Llisterri, J., & West, M. (1987). Los sistemas de conversión de texto a voz mediante síntesis por reglas: Una aproximación interdisciplinar. In C. Martín Vide (Ed.), *Lenguajes naturales y lenguajes formales II. Actas del II congreso de lenguajes naturales y lenguajes formales* (pp. 183-196). Barcelona: Promociones y Publicaciones Universitarias.

http://liceu.uab.cat/~joaquim/publicacions/Llisterri_West_87_Conversion_Texto_Voz.pdf

LOS SISTEMAS DE CONVERSIÓN DE TEXTO A VOZ MEDIANTE SÍNTESIS POR REGLAS: UNA APROXIMACIÓN INTERDISCIPLINAR

Joaquim Llisterri Boix
Laboratori de Fonètica
Facultat de Lletres
Universitat Autònoma de Barcelona

Martin West
Department of Applied Acoustics
University of Salford
-Gran Bretaña-

"There are no contemporary speech synthesis systems that begin to approach the level of surface phonetic variability observed in natural speech".

Allen, 1985b, pp. 1546-1547.

0. Introducción.

La exploración de las posibilidades de interacción con los ordenadores mediante la voz es tal vez uno de los temas privilegiados dentro de la investigación actual en el campo de la comunicación hombre-máquina. Si bien una parte considerable de los trabajos se orienta hacia la comprensión del habla por el ordenador, la generación automática de un texto oral a partir de una representación escrita -síntesis del habla- no es aún un problema completamente resuelto, tal como muestra la afirmación de uno de los principales investigadores del momento en este campo con la que encabezamos el presente trabajo. A fin de precisar los requisitos que debe satisfacer un sistema de conversión de texto a voz puede ser útil empezar por unas breves consideraciones sobre las ventajas de la salida vocal en ordenadores (Llisterri, 1985; Nadeu y Mariño, 1985; Witten, 1982):

(1) Libertad para la realización de otras tareas. Cuando el ordenador nos hace llegar sus mensajes mediante la voz, podemos tener las manos y la vista libres y el acceso a la información no está limitado a la existencia de un entorno bien iluminado; una salida sonora es omnidireccional y permite una mayor movilidad en el trabajo; por otra parte, los mensajes orales generados por la máquina no interfieren con otras actividades.

(2) Acceso telefónico. Para acceder por vía telefónica a una terminal de ordenador no se necesita un equipo complejo y se trata además de un método rápido e independiente de la distancia.

(3) Economía. Las salidas vocales requieren únicamente componentes electrónicos que pueden fabricarse a bajo costo. Desde el punto de vista del ahorro de tiempo, la cantidad de información que puede ofrecerse mediante el habla supera con mucho a la que se consigue leyendo un texto en la pantalla de una terminal.

(4) El habla es el modo de comunicación más natural y universal de la especie humana. La interacción con los ordenadores es mucho más accesible si se utiliza el lenguaje natural. No hay que olvidar tampoco que así se facilita la utilización de los equipos informáticos a los invidentes.

Existen actualmente diversas técnicas para la generación de mensajes orales por ordenador (varias de ellas se presentan en Cater, 1983; Flanagan, 1982; Holmes, 1981; Lee y Lochovsky, 1983; Mariño, Nadeu y Llisterri, 1987; Zue, 1982), aunque en el presente trabajo nos limitaremos a las características esenciales de los sistemas de conversión de texto a voz, centrándonos en los problemas específicos que requieren la colaboración entre lingüistas, especialmente fonetistas, e ingenieros.

1. La conversión de texto a voz en la comunicación con los ordenadores: necesidad y aplicaciones.

Un conversor de texto a voz puede definirse como un sistema que transforma cualquier texto escrito siguiendo las convenciones ortográficas de una determinada lengua en su equivalente hablado.

