un presupuesto de quince millones de dólares y una duración de 5 años.

Un grupo de expertos dirigidos por el profesor A. Newell propone una larga lista de especificaciones. El sistema debía aceptar discurso continuo, permitiendo una fácil adaptación a nuevos locutores, el diccionario de trabajo debía ser de 10.000 palabras con una sintaxis artificial adecuada a una tarea restringida. El número de errores en la comprensión de las frases debía ser menor del 10%, y el tiempo de respuesta de pocas veces tiempo real, en una máquina de la generación siguiente a la de la época en que fueron propuestos estos objetivos. El sistema debía manejar 100 millones de instrucciones por segundo y ser estar listo para su demostración en 1973.

Antes de que hubieran acabado completamente sus deliberaciones, hicieron algunas modificaciones. Por ejemplo, se paso a exigir un tamaño de vocabulario de 1.000 palabras en vez de 10.000 y se amplio el plazo de presentación de los resultados hasta 1976.

Aunque los objetivos no llegaron a alcanzarse plenamente, las aportaciones del proyecto *ARPA-SUR* contribuyeron de forma muy notable al mejor conocimiento de las propiedades del habla y de las limitaciones de los sistemas de reconocimiento automático del habla, así como a la toma de conciencia de la necesidad de una mayor investigación en este campo para salvar estas limitaciones.

Varías empresas y universidades participaron en el proyecto con mayor o menor fortuna. Fundamentalmente se perfilaban dos aproximaciones al problema. Los *modelos estructurales estocásticos* y en concreto los *modelos ocultos de Markov* según Rabiner [1989] y Torres [1993]. Con este tipo de aproximación la correlación entre las características espectrales propias de la señal acústica y las correspondientes unidades sub-léxicas es aprendida de forma automática a partir de un amplio conjunto de muestras de aprendizaje. Y los *sistemas basados en el conocimiento* en los que todo el conocimiento aportado al sistema se basa en *reglas* sintácticas, gramaticales, etc., dadas *a priori* u obtenidas tras un estudio del problema concreto.

El sistema que más se aproximó a los objetivos propuesto fue *HARPY* de la *Universidad Carnegie-Mellon* y desarrollado por D. H. Klatt. El sistema *HARPY* modelaba todas las fuentes de conocimiento (fonológica, léxica, sintáctica y semántica) en una única gran red de estados

finitos, la cual se obtenía previamente a partir de sub-redes que modelaban las distintas palabras. Aunque su antecesor, el sistema DRAGON, utilizó un modelo Markoviano y los métodos asociados, el sistema *HARPY* no se basaba explícitamente en *cadenas de Markov*, sino que asumía una estructura en red *particularizada* al problema específico, e introducía una importante modificación al *algoritmo de Viterbi* de reconocimiento: *la búsqueda en haz (Beam Search).* Este sistema alcanzó una tasa de errores semánticos de aproximadamente el 5%, con un vocabulario de 1011 palabras, una sintaxis artificial de baja *perplejidad*, con *adaptación* al locutor, y con tiempos de respuesta de aproximadamente 80 veces tiempo real.

La Universidad Carnegie-Mellon desarrolló, también en el marco del ARPA-SUR, otro sistema totalmente diferente; el *EARSAY-II*. continuación del *EARSAY-I*. Este sistema introdujo la arquitectura *pizarra (blackboard)* y el concepto de fuente de conocimiento, como procesos paralelos, independientes, cooperativos y asíncronos. Estos procesos eran controlados por un *planificador*, y se comunican entre sí a través de una estructura de datos global (la pizarra) donde se *anotaban* todas las hipótesis emitidas. Las fuentes de conocimiento se *autoactivaban* ante la satisfacción de ciertas condiciones de estado de la pizarra, especificadas en sus *tramas-estímulo (stimulus frame)*, lo que permitía realizar las acciones especificadas en sus correspondientes *tramas respuestas (response frame).* Las tramas-respuesta (programas), se dedicaban a introducir o actualizar ciertas informaciones de la pizarra, sobre las que tenían competencia. Este sistema consiguió tasas de error del 9% al nivel puramente semántico, y del 26% al nivel sintáctico-semántico, con un consumo de recursos de 3 a 4 veces superior que el HARPY. No obstante, los méritos de HEARSY-II no se deben valorar en base a estos resultados, sino que su mayor logro ha sido las aportaciones a la arquitectura de sistemas complejos *inteligentes*, llamados normalmente *sistemas basados en el conocimiento*.

