## 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.

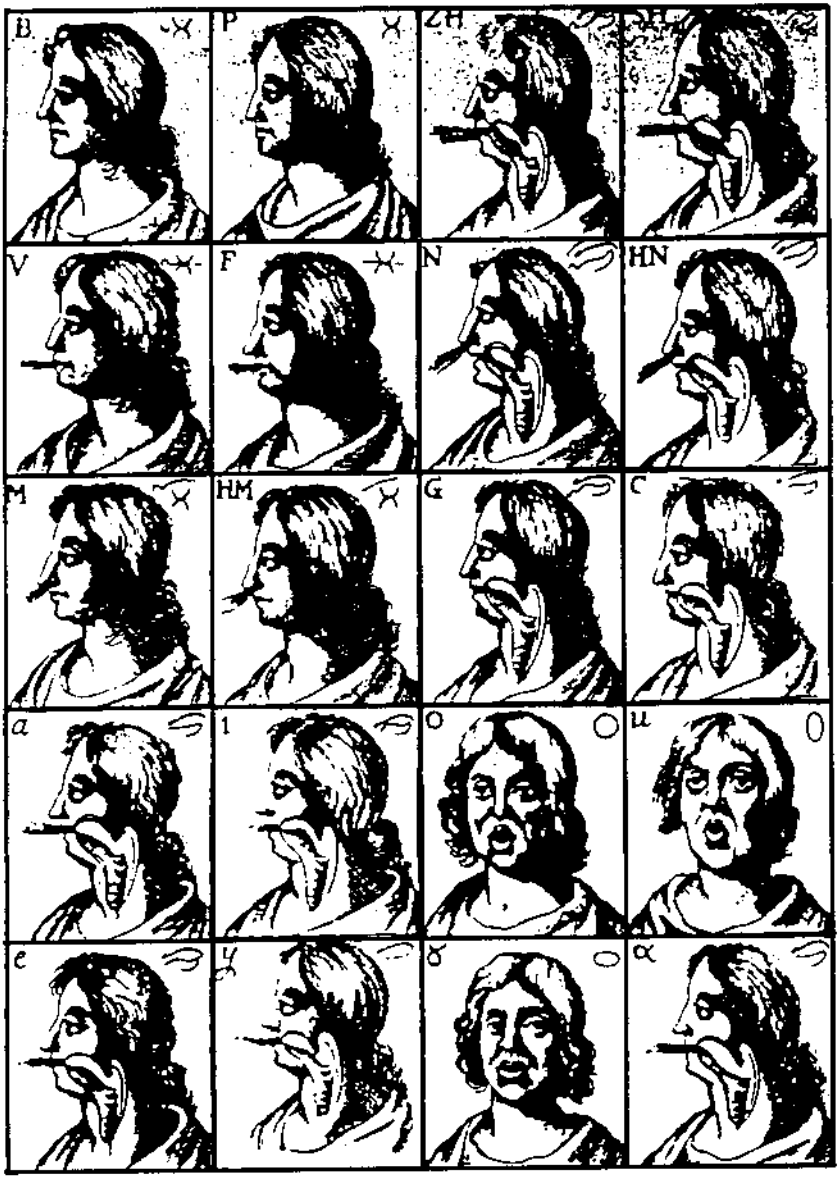


**Figura 1.** Diferentes posiciones del tracto vocal publicadas por B.J. Wilkins en 1668

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.