Los sistemas de conversión de texto a voz son necesarios cuando el ordenador debe generar una serie infinita de mensajes imposible de prever. Las técnicas de síntesis basadas en el almacenamiento de ondas sonoras codificadas tienen limitaciones inherentes en cuanto al número de enunciados que pueden llegar a producir y son de muy poca utilidad cuando la entrada al sistema está constituida por información ilimitada. Ciertas aplicaciones de la síntesis de voz requieren precisamente la posibilidad de generar en tiempo real un número infinito de mensajes impredictibles. A continuación nos referiremos brevemente a algunos de estos casos (Sclater, 1983; Stella, 1985).

(1) Máquinas lectoras para ciegos.

La construcción de máquinas lectoras para ciegos es uno de los objetivos prioritarios en tecnología de la voz (Allen, 1973; Cooper, Gaitenby, Mattingly y Umeda, 1969; Lee, 1969). La máquina lectora de Kurzweil, introducida en el mercado a finales de los setenta y actualmente en uso en bibliotecas públicas y otras instituciones, constituye el intento más avanzado en este campo; consiste en un sofisticado conversor de texto a voz asociado a un sistema de reconocimiento óptico que permiten leer en voz alta cualquier documento impreso.

(2) Ayudas para los disminuidos físicos.

Las personas que se encuentran incapacitadas para hablar y usan algún tipo de ayuda que les permite intercambiar información con sus semejantes deben ser capaces de producir cualquier mensaje que deseen, y no sólo aquellos que han sido almacenados previamente en su sintetizador. Sólo los conversores de texto a voz pueden ofrecer una posibilidad real de comunicación que permita responder a nuevas situaciones. Hoy en día están investigándose diversos métodos de control de la prótesis -por ejemplo, mediante movimientos oculares- diseñados especialmente para aquellos minusválidos a los que les es imposible utilizar un teclado convencional.

(3) Terminales hablantes.

En ciertos casos es necesario poder obtener de la máquina cualquier mensaje oral, aunque éste sea completamente nuevo, especialmente cuando se accede por teléfono a la información. Esta es la situación que se da cuando se consultan telefónicamente descripciones de productos, listas de precios, listas de existencias de almacenes, o cuando en el ámbito de la industria se reemplazan los mensajes escritos en la terminal por indicaciones verbales sobre el estado de un determinado proceso. Igualmente se requiere un sistema muy flexible para obtener información de grandes bases de datos -por ejemplo, la base lexicográfica del TERMCAT-.

Otra posibilidad actualmente utilizada en aplicaciones con un elevado grado de complejidad es la síntesis a partir de conceptos; mediante esta técnica es el mismo sistema el que genera el mensaje además de convertirlo en sonido. Un ejemplo lo constituye una estación de distribución de aguas controlada enteramente por computador en Birmingham (Fallside y Young, 1984).

Los sistemas de conversión de texto a voz se emplean también en enseñanza asistida por ordenador, como refuerzo de los estímulos visuales y como procedimiento para que los estudiantes que no pueden leer (invidentes o demasiado jóvenes) utilicen programas educativos (Gray, 1984).

Es fácil darse cuenta de que muchas de las aplicaciones que hemos citado sólo serán totalmente útiles cuando se combinen con un programa de reconocimiento de voz independiente del hablante y cuando la calidad del habla sintetizada se acerque suficientemente a la del habla natural y no se degrade al utilizar el teléfono como medio de transmisión. Cada una de las aplicaciones requiere una técnica de síntesis específica, pero la selección de ésta depende básicamente de dos factores: la calidad del habla que debe producir el sistema -es decir, la naturalidad y la inteligibilidad- y la flexibilidad que se necesite. Al mencionar aquí la flexibilidad nos referimos a la capacidad que debe tener el sistema para generar diversos mensajes a partir de un conjunto finito de elementos. Los sistemas de conversión de texto a voz existentes en la actualidad ofrecen un grado muy elevado de flexibilidad, pero en cambio no han llegado aún a alcanzar la calidad propia del habla natural.

El lenguaje, como bien saben los lingüistas, puede estructurarse en varios niveles, lo cual permite concebir diversas técnicas de síntesis del habla según la unidad empleada en el proceso de almacenamiento y producción de los mensajes hablados (Dilts, 1984):

(1) Frases.