También se desarrolló dentro del proyecto ARPA-SUR, el sistema cuyo nombre viene de la frase *oir lo que quiero decir* (*Hear What I Mean* (*HWIN*)) de la *Bolt Beranek and Newman Inc*. Fue el tercero en cuanto a prestaciones alcanzadas. HWIN estaba basado en una filosofía (no arquitectura) similar al HEARSY-II, pero haciendo especial hincapié en conseguir un óptimo en las estrategias de hipotetización y de los resultados de interpretación a obtener. Este sistema constaba de cuatro módulos básicos: el *procesador acústico*, el *módulo de acceso léxico*, el *componente lingüístico* y el *módulo de control*. El procesador acústico estaba en contacto directo con el módulo de acceso léxico, el cual se encargaba de suministrar, a petición del modulo de control, hipótesis sobre cuales eran las mejores palabras contenidas en un segmento dado de señal, junto con una medida de la calidad de la comparación. El componente lingüístico se encargaba de determinar, también a petición del modulo de control, si una secuencia de palabras dada podía ser interpretada como una sub-secuencia de alguna sentencia sintáctica, semántica o pragmática correcta. El objetivo principal del módulo de control era descubrir cual es la secuencia óptima de palabras que cubría la totalidad de la señal vocal, y podía ser aceptada a la vez por el componente lingüístico. Los resultados alcanzados por este sistema a la finalización del proyecto ARPA-SUR, fueron bastante peores que los alcanzados por los sistemas anteriores (56% de error semántico con un consumo de recursos aproximadamente 200 veces mayor que el HARPY)

En paralelo con el proyecto ARPA-SUR, varias empresas desarrollaron sus propios sistemas, de entre los cuales el mas significativo es el del grupo de tratamiento del habla del *Centro de investigación IBM J. Thomas (Thomas, J. Watson Research Center*) según Jelinek [1976] y Bahl [1983]. Básicamente proponen un sistema de reconocimiento de discurso con-

tinuo que consta de un *procesador acústico*, seguido por un *descodificado lingüístico*. El procesador acústico era el encargado de transcribir la señal acústica a una cadena de símbolos fonéticos y el descodificado lingüístico traducía dicha cadena fonética, posiblemente con errores, a una cadena de palabras. El reconocimiento del habla se formulaba como un problema de la *teoría de la comunicación*: el locutor y el procesador acústico eran considerados como si fueran un *canal de transmisión*, donde el locutor transforma el texto fuente en señal de voz, y el procesador acústico actúa como un compresor de datos. El canal proporcionaba una cadena ruidosa, a partir de la cual el descodificado lingüístico debía recuperar el mensaje; en este caso, el texto original. El descodificado lingüístico constaba de un *Modelo de Markov* del lenguaje, compuesto de los sub-modelos correspondientes a palabras y fonemas, y de un *algoritmo de descodificación*, cuyo objetivo era encontrar la cadena de palabras que con mayor probabilidad podía ser producida por el generador de Markov, y que fuera además compatible con la cadena observada. Como algoritmo de descodificación se utilizaba el *algoritmo de Viterbi* para redes de talla reducida, y el llamado *algoritmo a pila (stack decoding)* para grandes redes. Los resultados obtenidos con este sistema para tareas concretas se acercaban e incluso superaban a los de HARPY para tareas semejantes, con la ventaja de que gran parte del trabajo de construcción de las fuentes de conocimiento estaba totalmente automatizado.

