SITUACION ACTUAL DE LA SINTESIS DE VOZ

JOSE MARTI ROCA

Jefe del Departamento de Acústica de la
Escuela Universitaria de Telecomunicación
La Salle Bonanova, Barcelona.

1. ETAPAS HISTORICAS PREVIAS.

El tema de la producción de la voz humana desde muy antiguo se ha visto rodeado de un cierto misterio y de una especial curiosidad que se ha ido desvelando a medida que los conocimientos acústicos han ofrecido una explicación científica del fenomeno.

Durante la década de los cuarenta se descubrió y se implantó el espectrógrafo, instrumento que contribuyó notablemente al análisis de la voz y, por la tanto, también a los métodos de síntesis. La representación tridimensional de la energía en función de la frecuencia y el tiempo permite una visión muy clara de la evolución de la voz a lo largo del tiempo. El sistema sirvió no solamente para el análisis del habla, sino también para obtener las primeras experiencias de síntesis de voz a partir de una lectura óptica de los espectrogramas por el sistema denominado "Pattern Playback" (Cooper, 1951).

Los grandes progresos en el tratamiento del habla, tanto en el análisis como en la síntesis, se han realizado a partir del momento en que se ha podido tratar la señal debidamente digitalizada con métodos numéricos y con la ayuda del ordenador. El sistema más simple de almacenamiento digital consiste en muestrear la señal a una frecuencia, pongamos de 10 KHz (para asegurar una banda pasan-te de 5 KHz) con un mínimo de 10 bits. Esto representa un ritmo de almacenamiento de 100000 bits/s. Ahora bien, mucha de esta información es redundante, ya que la voz presenta muchos intervalos estacionarios y con una parametrización adecuada se puede reducir esta tasa hasta unos 2000 bits/s. Esta parametrización permite una edición posterior de la voz en tiempo real, con una gran versatilidad para la concatenación de unidades básicas.

2. SISTEMAS BASICOS DE PARAMETRIZACION.

Tradicionalmente se ha considerado que la unidad más elemental de la voz humana sería el fonema, como abstracción de las unidades acústicas elementales de una determinada lengua. Ahora bien, durante la producción de un fonema, la onda acústica está sometida a variaciones, tanto en la configuración como en la amplitud, de forma que, si la queremos describir de forma directa, habrá que acudir a intervalos de tiempo más cortos durante los cuales las características del sonido se puedan considerar prácticamente constantes. Estas unidades pueden recibir el nombre de microfonemas o tramas, y deberán ser suficientemente cortas para que en transiciones rápidas este escalonamiento no traiga problemas para la resolución temporal de nuestra percepción acústica. Cada microfonema quedará determinado por unos valores numéricos o parámetros que caracterizarán los valores de la señal acústica para generarla adecuadamente cuando interese. Veremos a continuación los principales sistemas de parametrización que se utilizan actualmente.

2.1. Vocoders de parámetros articulatorios.

Un modelo acústico que explica la función de filtrado que tiene el tracto bucal consiste en suponer una serie de conductos cilíndricos de diferentes secciones conectados desde las cuerdas vocales hasta los labios y, en su caso, también con una derivación lateral por las cavidades nasales. Cada uno de estos cilindros, atravesados por ondas planas, se puede tratar análogamente al circuito de la figura 1; donde la inductancia M corresponde a la masa acústica de la cavidad y la capacidad C a la elasticidad o "compliance" de la misma cavidad. R corresponde a la resistencia que ocasiona las pérdidas por viscosidad del fluido y G a la conductancia que da lugar a las pérdidas de energía acústica por las paredes del recinto. Las presiones de entrada p1 y salida p2 se corresponden análogamente a unas diferencias de potencial y las velocidades volumétricas U1 y U2 a unas corrientes eléctricas.