El almacenamiento de frases completas mediante cualquiera de los procedimientos existentes para la codificación de la onda sonora permite obtener voz de una calidad y una naturalidad prácticamente insuperables, pero se diferencia muy poco de las grabaciones analógicas. Es una técnica que presenta una falta total de flexibilidad, puesto que implica almacenar cada una de las frases que deben producirse.

(2) Palabras.

El almacenamiento de palabras permite una mayor flexibilidad y economía, pero presenta dos problemas graves: por un lado, es imposible reproducir las variaciones contextuales de los sonidos en función de su entorno fonético; por otro, se hace difícil mantener las variaciones de entonación según la frase y la posición de la palabra en la frase.

(3) Morfemas.

Como unidad de síntesis, los morfemas presentan problemas análogos a las palabras, pero aún más acusados. El habla así producida es muy poco natural debido a las dificultades de concatenación de los elementos constituyentes.

(4) Fonemas o alófonos.

Desde el punto de vista del lingüista, ésta parece ser la unidad más natural para la producción de habla y también es la más económica desde la perspectiva del ingeniero. El fonema se considera la mínima unidad abstracta en la que puede dividirse un enunciado, aunque se reconozca la existencia de elementos menores como los rasgos distintivos. Cualquier lengua natural puede describirse mediante unos 30-40 fonemas, cuya combinación en unidades de nivel más alto -morfemas, palabras y frases- permite la creación de cualquier enunciado en esta lengua.

En la teoría lingüística se han dado y se dan aún polémicas sobre la existencia misma del fonema, sobre su adecuación como nivel de análisis o incluso sobre su validez psicológica, pero es innegable que el fonema constituye una buena manera de reducir la información que debe emplearse en un sistema de conversión de texto a voz que deba combinar una memoria pequeña con una gran flexibilidad. Tal como ha señalado Allen (1976), cuando se necesita un procedimiento que permita construir enunciados a partir de unidades simples, debemos pensar en unidades situadas en un cierto nivel de abstracción. Mediante el conjunto de fonemas propio de una lengua y las reglas para su combinación debería ser posible generar cualquier enunciado en dicha lengua. Desafortunadamente, como veremos más adelante, el problema es mucho más complejo. La técnica a la que nos estamos refiriendo, conocida como síntesis por reglas, se usa en la actualidad en conversores de texto a voz utilizados en aplicaciones que requieren un alto grado de flexibilidad, como muchas de las que hemos enumerado anteriormente.

Es importante señalar aquí que todo lo dicho en este punto constituye un ejemplo claro de la interacción entre el conocimiento teórico de la estructura de la lengua y el diseño y la realización de una aplicación tecnológica encaminada a la resolución de un problema práctico.

2. El desarrollo de los sistemas de conversión de texto a voz mediante síntesis por reglas.

El concepto de síntesis por reglas nació a finales de los años cincuenta en los Laboratorios Haskins, en Estados Unidos. El proyecto inicial consistía en desarrollar el aprendizaje de la lectura de espectrogramas en sordos de modo que pudieran comprender la lengua hablada sin necesidad de recurrir a la utilización del lenguaje de los signos o a transcripciones escritas;

esta idea inicial desembocó en una investigación sobre la estructura acústica del habla y, en última instancia, sobre la codificación de la información acústica que nos permite interpretar como un código lingüístico los mensajes sonoros que recibimos (Potter, Kopp y Green, 1947). El equipo de Haskins desarrolló, intentando encontrar la información acústica esencial para la comprensión del habla, un método para la generación artificial de enunciados.

El llamado *Pattern Playback* es un instrumento que "lee en voz alta" representaciones de la evolución de la frecuencia y la amplitud de la onda sonora en función del tiempo (espectrogramas) que dibuja el propio investigador. El objetivo de este trabajo consistía en eliminar toda la información no significativa que se encuentra en las señales acústicas del habla a fin de reducirla a sus unidades mínimas. Se desarrollaron así una serie de reglas para crear las representaciones espectrográficas que, producidas según este conjunto de reglas, dieran como resultado un mensaje identificable. Con un conjunto finito y limitado de reglas pudo llegar a sintetizarse cualquier frase (Liberman, et al., 1959).