Aparte de los Estados Unidos, varios son los países en los que se han desarrollado proyectos de reconocimiento automático del habla dignos de mención. Uno de los más constructivos fue el propuesto conjuntamente por investigadores de la *Universidad de Concordia* (Canadá) y de la *Politécnica de Torino* (Italia), consistente en una interesante extensión de la idea de *sistema experto* que en cierto modo actualizaba las ideas introducidas en HEARSY-II. Esta extensión se basaba en el concepto de *sociedad de expertos*, según la cual se constituían como (micro-)sistemas expertos las actividades cooperativas de diversos niveles de comprensión. De esta manera los diferentes expertos en sociedad podían ejecutar en paralelo los distintos algoritmos que resultaban como consecuencia de una descomposición del problema general de interpretación en tareas. En su última versión, la sociedad de expertos se componía de 3 expertos (acústico, fonético-pseudosilábico y léxico). La estrategia de control de los dos primeros expertos utilizaba métodos de *planificación* para establecer *planes* de actuación de los micro-expertos que componían cada experto, de forma que el trabajo conjunto podía realizarse de forma más eficiente. Este paradigma, se estaba aplicando en un sistema de reconocimiento que pretendía realizar la difícil tarea de reconocer con precisión secuencias de dígitos y letras inglesas conectadas sin ninguna restricción sintáctica. Según los resultados preliminares era un sistema muy prometedor.

En Francia, el *Centro de investigación en informática de Nancy* (*Centre de Recherche en Informatique de Nancy (CRIN)*) lanzó en la época de los 70 el proyecto *MYRTILLE* cuyo primer sistema (*MYRTILLE-I*) permitía el uso de lenguajes artificiales, modelados por gramáticas independientes de contexto, para soportar pequeñas aplicaciones. El segundo de los sistemas *MYRTILLE-II*, concebido con unos planteamientos mucho más ambiciosos, permitía una sintaxis *pseudo-natural* para una aplicación concreta de información meteorológica. MARTILLE-II se componía de tres niveles: en primer lugar el descodificado acústico-fonético; en segundo lugar el correspondiente al reconocimiento de frases que utilizaba redes como representación del lenguaje, y uno léxico jerarquizados en 4 niveles. El último nivel era el de interpretación, que aplicaba restricciones semánticas y pragmáticas locales. Posteriormente

el proyecto MYRTILLE desembocó en un replanteamiento del problemas de la *maquina de dictado automática*, utilizando la metodología de sistemas expertos para la descodificación acústico fonética.

Demostrada en los primeros 80 la ineficacia de los sistemas basados en conocimiento, se invirtió todo el esfuerzo en desarrollar sistemas capaces de extraer conocimiento de forma inductiva, es decir, a partir de muestras. Se utiliza a partir de entonces, siguiendo los trabajos de IBM, una modelización acústica basada en *Modelos Ocultos de Markov* (*Hidden Markov Models*), discretos y continuos, y se optiman los algoritmos de aprendizaje para entrenar los sistemas a partir de grandes bases de datos. Se mejoraron también los sistemas de alineamiento temporal no lineal para el reconocimiento de palabras conectadas, más concretamente se desarrollaron algoritmos de búsqueda eficientes con los que determinar la secuencia óptima de patrones para una secuencia de vectores acústicos.

Mediada la década de los 80 se presentó la aproximación conexionista como alternativa a la aproximación estadístico-probabilística. Las *redes neuronales artificiales (Artificial Neural Networks*) compartían con los modelos ocultos de Markov su carácter inductivo, es decir, el aprendizaje a partir de muestras, pero sus configuraciones clásicas -como los *perceptrones multicapa (Multi-Layer Perceptron)*- no eran capaces de representar fenómenos dinámicos como la señal de voz, por lo cual tuvieron que desarrollarse arquitecturas recursivas específicas, con objeto de superar estas limitaciones. Otros autores han optado desde entonces por configuraciones híbridas en las que los perceptrones multicapa se utilizan para estimar las probabilidades de emisión de un modelo oculto de Markov.

Actualmente la investigación se concentra, por un lado, en mejorar el rendimiento de la modelización acústica (generación automática de unidades acústicas contextuales, entrenamiento discriminativo de los modelos) y, por otro lado, en la integración de niveles de conocimiento superiores (estrategias de búsqueda heurísticas en grandes autómatas que representan modelos del lenguaje). También se está invirtiendo un gran esfuerzo en diseñar y adquirir grandes bases de datos para el entrenamiento de sistemas de reconocimiento de discurso continuo.