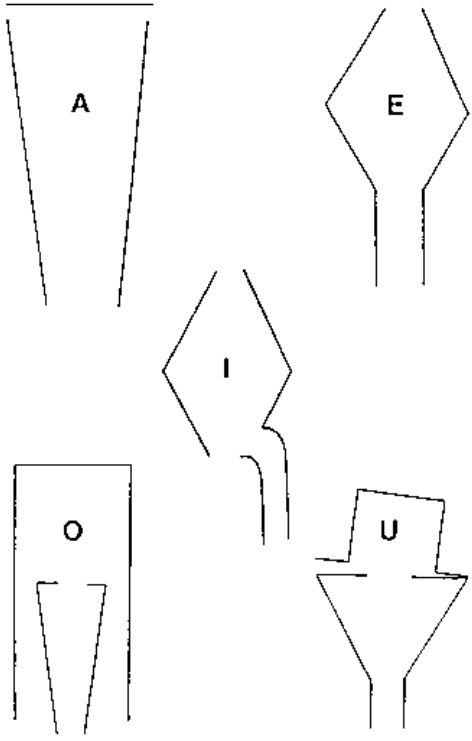


**Figura 2 .** Tubos construidos por C.G. Kratzenstein en 1779 para reproducir los sonidos vocálicos

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].

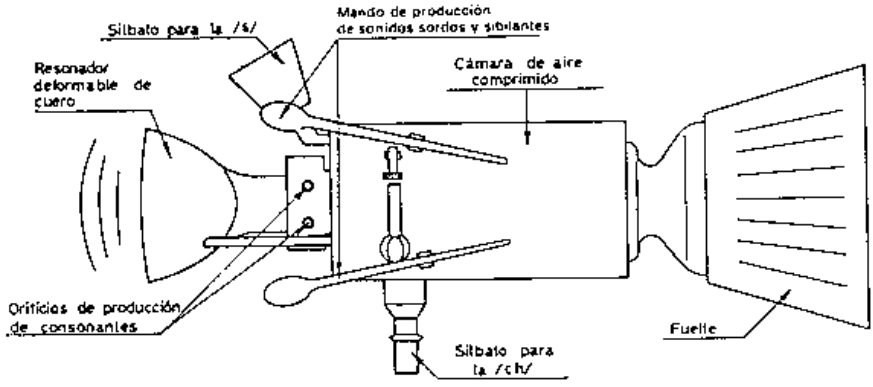


**Figura 3.** "Máquina parlante" de W. Kempelen (siglo XVIII)

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).

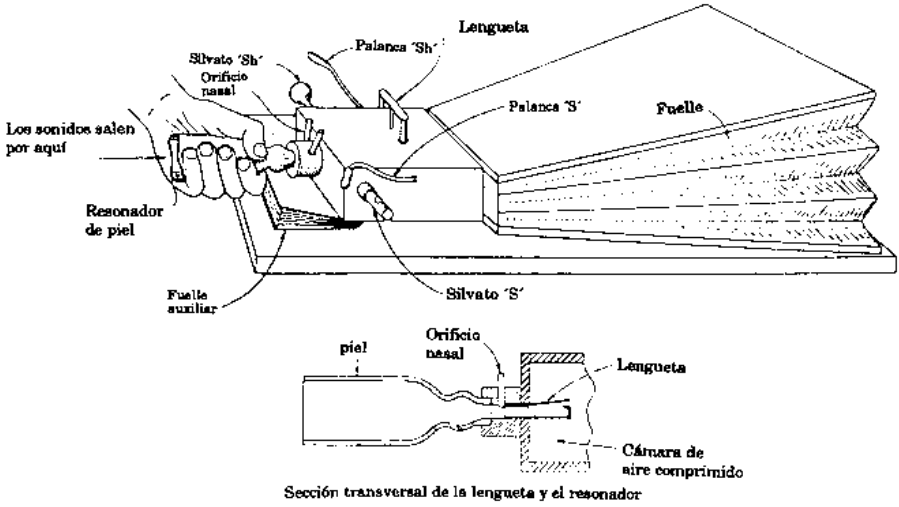


**Figura 4.** Máquina parlante de G. Wheatstone presentada en 1835

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.