Es importante señalar en este punto el hecho de que se desprende de estos estudios la necesidad de trabajar con unidades menores que los fonemas. Por otra parte, las investigaciones en otros campos, especialmente en fonología estructuralista, llevaron al concepto de rasgo distintivo, unidades en las que puede dividirse el fonema. La conjunción de las investigaciones en fonética acústica y en teoría lingüística llevó a comprobar la necesidad de recurrir a elementos subfonémicos para explicar de qué manera se codifica un mensaje en la onda sonora, ya que un determinado fonema no mantiene una relación biunívoca con un único punto de la cadena sonora (Jakobson, Fant y Halle, 1952; Jakobson y Halle, 1956).

Simultáneamente, el científico sueco C.G.Fant desarrolló un modelo acústico de producción del habla (Fant, 1960) que constituye la base de muchos de los sintetizadores actuales. La idea esencial consiste en considerar el tracto vocal como un filtro variable y en tratar la señal acústica producida como el resultado de la interacción entre una fuente y un filtro. Esto hizo posible la descripción del habla a partir de un conjunto limitado de parámetros relacionados con el comportamiento acústico del tracto vocal, a la vez que permitió disponer de un modelo fácilmente implementable sea en forma de circuito electrónico o de programa.

Desde el momento en que fue posible describir el habla como una representación parametrizada y contando con un cuerpo de conocimientos sobre las propiedades acústicas de las señales sonoras en relación con unidades lingüísticas como los rasgos distintivos, no había, en teoría, barreras para poner a punto un mecanismo que transformara en una representación sonora los valores de los parámetros acústicos relacionados con el habla, sistema que constituye la base de un conversor de texto a voz.

Probablemente fue Jonathan Allen (1976) uno de los primeros investigadores en sugerir la idea de llevar a cabo la síntesis automática de un texto escrito sin restricciones de ninguna clase. Este proyecto contribuyó a establecer un conjunto de reglas que convierten las representaciones ortográficas convencionales en su representación fonológica o alofónica como primer paso para su conversión en sonido. Hay que destacar que las concepciones de Allen estuvieron fuertemente influidas por el desarrollo de la lingüística generativa en el Massachusetts Institute of Technology . El conversor de texto a voz por él realizado se concibe como una transformación entre dos formas superficiales -texto escrito y cadena sonora- que se lleva a cabo a partir de una representación lingüística subyacente común a ambas. Las reglas de conversión de letra a fonema utilizaron al principio los trabajos en fonología generativa de Chomsky y Halle (1968), hasta integrar en las últimas versiones del sistema los resultados de las investigaciones sobre fonología y morfología que se llevan a cabo en el marco de diversos modelos lingüísticos de inspiración chomskiana. Esta interacción entre la teoría lingüística y las necesidades de la tecnología se ha llevado a cabo de forma bidireccional: en ciertos casos la necesidad de explicitar las reglas ha llevado a estudiar en profundidad ciertos problemas de la lengua para los que no se disponía de una descripción sistemática.

Los conversores de texto a voz pueden dividirse en la actualidad en dos grandes grupos:

(a) Versiones de laboratorio en forma de prototipo diseñadas especialmente como herramienta de investigación, aunque algunos de ellos hayan sido comercializados. En esta categoría se incluyen sistemas como el conversor de texto a voz desarrollado por Allen en el MIT (MITalk, descrito en Allen, 1981; 1985b) o el sintetizador de formantes en paralelo creado en la Joint Speech Research Unit en Gran Bretaña (Holmes, 1979). En España, es imprescindible mencionar los trabajos conjuntos del Departamento de Electrónica de la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universidad Politécnica de Madrid y del Laboratorio de Fonética del Consejo Superior de Investigaciones Científicas, que han dado lugar a la realización de un conversor de texto a voz para el castellano utilizando el sintetizador de Klatt (Rodríguez, et al., 1984; Martínez, et al., 1986), y el conversor de texto a voz para el catalán diseñado por J.Martí en el Laboratorio de Acústica de la Escuela Técnica de Ingenieros de Telecomunicación de la Salle en Barcelona partiendo del difonema como unidad de síntesis (véase su comunicación en este mismo volumen).

