La conversión de texto en habla

Text-to-speech synthesis

Joaquim Llisterri

La conversión de texto en habla permite a un ordenador transformar automáticamente un texto en su correspondiente forma sonora, lo más parecida posible a como lo haría un ser humano. Entre otras aplicaciones, esta tecnología lingüística permite acceder a través del teléfono a textos digitales (como una página web o un mensaje de correo electrónico), o facilitar el uso de los sistemas informáticos a personas con necesidades especiales.

The conversion of text into speech allows computers to automatically transform a text into its corresponding sounds, as close as possible to the way humans read. Among other applications, this linguistic technology provides telephone access to digital texts (such as web pages or e-mail messages), and facilitates the use of computerized system to people with special needs.

Conversar con los ordenadores

Básicamente, la interacción mediante el habla entre las personas y los ordenadores requiere tres componentes: un sistema que permita que el ordenador procese la lengua oral, otro que se encargue de proporcionar información verbal al usuario y un tercero que gestione la comunicación. Para la primera tarea se utiliza el reconocimiento del habla, que facilita a los sistemas informáticos la conversión de un enunciado pronunciado por una persona en una representación simbólica. En una aplicación de dictado automático, por ejemplo, esta representación aparecería en la pantalla como el texto escrito correspondiente al enunciado original.
El segundo componente, al que nos referiremos específicamente en este trabajo, realiza exactamente la operación inversa: transforma un texto escrito en su forma oral. Para ello utiliza las técnicas de síntesis del habla y, en concreto, la denominada *conversión de texto en habla* (CTH o TTS, *text-to-speech* en inglés).
Finalmente, la interacción entre la persona y la máquina se dirige desde un módulo denominado habitualmente gestor del diálogo, que controla la secuencia, en general de preguntas y respuestas, producidas por quien habla y por el ordenador.
Estas tres tecnologías (reconocimiento del habla, síntesis del habla y gestión del diálogo) se integran en los llamados sistemas de diálogo o sistemas conversacionales, mediante los cuales se puede obtener información o realizar transacciones mediante el habla con la ayuda de un sistema informático (fig. 1).

Las «máquinas parlantes»

La posibilidad de producir artificialmente habla similar a la humana ha despertado desde siempre el interés de los científicos. Una muestra de los esfuerzos realizados para conseguir este objetivo la encontramos en la «máquina parlante» del barón Wolfgang von Kempelen, descrita en 1791. La «máquina» intentaba reproducir fielmente la anatomía del aparato fonador humano, al igual que lo haría más tarde el sistema diseñado por sir Charles Wheatstone en 1835.
Más adelante, los componentes mecánicos se sustituyeron por los circuitos eléctricos, como en el Voder (*Voice Demonstrator*) de Homer Dudley, presentado en dos ferias mundiales en 1939. Un paso crucial se dio en los años sesenta, cuando el control de los ya denominados *sintetizadores de habla* empezó a realizarse desde un sistema informático completamente automático, abriendo así el camino a los productos actuales, cuyas primeras versiones se comercializaron en la década de los ochenta.
Hemos pasado, pues, de los fuelles, las lengüetas y las palancas a programas fácilmente accesibles a todos, que permiten que un ordenador lea en voz alta un texto cualquiera, cada vez de una forma más natural y en un creciente número de lenguas. En los apartados siguientes se intenta exponer sucintamente qué conocimientos y tecnologías son necesarios para convertir automáticamente un

texto escrito en su manifestación oral.