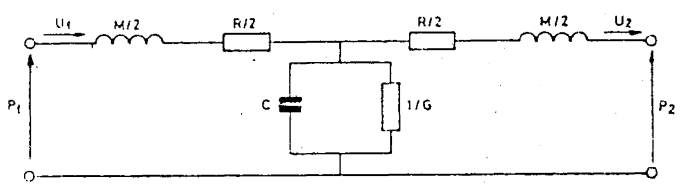


Fig. 1. Modelo eléctrico equivalente de un conducto cilíndrico.

Este método de trabajo nos permite deducir el comportamiento acústico del tracto bucal a partir

de su propia geometría que se cuantifica mediante la denominada función de área, o valor de la sección del conducto vocal en función de la distancia. C.H. Coker (Coker, 1968) propuso una interpretación del tracto bucal (fig.2) con un número más reducido de parámetros a partir de las coordenadas de posición de la lengua, la mandíbula, los labios y el velo del paladar. Conociendo la evolución a lo largo del tiempo de estas coordenadas y con un programa de cálculo adecuado se deduce la función de área que permite calcular las tres primeras resonancias de la cavidad bucal e introducirlas en un sintetizador. Actualmente, estos modelos articulatorios se trabajan por simulación mediante filtros digitales variables en el tiempo.

2.2. Vocoders de canales paralelos.

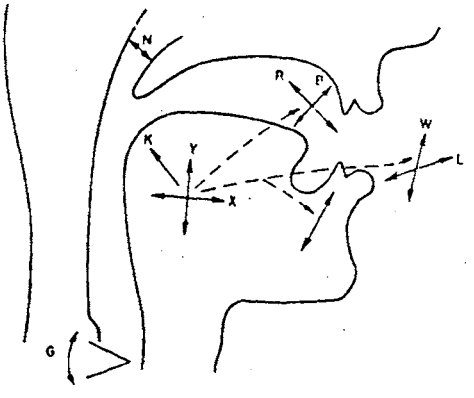


Fig.2. Modelo articulatorio propuesto por Coker.

El primer vocoder fue diseñado por H.Dudley (1939), en un intento de reducir la banda pasante de las comunicaciones telefónicas a la décima parte, sin afectar notablemente a la inteligibilidad. Así creó un modelo de análisis-síntesis del habla que utilizaba un banco de once filtros en paralelo. La señal de cada filtro reducida a una banda de 25 Hz, era enviada junto con las demás como una señal única dentro de una banda de 11 x 25=275 Hz, que no llega a la décima parte de la banda telefónica utilizada en aquel tiempo (3000 Hz). Para realizar la síntesis, la señal correspondiente a cada canal servía para controlar otro banco de filtros yuxtapuestos, cada uno con una banda pasante de 300 Hz y excitados por un oscilador de relajación o un ruido blanco. El sistema dio lugar a un tipo de palabra inteligible, pero de calidad mediocre.

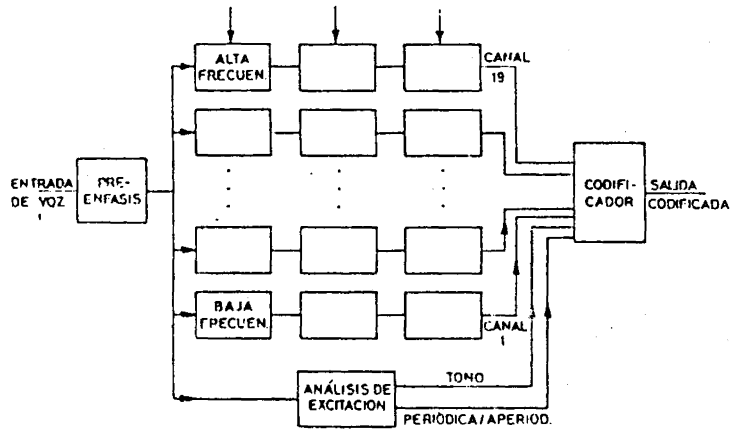


Fig. 3. Diagrama de bloques del vocoder de Holmes. Parte correspondiente a la codificación.