(b) Sistemas comercializados que se encuentran en el mercado en forma de módulos acabados o programables según las especificaciones del usuario. Tales sistemas tienen como requisitos básicos un bajo costo, un pequeño tamaño y una baja consumición de energía, al mismo tiempo que deben hacer posible la síntesis en tiempo real y deben disponer también de los interfaces estándar que permitan su utilización con diversos modelos de orde-

nadores personales. En conjunto constituyen versiones menos sofisticadas de los sistemas que entrarían en el primer apartado, faltándoles la flexibilidad requerida para los sistemas de investigación. La literatura sobre síntesis del habla contiene abundantes descripciones y referencias sobre tales sistemas (Cater, 1983; Moller, Stork, Gunawardana y Gilblom, 1984; Poulton, 1983; Sclater, 1983), por lo que no vamos a proceder aquí a una revisión detallada de todos ellos. Citaremos sólo entre los más conocidos el *Type-n-Talk* de Votrax, el sistema *Prose 2000* de Telesensory o *DECtalk* de Digital, entre otros.

3. Estructura de los sistemas de conversión de texto a voz.

En este apartado resumiremos las características esenciales de los conversores de texto a voz, señalando brevemente cuáles son las principales etapas por las que debe pasar un sistema de este tipo para transformar las representaciones ortográficas en habla. La estructura general de un sistema de conversión de texto a voz puede esquematizarse tal como se presenta en la figura 1:

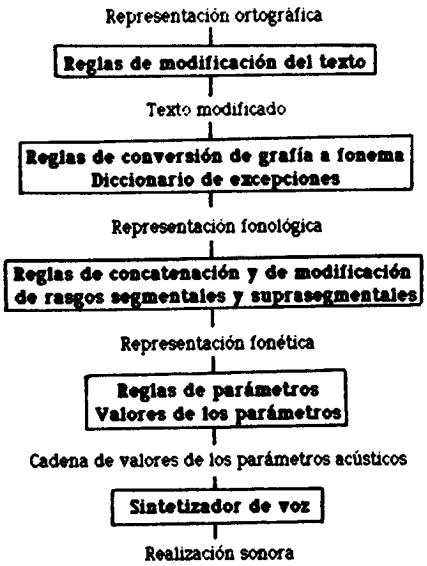

Fig. 1, Estructura general de un sistema de conversión de texto a voz.

Cabe señalar que en el presente trabajo no haremos ninguna referencia a la síntesis de los elementos suprasegmentales, que requeriría un tratamiento más detallado del que permite el espacio disponible aquí.

3.1. Procesamiento previo del texto escrito.

Un texto escrito que deba ser utilizado como entrada de un conversor de texto a voz requiere un procesamiento previo. Las operaciones necesarias son similares a las que lleva a cabo un hablante humano cuando lee en voz alta un texto, y comprenden el tratamiento de las siglas, de los números y de las abreviaciones, que deben aparecer en su forma completa.

3.2. Conversión de letras a fonemas.

El primer paso, una vez han operado las reglas de procesamiento del texto, es la conversión de letras a fonemas o a alófonos, según el tipo de sistema de que se trate. Para llevar a cabo esta operación el lingüista debe escribir primero una gramática, que tendrá la forma de un conjunto de reglas que relacionan las letras y los sonidos de una lengua dada. A continuación, el ingeniero debe dotarla de una forma computacionalmente aceptable. Normalmente, el conocimiento de las reglas de asociación entre grafías y sonido y la capacidad de programar se dan en especialistas distintos, por lo que es necesario llegar a un cierto grado de colaboración. Es posible que los usuarios de los sistemas de conversión de texto a voz diseñados más específicamente para la investigación deseen efectuar cambios en las reglas y verificar inmediatamente los resultados en la inteligibilidad de las producciones; si el usuario no es programador, se requiere así el uso de herramientas que faciliten la interacción con el ordenador (Hertz, 1982). El formalismo utilizado en las reglas es muy similar al de Chomsky y Halle (1968), ampliamente conocido de los lingüistas. Puede señalarse también que los cambios introducidos en estas reglas pueden producir diversas variedades de una misma lengua.