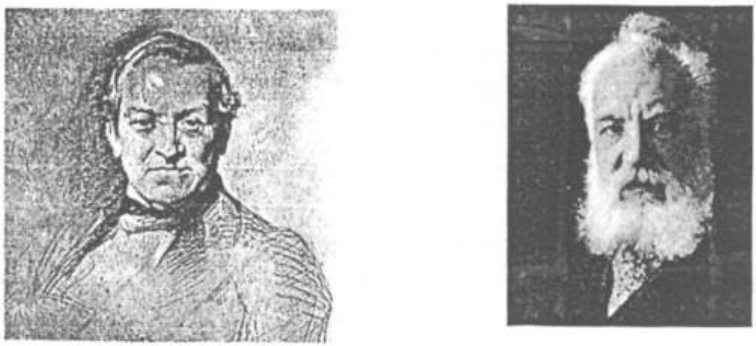

**Figura 5.** Retratos de G. Wheastone a la izquierda y A.G. Bell a la derecha

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].

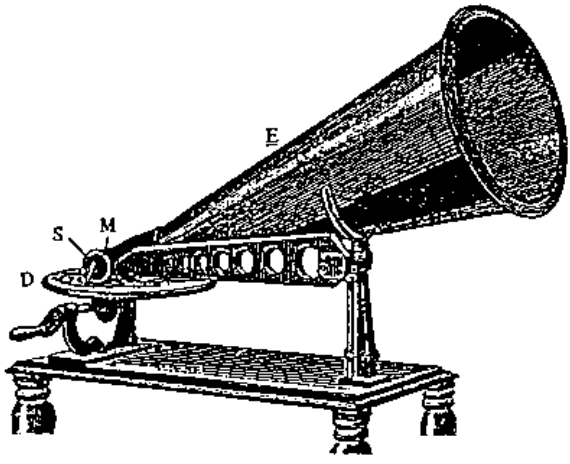


**Figura 6.** Gramófono de Berliner (1888). Comienzos del registro sonoro

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.

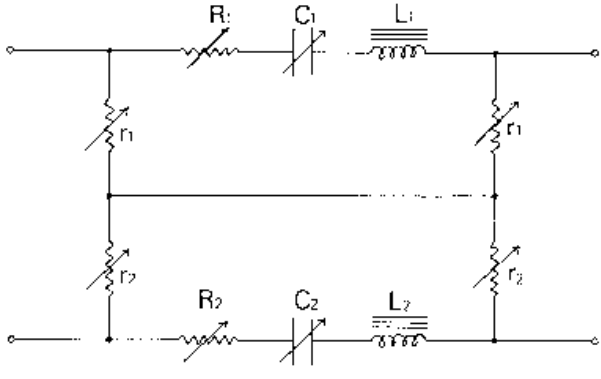


**Figura 7.** Sintetizador eléctrico de Stewart (1922).

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.

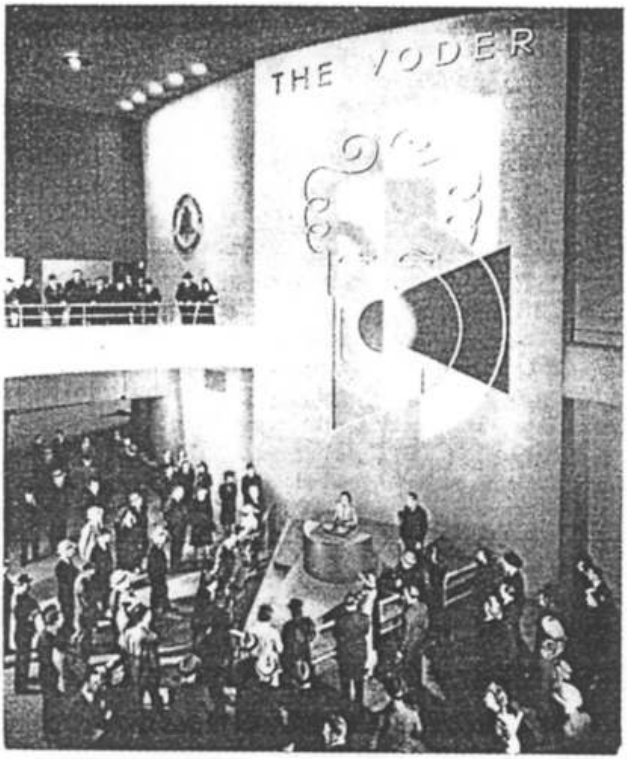


**Figura 8.** Demostración del Voder en la feria mundial de Nueva York, en 1939

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.