Posteriormente, se han utilizado vocoders con diferentes variantes y un progresivo perfecciona-

miento de la calidad. Una de las realizaciones más conocidas es la de J.N.Holmes (Holmes, 1980) que utiliza un banco de filtros (Fig.3).

2.3. Sintetizador por formantes.

Los sintetizadores por formantes parametrizan los valores de las resonancias del tracto bucal (formantes), la fuente de excitación periódica (con tono) o aperiódica (ruido) y el nivel energético. Estos parámetros se actualizan en tiempo real para una generación de voz continua (Fig.4).

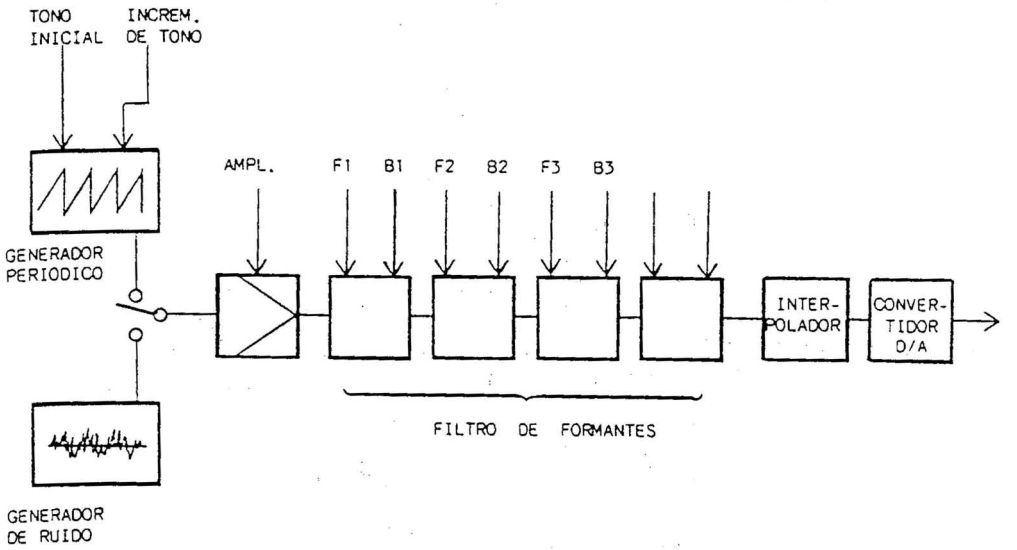


Fig.4. Diagrama de un sintetizador de 4 formantes.

Fundamentalmente, hay dos formas de trabajar según la forma de instalación de los filtros que dan lugar a cada una de las resonancias. Estos pueden estar en serie o en paralelo.

En los casos de sintetizadores serie se utilizan filtros en cascada de forma que la salida de uno se convierte en entrada del siguiente. Tienen el inconveniente de no permitir el control estrictamente independiente de cada una de las resonancias pero, por otro lado, se acomodan más a la forma física con que el tracto bucal modula la señal acústica.

Los sintetizadores paralelos generan las resonancias o antiresonancias a través de filtros independientes que suman sus efectos a la salida. Tienen ventajas en cuanto a la generación de sonidos que presentan ceros en su espectro como, por ejemplo, las consonantes nasales.

También se han diseñado sintetizadores que admiten las dos posibilidades, trabajando por uno de los dos métodos según las características de los fonemas que se quieren generar. Así, por ejemplo, el sintetizador de Klatt es una buena realización de esta filosofía, implantada por un sistema totalmente informatizado cuyos programas son bien conocidos (Klatt, 1980). Este sintetizador se ha utilizado en la Escuela Superior de Telecomunicaciones de Madrid para la síntesis del castellano con unos resultados bien satisfactorios (Rodríguez, 1984a).