Algunos sistemas como el MITalk introducen un estadio intermedio en el que se analiza la estructura interna de la palabra; en tal caso suele existir un diccionario de morfemas a los que se asocia su pronunciación, ortografía y reglas de combinación con otros elementos morfológicos en forma de diccionario de excepciones, tratándose las palabras polimorfémicas como una cadena de morfemas. En tales sistemas el procesamiento morfológico es previo a la conversión de letra a sonido, que sólo se lleva a cabo cuando ya se ha realizado el análisis morfológico.

3.3. Reglas alofónicas.

Las reglas alofónicas son los mecanismos por los que se definen las características fonéticas precisas de los fonemas que aparecen en el nivel descrito anteriormente. De su aplicación resulta un reajuste de las representaciones fonológicas generadas por la aplicación de las reglas de conversión de letra a sonido. La necesidad de estas reglas responde a que en el habla natural cada sonido está influido por los sonidos adyacentes y parece lógico pensar que un conversor de texto a voz que aspire a generar habla natural debe operar de forma parecida. La función de las reglas alofónicas es seleccionar el alófono apropiado a un contexto fonético particular, es decir, en términos lingüísticos, modelar la coarticulación.

Este segundo conjunto de reglas se aplica junto con una tabla en la que se definen los valores de cada uno de los parámetros acústicos que especifican la realidad sonora de los alófonos; para obtener tales valores, se parte del análisis acústico del habla natural de hablantes de la lengua que se pretende sintetizar (Llisterri y West, 1983).

Comentaremos aquí con cierto detalle el funcionamiento de esta parte del sistema en el conversor de texto a voz que se está desarrollando en la Universidad de Salford (Crompton, 1982; Crompton y West, 1983). Para su operación, el sistema utiliza tres archivos de datos: RSPAR, RSALF y RSTIF. El fichero principal RSPAR consta de 114 líneas con 18 valores de parámetros en cada una. RSPAR contiene los valores de cada uno de los 15 parámetros acústicos que para operar necesita el sintetizador, de modo que en cada línea se especifica una entrada para cada uno de los 114 fonemas o alófonos del inglés usados por el sistema. Los parámetros acústicos por los que se definen las características sonoras de cada alófono son los siguientes:

1. AV: amplitud de la sonoridad.
2. AH: amplitud de la aspiración, para aquellos casos en los que se da una fricción en la cavidad bucal sin estar acompañada de sonoridad.
3. AN: amplitud de la nasalidad.
4. F_0 : frecuencia del fundamental.
5. F_1 : frecuencia del primer formante.
6. F_2 : frecuencia del segundo formante.
7. F_3 : frecuencia del tercer formante.
8. N_1 : frecuencia del formante nasal.
9. B_1 : amplitud de banda del primer formante.
10. B_2 : amplitud de banda del segundo formante.
11. B_3 : amplitud de banda del tercer formante.
12. AC: amplitud de la fricción, especialmente en bandas superiores a 1.500 Hz.

13. K_1 : primer formante fricativo.
14. K_2 : segundo formante fricativo.
15. AK: relación entre polo y cero en la fricción.

Los parámetros temporales son los siguientes:

16. TI: duración de la transición de la consonante a la vocal.
17. TS: duración de la parte estacionaria de la vocal.
18. TO: duración de la transición de la vocal a la consonante.

Debe señalarse aquí que los valores de las transiciones de la consonante a la vocal se consideran, desde el punto de vista del proceso de síntesis, como integrantes de la consonante y no como propiedades de las vocales.

RSALF es un fichero de alófonos, y RSTIF es otro fichero que contiene información temporal que complementa la que se encuentra en RSPAR; durante la operación del sistema ambos son examinados conjuntamente con el fichero principal.

Para generar una secuencia sonora a partir de la transcripción fonética, la subrutina correspondiente del programa de síntesis toma el primer fonema entrado junto con la parametrización que de él se encuentra en el RSPAR y busca en la tabla de RSALF si existe algún alófono de este fonema en función del que le sigue en la cadena. Se seleccionan después los valores de los parámetros acústicos propios de ambos fonemas y se interpola entre cada uno de los parámetros de ambos alófonos, creando tantos puntos de interpolación como requiera el valor de TS, es decir, de la duración de la parte estable de la vocal. A continuación se añaden el resto de los parámetros temporales, repitiéndose el proceso en intervalos regulares.