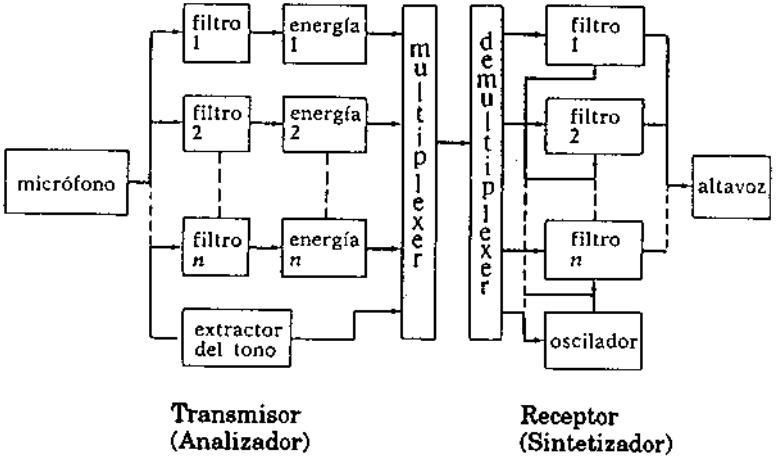


**Figura 9.** Diagrama del Vocoder

## 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.

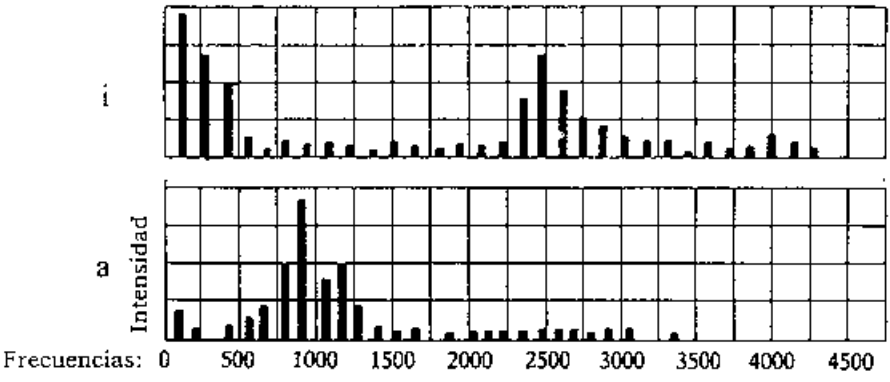


**Figura 10.** Dos espectros de vocales

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).

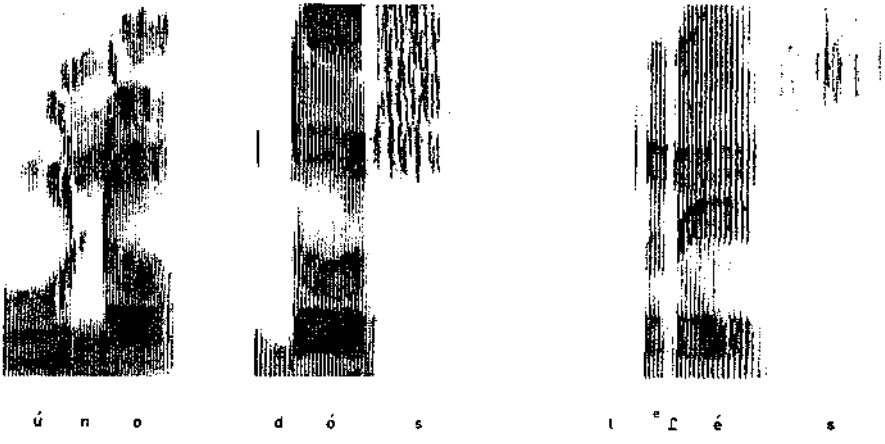


**Figura 11.** Espectrograma obtenido para las pronunciaciones de varios dígitos castellanos.

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.