La codificación del habla en un sintetizador por formantes se puede reducir a unos pocos parámetros como son: la frecuencia central y la banda de los primeros formantes, el nivel energético global, la presencia o ausencia de periodicidad y la frecuencia fundamental. Todo esto se puede codificar a un ritmo de unos 1000 bits/s (Flanagan, 1970). Aunque para una muy buena calidad de la voz habría que ampliar esta tasa hasta unos 4000 bits/s.

2.4. Sintetizadores de predicción lineal.

La técnica de predicción lineal (LPC), introducida en el campo de la voz por B.S.Atal y Suzane L.Hanauer (Atal, 1971) y por los japoneses F.Itakura y S.Saito (Itakura, 1972), parte de un tratamiento temporal de la señal acústica con ciertos

parámetros que permiten ahorrar la redundancia de información que se da en segmentos próximos de la voz. Esta técnica constituye una buena herramienta para la parametrización de la señal de voz, pero al mismo tiempo, por un proceso inverso, permite regenerar la señal acústica previamente parametrizada por un algoritmo LPC.

En cuanto a la eficiencia de la codificación por predicción lineal con 10 coeficientes y microfonemas de 10 ms, hay que decir que se obtiene una tasa de unos 5000 bits/s, que se pueden reducir a unos 3250 bits/s trabajando con coeficientes de reflexión. Con una más adecuada codificación se pueden alcanzar tasas de 2400 bits/s, como establece la norma americana LPC-10, para las transmisiones serie con 10 parámetros y microfonemas con una duración de 22,5 ms.

Este es uno de los sistemas con más posibilidades de futuro para la transmisión del habla codificada con una buena relación calidad-precio.

3. UNIDADES BASICAS DE CONCATENACION.

3.1. Síntesis a partir de fonemas.

El primer paso para la síntesis del habla estriba en la opción por un sistema concreto de parametrización, tal como se ha considerado en el apartado anterior. Ahora bien, la segunda opción se ha de referir al tipo de unidades básicas que habrá que retener en memoria para su posterior concatenación. Aquí hay que recordar un principio fundamental de la fonética de cualquier idioma que consiste en la coarticulación de fonemas adyacentes. Estamos acostumbrados a identificar la cadena fonética con un conjunto de símbolos gráficos discretos (las letras) como si los sonidos tuvieran también este carácter discreto. Esto no corresponde a la realidad acústica de una lengua hablada. Los sonidos correspondientes a una cadena fonética están fuertemente interrelacionados, de forma que pueden ser alterados por los fonemas anteriores y posteriores. La solución de guardar los fonemas y yuxtaponerlos acústicamente a lo largo del tiempo

conduce a un resultado que tiene muy poco que ver con la voz humana.

Si se opta por guardar en memoria los rasgos esenciales que definen la acústica de un fonema, habrá que disponer también de un conjunto de reglas que determinen la forma concreta de hacer la transición entre fonemas en cada caso particular. Este sistema es el que normalmente se denomina "síntesis por reglas". Es el que supone unidades más simples y, por tanto, también un conjunto más reducido de las mismas. En contrapartida, supone la aplicación de un sistema muy extenso de reglas para definir todas las posibles transiciones. En términos informáticos podemos decir que el fichero de unidades básicas sería muy reducido (entre 30 y 50 para idiomas occidentales), mientras que el conjunto de reglas o programas para decidir el tipo de coarticulación sería mucho más extenso.

Entre los que han trabajado la síntesis a partir de fonemas hay que recordar a J.N.Holmes (Holmes, 1964), que procede a una interpolación lineal de los formantes, atendiendo también a reglas de duración de las transiciones. Una realización más precisa es la de L.Rabiner (Rabiner, 1968), que introduce funciones exponenciales para la interpolación de formantes, con una constante de tiempo dependiente del orden del formante y de los fonemas adyacentes. Introduce también reglas suprasegmentales para generar el contorno de la frecuencia fundamental en función de la duración de la frase, de la situación de las vocales tónicas y de la perturbación introducida debido al carácter sordo o sonoro de las consonantes. El mismo autor habla de unos porcentajes de inteligibilidad del 80% y del 90% con este sistema.