Hay que hacer notar aquí, aunque sin duda el lector lingüista ya lo habrá observado, que los alófonos contenidos en los ficheros no se seleccionan en términos estrictamente lingüísticos, sino que representan más bien variantes fonéticas determinadas a partir del habla natural, aplicando diversas técnicas de análisis acústico a un corpus estadísticamente representativo de combinaciones consonante-vocal. Los alófonos se definen pues en términos de variaciones de los puntos de inicio de las transiciones de consonante a vocal.

3.4. Conversión de parámetros a sonidos.

Una vez se ha generado una cadena con los valores de los parámetros asociados a cada alófono, ésta debe convertirse en una señal analógica; este proceso lo lleva a cabo habitualmente un sintetizador -el OVEIIId contro-

lado por un ordenador PDP/11-03 en la Universidad de Salford o el programa analógico de síntesis por formantes de Klatt en otros sistemas-, enviándose el resultado a un amplificador y un altavoz, con lo cual se obtiene la versión sonora de la cadena de caracteres.

4. Consideraciones finales.

Se ha visto, pues, que el objetivo de la investigación en síntesis por reglas puede ser tanto la construcción de un sistema capaz de resolver los problemas que se le plantean a un ingeniero cuando se enfrenta al problema de generar de la forma más efectiva posible un conjunto infinito de mensajes, como la adquisición de nuevos conocimientos de naturaleza básica sobre los procesos de producción y percepción del habla. Los modelos realmente útiles en síntesis deben combinar ambos objetivos. Creemos que los conversores de texto a voz constituyen un marco extraordinariamente útil para promover la interacción entre dos especialistas del estudio del habla: el lingüista y el ingeniero. Estamos convencidos de que dichos sistemas sólo llegarán al nivel al que aspira Allen cuando sean el resultado de una colaboración fructífera entre ambos expertos.

Referencias.

Allen, J. (1973), "Reading Machines for the Blind: The Technical Problems and the Methods Adopted for Their Solution". IEEE Transactions on Audio and Electroacoustics, AU-21, 3, pp. 259-264.

Allen, J. (1976), "Synthesis of Speech from Unrestricted Text", Proceedings of the IEEE, 64, 4, pp. 433-442.

Allen, J. (1981), "Linguistic-based algorithms offer practical text-to-speech systems", Speech Technology, 1, 1, pp. 12-16.

Allen, J. (1985a), "Speech Synthesis from Unrestricted Text", en F.Fallside y W.A.Woods, eds., Computer Speech Processing, Prentice-Hall Int., Londres, pp. 461-478.

Allen, J. (1985b), "A Perspective on Man-Machine Communication by Speech", Proceedings of the IEEE, 73, 11, pp. 1541-1550.

Cater, J.P., (1983), Electronically Speaking: Computer Speech Generation, Howard W. Sams, Indianapolis.

Chomsky, N. y Halle, M. (1969), The Sound Pattern of English, Harper & Row, Nueva York.

Cooper, F.S., Gaitenby, J.H., Mattingly, I.G. y Umeda, N. (1969), "Reading Aids for the Blind: A Special case of Machine-to-Man Communication", IEEE Transactions on Audio and Electroacoustics, AU-17, 4, pp. 266-270.

Crompton, C.K. (1982), The Extraction of Parameters from Natural CVs (Consonant-Vowel Pairs) Suitable for the Production of a Copy Synthetic Objective, MA. Thesis, Department of Modern Languages, University of Salford, Salford.

Crompton, C.K. y West, M. (1983), "The Extraction of Parameters from Natural Consonant-Vowel Pairs for the Production of Copy Synthetic Equivalents", en 11th International Congress on Acoustics, vol. 4, pp. 179-182.