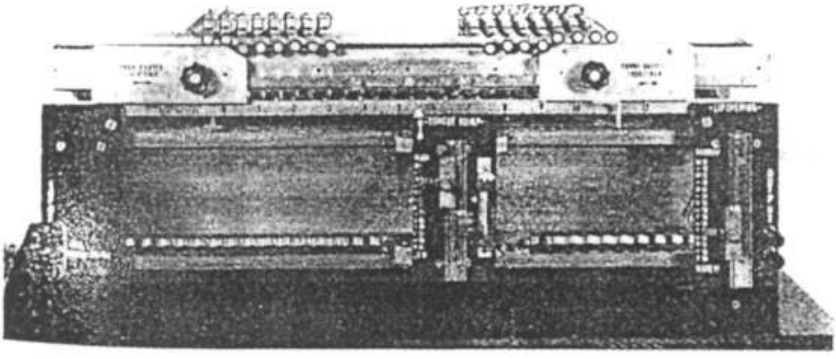

**Figura 12.** Sistema construido por H.K. Dunn en 1950 para modelar el tracto vocal

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).

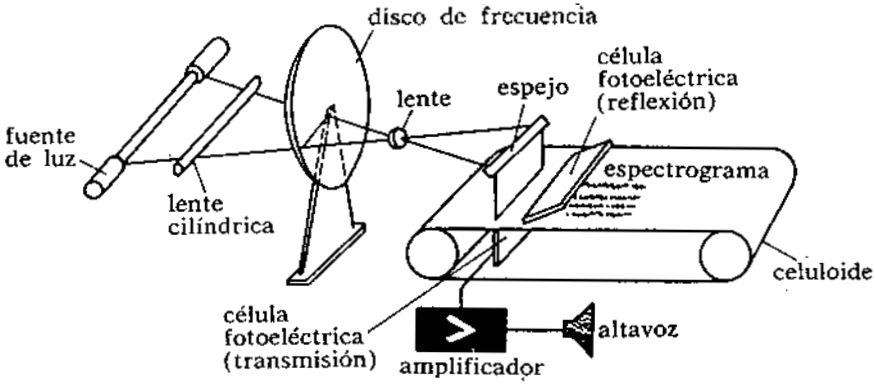


**Figura 13.** Esquema del Pattern Playback

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.