En nuestro país hay que citar los trabajos de síntesis por reglas realizados en la Escuela Superior de Telecomunicaciones de Madrid, con unos resultados de inteligibilidad y calidad del habla muy satisfactorios. (Iglesias, 1981), (Rodríguez, 1984a).

También se ha trabajado en síntesis a partir de fonemas mediante parámetros articulatorios, de forma que cada fonema queda definido por una dis-

tribución de las cavidades del tracto bucal, y las reglas establecen la forma concreta con que se realiza el movimiento articulatorio. De hecho, estos parámetros coinciden de alguna forma con la función de área tal como se ha definido en el apartado 2.1). Uno de los investigadores que más han contribuido a sentar las bases teóricas de la síntesis por reglas es I.G.Mattingly (Mattingly, 1974).

3.2. Síntesis a partir de difonemas.

Para ahorrarse el problema de la coarticulación, se puede acudir a unas unidades básicas que incluyan ya toda la transición entre dos difonemas cualesquiera desde la parte estacionaria del primero hasta la parte estacionaria del segundo, de manera que la concatenación de unidades se haga precisamente por la parte central de cada fonema.

En algunos casos (Dixon, 1968) se han utilizado los difonemas con la ayuda de un sintetizador de formantes, de manera que el inventario contiene únicamente los parámetros que controlan las transiciones de cada difonema. En el momento de la edición se superponen también los valores paramétricos que controlan la frecuencia fundamental, la intensidad y la duración.

3.3. Síntesis a partir de semisílabas.

La semisílaba es una unidad diferente del difonema aunque, a primera vista, pudieron parecer semejantes. Si partimos del hecho de que la sílaba constituye un segmento suficientemente delimitado dentro de una palabra, podemos generar voz por yuxtaposición de sílabas, siempre que, posteriormente, podamos actuar sobre las variables de frecuencia fundamental, intensidad y duración que marcan los rasgos suprasegmentales indispensables para una buena calidad del habla. El inventario de sílabas en un idioma puede oscilar entre 4000 y 10000 unidades (Witten, 1982 p.162).

La estructura de cualquier sílaba presenta una parte vocálica central que puede ir precedida

y seguida por una consonante o un grupo consonántico. Si optamos por una segmentación que corte por la parte estacionaria de la vocal obtendremos dos segmentos: la semisílaba inicial y la semisílaba final. Con ello podemos reducir notablemente el inventario de unidades básicas. Para trabajos en inglés (Rosenberg, 1983) son suficientes unas 1000 semisílabas.

Diferentes experiencias de síntesis a partir de semisílabas han dado muy buenos resultados tanto en lo referente a la inteligibilidad como a la naturalidad del habla (Fujimura, 1976), (Macchi, 1977).

El sistema permite también disponer de ciertos prefijos y sufijos muy corrientes en el lenguaje y que se pueden tener ya préviamente sintetizados como un todo.

4. LOS CONVERSORES TEXTO-VOZ.

4.1. El paso de texto a sonido.

Para realizar una lectura correcta en cualquier idioma no es suficiente un conocimiento del alfabeto. La pronunciación correcta supone mucho más que una simple yuxtaposición o concatenación de sonidos. Los efectos de coarticulación entre fonemas próximos producen una alteración de sus rasgos acústicos en función del contexto. Un sistema conversor texto-voz ha de ser capaz de pronunciar correctamente un texto a partir de la cadena de caracteres ortográficos en código ASCII. La entrada de estos caracteres puede hacerse directamente por teclado o a partir de ficheros de texto previamente introducidos en el computador. No contemplamos aquí el paso previo que podría realizar la lectura óptica de estos caracteres a partir de texto impreso, por técnicas de reconocimiento de imagen.