La producción de los sonidos del habla

Un modelo acústico de producción del habla
Un avance fundamental para llegar a los actuales sistemas de síntesis fue la posibilidad de modelar el proceso de producción de los sonidos del habla mediante un procedimiento relativamente simple, distinguiendo dos componentes principales: la fuente y el filtro.
En el habla humana, los sonidos se originan gracias al aire que sale de los pulmones y que llega al exterior a través de la laringe y del tracto vocal. En la laringe existen dos pliegues musculares (las cuerdas vocales) que al abrirse y cerrarse provocan la vibración de las moléculas del aire, emitiendo de este modo sonidos llamados *sonoros* como, por ejemplo, las vocales; en otras clases de sonidos, como en ciertas consonantes ([f] o [s], por ejemplo), el aire simplemente pasa entre las cuerdas vocales abiertas.
Este mecanismo permite dividir los sonidos en dos clases: los denominados *sonoros*, en los que vibran las cuerdas vocales –como se observa colocando las yemas de los dedos a la altura de la laringe y pronunciado una vocal alargada– y los *sordos*, en los que no se produce esta vibración (y que puede comprobarse por el mismo procedimiento, pronunciando una [s] larga).
La principal propiedad de los sonidos sordos es que acústicamente consisten en una onda sonora periódica, en la que se repite un patrón de vibración que corresponde al ritmo de abertura y cierre de las cuerdas vocales, mientras que en los sonidos sonoros la onda sonora resultante es aperiódica, sin una regularidad apreciable en su estructura acústica.
Para producir artificialmente estos dos tipos de sonidos basta con disponer de un mecanismo, conocido en síntesis como la *fuente*, capaz de generar ondas sonoras periódicas y ondas sonoras aperiódicas. De este modo puede imitar las características acústicas de los sonidos sordos y sonoros (fig. 2).
El siguiente elemento necesario para la síntesis es un procedimiento para reproducir la acción del tracto vocal (la faringe, la cavidad bucal y la cavidad nasal) en la producción de los sonidos. El tracto vocal puede considerarse como una cavidad de resonancia, similar a la de un instrumento musical. Además de su periodicidad, las ondas sonoras correspondientes a los sonidos sonoros tienen otra propiedad: poseen una frecuencia, denominada *fundamental*, que corresponde al tono agudo o grave con el que los percibimos y que está relacionada con el tamaño de las cuerdas vocales y la velocidad a la que se abren y cierran y una serie de armónicos, que son los responsables de que cada sonido se perciba con un timbre (claro, oscuro) diferente. Simplificando, se puede equiparar las cuerdas vocales con las cuerdas de un instrumento, y el tracto vocal con su caja de resonancia, de modo que una misma nota producida por un violín o por un contrabajo, por ejemplo, se percibe con un timbre diferente, lo que nos permite distinguir entre ambos instrumentos, aunque la frecuencia fundamental o tono (la *nota*, en términos musicales) sea la misma.
El tracto vocal no adopta la misma configuración en el momento de producir cada sonido; esto puede observarse, por ejemplo, comprobando que en una vocal como [i] la lengua se sitúa hacia la parte anterior de la cavidad bucal (al introducir un lápiz entre los labios se llega de inmediato a tocar la punta de la lengua), mientras que en una vocal como [u] la lengua se sitúa en la parte posterior (el mismo experimento permite observar que el lápiz puede introducirse mucho más entre los labios). Por otra parte, una vocal como [a] exige que la mandíbula inferior se desplace hacia abajo, abriendo bien la cavidad bucal; si bien se puede pronunciar con relativa facilidad [i] o [u] sosteniendo un cigarrillo entre los dientes, la misma operación es más difícil con la vocal [a]. Técnicamente, se habla en fonética de vocales anteriores en el caso de [i], posteriores en el caso de [u] y abiertas en el caso de [a], frente a [i] e [u], que se consideran cerradas (fig. 3).
Estos cambios en la posición de la lengua y la mandíbula hacen que la forma del tracto vocal sea diferente para cada sonido. En cierto modo, se puede equiparar a lo que sucedería con un instrumento que dispusiera de las mismas cuerdas y de diferentes cajas de resonancia, o de una caja que pudiera variar su tamaño y forma.
Desde el punto de vista acústico, este hecho tiene consecuencias importantes, pues la caja de resonancia del tracto vocal altera las propiedades de la onda sonora que se ha creado en las cuerdas vocales, modificando la amplitud de los armónicos, responsables del timbre con el que percibimos