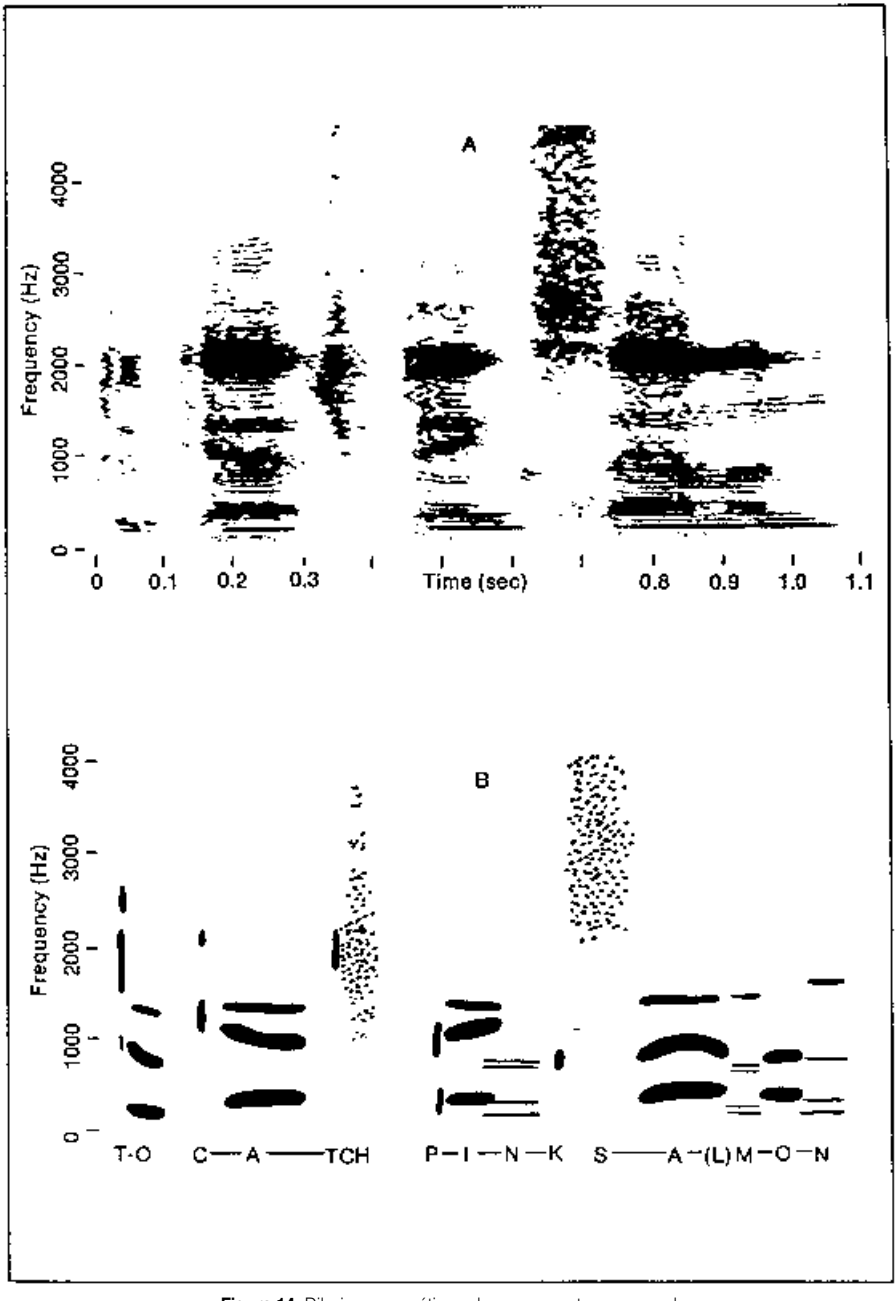


**Figura 14.** Dibujo esquemático sobre un espectrograma real

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.

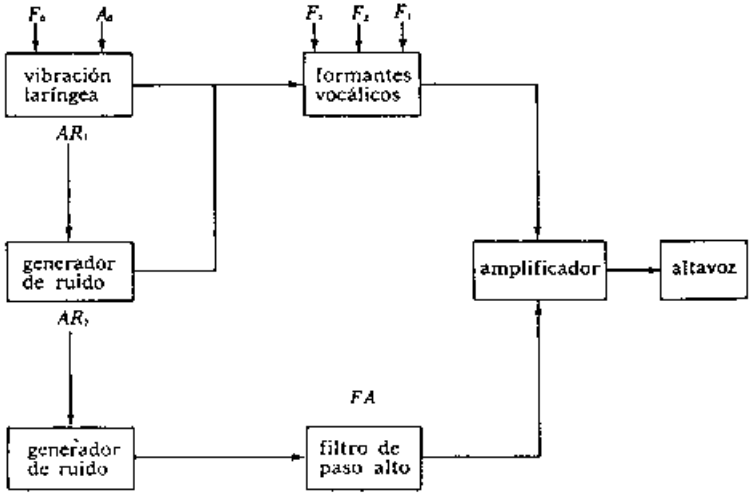


**Figura 15.** Esquema del PAT desarrollado en Edimburgo

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.