En catalán y castellano la correspondencia entre caracteres ortográficos y sonidos es mucho más estrecha que en otras lenguas, como el caso del inglés. Pero, en cualquier caso, el paso de

texto a sonido dista mucho de ser algo inmediato, particularmente cuando se trata de conseguir una lectura con un alto grado de naturalidad e inteligibilidad.

4.2. La conversión de caracteres a fonemas.

La codificación de sonidos ha de realizarse según unos símbolos distintos de los caracteres ortográficos. En fonética se utilizan unos símbolos aceptados por la Asociación Fonética Internacional (AFI). La adaptación de estos símbolos a caracteres ASCII no está normalizada, pero se suelen utilizar tablas de conversión más o menos generalizadas como las del proyecto ARPA (SHOUP, 1980); aunque su utilización está pensada fundamentalmente para la lengua inglesa.

El número de fonemas (unos 40 para el catalán y unos 30 para el castellano) suelen ser algo mayor que el de caracteres; pero el número de sonidos distintos (alófonos) es aún mayor, por cuanto se distinguen diferentes formas de pronunciar un mismo fonema segun el contexto en que se encuentra.

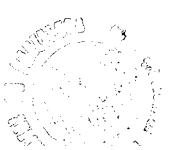
El proceso más sencillo de conversión automatizada de caracteres a sonidos se realiza generalmente por una comparación con los caracteres inmediatamente anteriores y posteriores de la secuencia ortográfica. Se obtiene así un conjunto de reglas que puede ser relativamente reducido, particularmente en lenguas más fonéticas como es el caso del castellano. En inglés se necesitan más de 500 reglas para realizar este proceso (Klatt, 1987). La concatenación de estos sonidos dentro de una sílaba se ha de hacer teniendo en cuenta las reglas de coarticulación que modifican notablemente las características propias de cada fonema. Generalmente se acude también a inventarios de excepciones más o menos extensos para tratar los casos más anómalos.

4.3. Tratamiento suprasegmental.

Una vez obtenidos los segmentos fundamentales (difonemas, semisílabas, sílabas, etc.) hay que tratar los rasgos que dan la continuidad a una frase determinada. Así, por ejemplo, las pausas no siempre coinciden con el espacio entre palabras, sino que se dan muchos casos de palabras encadenadas en una sola emisión de voz, lo cual se deduce a partir de la estructura sintáctica de la frase. Las pausas más fáciles de implementar son las que se deducen directamente a partir de los signos de puntuación.

Otro aspecto suprasegmental es la distribución adecuada de la intensidad con las atenuaciones correspondientes a los finales de palabra y de frase; aunque una buena parte de los cambios de intensidad vienen incluidos ya en la definición intrínseca de cada segmento o de cada fonema. Así, por ejemplo, los núcleos vocálicos coinciden siempre con un máximo de energía dentro de cada sílaba. La intensidad puede afectar también de un modo particular a las sílabas tónicas, aunque no conviene exagerar este rasgo para no dar una sensación de énfasis excesivo sobre las sílabas acentuadas, que rompe la fluidez y la continuidad de la emisión.

El rasgo suprasegmental que más contribuye a la naturalidad de la voz es la entonación o cambio de frecuencia fundamental. Dentro de cada palabra la subida de tono corresponde a la posición de la sílaba tónica, aunque en algunos casos también se puede señalar el acento por un descenso de la curva melódica. A nivel de frase existe también una acentuación que está igualmente ligada al cambio de tono y que contribuye decididamente a dar el sentido de la frase. Las subidas y bajadas de la frecuencia fundamental por razón del acento se han de combinar con una evolución global del tono que es propia del tipo de frase: enunciativa, interrogativa, admirativa, etc. Naturalmente todo este proceso no se puede introducir atomáticamente a partir de un texto determinado sin una determinada "comprensión" del sentido del mismo. Lo más fácil es introducir un tipo de entonación sistemática para frases interrogativas o exclamativas (clara-



mente detectables por los signos de interrogación o exclamación).