Dilts, M. (1984), "Text to Speech", en G.Bristow, ed., Electronic Speech Synthesis. Techniques, Technology and Applications, Granada, Londres, pp. 94-113.

Fallside, F. y Young, S. (1984), "Speech output from complex systems", en G.Bristow, ed., Electronic Speech Synthesis. Techniques, Technology and Applications, Granada, Londres, pp. 274-287.

Fant, C.G. (1960), Acoustic Theory of Speech Production, Mouton, La Haya.

Flanagan, J. (1982), "Talking with Computers: Synthesis and Recognition of Speech by Machines", IEEE Transactions on Biomedical Engineering, BME-29, 4, pp. 223-232.

Gray, T. (1984), "Talking Computers in the Classroom", en G.Bristow, ed., Electronic Speech Synthesis. Techniques, Technology and Applications, Granada, Londres, pp. 243-259.

Hertz, S. (1982), "From text to speech with SRS", Journal of the Acoustical Society of America, 72, 4, pp. 1.155-1.170.

Holmes, J. (1979), "Synthesis of natural-sounding speech using a formant synthesizer", en B.Lindblom y S. Öhman, eds., Frontiers of Speech Communication Research, Academic Press, Londres, pp. 275-285.

Holmes, J. (1981), "Machines that Speak", Acoustics Bulletin, octubre, pp. 16-21.

Jakobson, R., Fant, C.G. y Halle, M. (1952), Preliminaries to Speech Analysis, MIT Press, Cambridge, Mass.

Jakobson, R. y Halle, M. (1956), Fundamentals of Language, Mouton, La Haya.

Lee, F.F. (1969), "Reading Machine: From Text to Speech", IEEE Transactions on Audio and Electroacoustics, AU-17, 4, pp.275-282.

Lee, D.L. y Lochovsky, F.H. (1983), "Voice Response Systems", Computing Surveys, 15, 4, pp. 351-373.

Liberman, A.M., et al. (1959), "Minimal rules for Synthesizing Speech", Journal of the Acoustical Society of America, 31, 11, pp. 1.490-1.499.

Llisterri, J. (1985), "Sobre màquines parlants", Papers de Batxillerat, 3, 8, pp. 216-220.

Llisterri, J. y West, M. (1983), "Analysis of stop-vowel transitions in Catalan", en 11th International Congress on Acoustics, vol. 4, pp. 279-283.

Mariño, J.B., Nadeu, C. y Llisterri, J. (1987), "Síntesis automática del habla", en Inteligencia artificial: conceptos, técnicas y aplicaciones, Marcombo, Barcelona.

Martínez, M., et al., (1986), "Conversión automática texto-habla y su relación con el procesamiento del lenguaje natural", en C.Martín Vide, ed., Lenguajes naturales y lenguajes formales, I, Universitat de Barcelona, Barcelona, pp. 366-375.

Moller, C., Stork, J., Gunawardana, R. y Gilblom, D. (1984), "Ready-to-use speech systems", en G.Bristow, ed., Electronic Speech Synthesis. Techniques, Technology and Applications, Granada, Londres, pp. 192-211.

Nadeu, C. y Mariño, J.B. (1985), "Comunicación oral con el computador", Mundo Electrónico, 149, pp. 108-116.

Potter, R., Kopp, G.A. y Green, H. (1947), Visible Speech, Bell Telephones Laboratories, Nueva York.

Poulton, A.S. (1983) Microcomputer Speech Synthesis and Recognition, Sigma Technical Press , Wilmslow.

Rodríguez, M., et al. (1984), "Visión panorámica de la respuesta oral de máquinas", Mundo Electrónico, 144, pp. 57-66.

Sclater, N. (1983) Introduction to Electronic Speech Synthesis, Howard W. Sams, Indianapolis.

Stella, M. (1985), "Speech Synthesis", en F.Fallside y W.A.Woods, eds., Computer Speech Processing, Prentice-Hall Int., Londres, pp. 421-460.

Witten, I.H. (1982), Principles of Computer Speech, Academic Press, Londres.

Zue, V. (1982), "Computer Voice Response and Speech Synthesis", Trends and Perspectives in Signal Processing, 2, 4, pp. 7-9.