4. CONCLUSIONES

La síntesis de voz tiene una larga historia. El aparato bucal humano es un sistema acústico maravillosamente controlado y se han hecho muchos intentos de reproducirlo artificialmente. Los primeros aparatos mecánicos fueron sustituidos por métodos eléctricos y los sintetizadores de voz actuales están basados en circuitos integrados y se puede encontrar una gran variedad en el mercado. Aunque aún se están haciendo mejoras en los dispositivos se puede decir que ya es posible encontrar sintetizadores que producen voz muy natural.

El reconocimiento automático del habla por el contrario, ha experimentado un gran avance en los últimos años. De no poder reconocer más que un conjunto pequeño de palabras aisladas (típicamente dígitos y unos pocos comandos o palabras clave) pronunciadas por un único locutor, se ha pasado a sistemas de vocabularios medios y grandes, capaces de reconocer palabras conectadas, o de identificar palabras clave en discurso continuo, con bastante independencia del locutor. Hoy día existen máquinas de dictado automático, aunque no demasiado sofisticadas; los sistemas de seguridad mediante voz, de reconocimiento o verifi-

cación del locutor, están en plena expansión comercial, y algunos de los ordenadores personales *multimedia* aparecidos recientemente, capaces de combinar voz, imagen y datos en una pantalla, son también capaces de procesar órdenes habladas.

Pero quedan muchas tareas sin resolver. Actualmente, están en proceso de desarrollo sistemas de discurso continuo, con lenguajes restringidos, para el acceso automático a bases de datos, lo cual puede permitir aplicaciones como la venta automática de billetes en un aeropuerto, o la navegación interactiva por *Internet* mediante voz. Las investigaciones apuntan a sistemas de discurso continuo, multilocutor, con grandes vocabularios, tal vez con gramáticas restrictivas, que poco a poco irán acercándose más al lenguaje natural. Tareas como el diálogo, que mejoraría el rendimiento de los sistemas interactivos, y la traducción automática, de particular interés para la Comunidad Económica Europea, también están en fase de investigación.

5. BIBLIOGRAFÍA

AINSWORTH, W.A. (1973) *Mechanism of Spech Recognition*. Oxford, Pergamon Press.

BAHL, L.R., JELINK, F., MERCER, R.L. (1983) "A Maximum Likelihood Approach to Continuous Speech Recognition". *IEEE Transactions on Pattern Analysis and Machine Inteligence*, 5(2), 179-190.

BRIGHAM, E.O. (1974) *The fast Fourier transform*. Englewood Cliffs, N.J. Prentice-Hall.

CASACUBERTA. F, VIDAL. E. (1987) *Reconocimiento Automático del Habla*. Madrid, Marcombo.

DAVIS, K. H., BIDDULPH, R., and BALASHEK, S. (1952) "Automatic recognition of spoken digits". *Journal of the Acoustical Society of America*, 24(6), 637-642.

FLANAGAN, J.L. (1972) *Speech Analysis, Synthesis, and Perception*. Segunda edición, New York, Springer-Verlag.

GENDRE. C. (1990) *Magnetófonos*. Madrid, Paraninfo.

KLATT, D. (1977) "Review of the ARPA speech understanding project". *Journal of the Acoustical Society of America*, 62, 1324-1366.

JELINEK, F. (1976) "Continuous Speech Recognition by Statistical Methods". Proc. of the IEEE, 64(4), 532-566.

MARTINEZ. E (1986) *Fonética*. Barcelona, Teide.

PEINADO. A.M. (1994) *Selección y estimación de parámetros en sistemas de reconociminto de voz basados en modelos ocultos de Markov*. Granada. Tesis doctoral.

POULTON, A.S. (1983) *Microcomputer Speech Synthesis and Recognition*.. London, Sigma Thecnical Press.

RABINER, L.R. (1989) "A Tutorial on Hidden Markov Models and Selectec Aplications in Speech Recognition". Procc of the IEEE, 77(2), 257-286.

RABINER, L.R., JUANG, B.H. (1993) *Fundamental of speech recognition*. New Jersey, PTR Prentice-Hall.

TORRES, I., CASACUBERTA, F., VARONA, A. (1993) "Acoustic-Phonetic Decoding of Spanish Continous Speech with Hidden Markov Models". *NATO-ASI, News Advanced and Trenck in Speech Recognition and Coding,* 43-46.