Otro aspecto claramente suprasegmental es la dilatación o contracción temporal de los segmentos elementales según su situación en la palabra y en la frase. Para indicar la presencia de sílabas tónicas es fundamental una dilatación del grupo vocálico correspondiente (un aumento de la cantidad en lenguaje fonético). Igualmente pueden estudiarse reglas de relentización del final de las palabras y del final de las frases.

4.4. Investigación actual sobre sistemas conversores texto-voz modelos comerciales.

Recientemente (Klatt, 1987) se ha presentado un estudio bastante completo del estado actual de la investigación en torno al tema y de sus aplicaciones comerciales. El estudio refleja particularmente las aplicaciones correspondientes a la lengua inglesa que, por otra parte, es la lengua en la que se han logrado resultados más elaborados. En dicho estudio pueden verse varias líneas de investigación, desde los primeros estudios teóricos, pasando por las respectivas reglas de conversión fonética y las experiencias de laboratorio, hasta las aplicaciones comerciales.

Es de advertir que la línea de investigación iniciada con modelos articulatorios (Dunn, 1950; Rosen, 1958; Coker, 1968...) no ha llegado al estadio de comercialización y se ha quedado únicamente en modelos de laboratorio. La línea de sintetizadores por formantes (Fant, 1953; Holmes, 1964; Mattingly, 1968; Rabiner, 1968; Rabiner, 1968; Klatt, 1970 ...) es la que ha aportado más resultados comerciales con un elevado nivel de calidad: los sistemas Prose-2000 comercializado por Speech Plus (1982), Dectalk comercializado por Digital Eq. a un precio de unos 4000 $ (1983) y el sistema sueco en varios idiomas Infovox (1983). El sistema Type-n-talk de Votrax en uno de los resultados más económicos pero con un nivel de calidad inferior. A nivel mucho más ambicioso se sitúan las máquinas de lectura Kurzweil a partir de ca-

racteres escritos, asequibles únicamente en bibliotecas públicas para uso de invidentes.

La concatenación a partir de difonemas o semisílabas arranca desde las investigaciones de Peterson (1958), Fujimura (1978), Dixon (1968), Olive (1977) y ha dado lugar a buenos resultados como los del sistema Echo (1983) y Conversant Systems (1987).

Respecto a los resultados obtenidos en otras lenguas se pueden mencionar la marca alemana AEG que utiliza el sistema Votrax. En francés se han realizado numerosos trabajos con resultados que dejan a esta lengua en un segundo puesto después del inglés. Así por ejemplo, la compañía Ferma ha comercializado el sistema F 5000 A para la conexión a la red telefónica (precio: unos 2000 $). La casa Vecsys comercializa la placa ICO85 en formato multibus (1000 $) y el sistema SYMPA, más económico (500 $), con conexión vía RS 232 para la conversión texto-voz. Igualmente XCOM tiene la placa CPS100 (1000 $).

En Suecia se ha desarrollado el ya mencionado sistema Infovox (1983) con las respectivas versiones en inglés, sueco, alemán, francés, italiano, español, ... con el modelo SA101PC para IBM-PC o compatibles y la máquina de escribir parlante portátil MULTI-TALK con multitud de aplicaciones para invidentes y personas con dificultades en el habla.

En España no se desarrollan aún productos comerciales referentes a la conversión texto-voz, pero sí que se han realizado estudios de investigación en torno a centros universitarios. En la Escuela Superior de Telecomunicaciones de Madrid se ha desarrollado un conversor texto-voz en castellano basado en el sintetizador de Klatt (Rodríguez, 1984a y 1984b). En la Escuela Universitaria de Telecomunicaciones la Salle Bonanova de Barcelona se ha desarrollado un primer modelo de conversor texto-voz en catalán (SINCAT) (Martí, 1986) y en castellano (SINCAS) (Martí, 1989).

5. BIBLIOGRAFIA.

Atal, B.S. & Hanauer, S.L. 1971. Speech analysis and synthesis by linear prediction of speech wave, *JASA* 50 pp. 637-655.

Coker, C,H 1968. Speech synthesis with a parametric articulatori model Speech Symposium, Kyoto. Paper a-4.

Cooper, F.S, Liberman, A.M. & Borst, J.M. 1951. The interconversion of audibles and visible patterns as a basis for research in the perception of speech, *Proc. Natl. Acad. Sci.* 37, pp. 318-325.

Dixon, N.R. & Maxey, H.D.1968. Terminal analogue synthesis of continuous speech using the diphone methos of segment assembly. IEEE Trans. Audio and Electroacoustics. AU-16, pp.40-50.

Dudley, H. 1939. Remaking speech, *JASA* 11, pp. 169-177.

Flanagan, J.L, Coker, C.H, Rabiner, L.R., Shafer, R,H & Umeda, N. 1970. Synthetic voices for computers, *IEEE* Spectrum 7, pp.22-45.

Fujimura, D. 1976. Sillables as concatenated demisillables and afixes, *JASA*,59 supl. 1 p. S55.

Holmes, J.N., Mattingly, I.G. & Shearne, J.N. 1964. Speech synthesis by rule, *Language and Speech* 7, pp. 127-143.

Holmes, J.N. 1980. The JSRU chanel vocoder, *Proc. Institute of Electircal Engineers,* 127 (F1), pp. 53-60.

Iglesias, E. & Meneses, J. 1981. Phonetical processing of ortographical text in a speech synthesis by rule system for Spanish, *Vigo Workshop on Signal Processing and its Applications*. Vigo, julio 1981.

Itakura, F. & Saito, S. 1972. On the optimum quantization of feature parameters in the parcor speech synthesizer". *Proc. Conf. Speech Commun. Process*. pp.434-437.

Klatt, D.H. 1980. Software for a cascade/parallel formant synthesizer, *JASA* 67, pp.971-995.

Klatt, D.H. 1987. Text-to-speech conversion, *JASA* 82 pp. 737-793.

Lovins, J.B. & Fujimura, O. 1976. Synthesis of English monosyllables by demisyllable concatenation, *92nd. Meeting of Acoustical Society of America*. San Diego, California. Noviembre, 1976.

Macchi, M.J. & Nigro, G. 1977. Sillable affixes in speech synthesis, *93ed. Meeting of Acoustical Society of America*. Pensylvania State University, junio 1977.

Martí, J. 1986. SINCAT. El sintetizador català de veu, *Quaderns Tècnics* 7, pp. 13-17.

Martí, J. 1989. SINCAS-SINCAT: Un conversor text-veu en castellà i català, Comunicación presentada a la Convención Informática Latina, marzo 1989.

Mattingly, I.G. 1974. *Current trends in linguistics*. Vol. 12. Speech synthesis for phonetic and phonological models, pp. 2451-2487. Ed. Mouton the Hague.

Rabiner, L. 1968. Speech synthesis by rule: an acoustic domain approach, *Bell System Tech. J.* 47, pp 17-37.

Rodríguez, M., Olabe, J.C., Santos, A., Muñoz, P., Villaseca, I. & Muñoz, E. 1984a. Visión panorámica de la respuesta oral de máquinas, *Mundo Electrónico* 144, pp 57-66.

Rodríguez, M., Iglesias, E., Martínez, R. & Muñoz, E. 1984b. Alternativas para síntesis de voz, *Mundo Electrónico* 144, pp. 67-79.

Rosenberg , A.E., Rabiner, L.R., Wilpon,J.G. & Kahn, D. 1983. Demisyllable-based isolated word recognition system, *IEEE ASSP* 31, pp. 713-726.

Shoup, J.E. 1980. Phonological aspects of speech recognition, in *Trends in speech recognition* edited by W.A.LEA. Prentice-Hall pp. 125-138.

Witten, I.H. 1982. *Principles of computer speech*. Ed. Academic Press.