cada sonido.
Esta acción del tracto vocal sobre los armónicos de la onda sonora que se crea en la laringe puede imitarse en la síntesis con lo que se denomina un filtro, es decir, un mecanismo que aumenta la amplitud de determinados armónicos y reduce la de otros, dotando así a cada sonido de una configuración acústica diferente (fig. 4).
Con ello tenemos ya los dos elementos básicos para imitar artificialmente el mecanismo de producción del habla: una fuente, con una acción análoga a la de las cuerdas vocales, y un filtro que hace las veces de tracto vocal. En 1960, Gunnar Fant formuló detalladamente este modelo (conocido como el *modelo de la fuente y el filtro* y que todavía se utiliza en la actualidad en los sistemas de conversión de texto a habla) y lo puso en práctica en el diseño de uno de los primeros sintetizadores modernos.

Las propiedades acústicas de los sonidos del habla
Sin embargo, aunque dispongamos de un medio de replicar la acción de la laringe y del tracto vocal, para llevar a cabo la síntesis necesitaremos también un cierto conocimiento sobre las propiedades de cada uno de los sonidos que forman el habla, tanto desde el punto de vista de su articulación como de sus características acústicas.
Asimismo, es preciso conocer cuáles de estas características acústicas contribuyen a distinguir perceptivamente los sonidos entre sí puesto que, en última instancia, en el proceso de comunicación entre personas se transmiten ondas sonoras que interpretamos como sonidos de la lengua.
Un avance muy significativo en este sentido se realizó a finales de los años cincuenta, con el *Pattern Playblack* (fig. 5) de los Laboratorios Haskins. En esencia, consistía en un sistema que permitía «dibujar» representaciones acústicas esquematizadas de los sonidos –en forma de espectrogramas, representaciones utilizadas habitualmente en fonética en las que se puede observar la frecuencia, la intensidad y la duración de los sonidos que componen un enunciado– y reproducirlas para observar los cambios en la percepción que se producían en función de las variaciones en los parámetros acústicos. De este modo se pudo determinar cuáles son las propiedades acústicas esenciales que distinguen diversas clases de sonidos, propiedades que se relacionan, en la síntesis, con la información que debe llegar a la fuente y al filtro para la producción de cada sonido.
Los estudios en fonética acústica y perceptiva, apoyados en el modelo de la fuente y el filtro y en herramientas como el espectrógrafo y el *Pattern Playback*, han permitido llegar a una caracterización de las propiedades de cada uno de los sonidos, información esencial si se desean reproducir de forma artificial.
Por ejemplo, una vocal puede caracterizarse acústicamente en términos de su frecuencia fundamental –responsable del tono o altura tonal con que la percibimos–, de su duración y de la frecuencia e intensidad de sus componentes principales. Estos componentes se denominan formantes y se identifican como F1 (primer formante), F2 (segundo formante), etcétera, y son responsables del timbre, propiedad que nos permite diferenciar la percepción, pongamos por caso, de una [a] o de una [i].
Así pues, la síntesis de una vocal requeriría disponer de información sobre su frecuencia fundamental y la frecuencia de sus formantes. El primer aspecto se relaciona con la fuente, mientras que el segundo depende del filtro, ya que los formantes no son más que armónicos cuya amplitud se realza en función de la forma que adopte el tracto vocal, creando diferentes cajas de resonancia para cada sonido, tal como veíamos en el ejemplo anterior con [a], [i] y [u].

La producción de enunciados

Cabría pensar que para que funcione un sistema de síntesis del habla basta con disponer de la información sobre los parámetros acústicos necesarios para que la acción conjunta de una fuente y un filtro proporcione cada uno de los sonidos del enunciado a reproducir.
Se podría imaginar la posibilidad de que un locutor grabara cada uno de los sonidos de la lengua en la que deseamos crear enunciados sintetizados; a continuación, se analizarían las propiedades acústicas de cada sonido, y se guardarían esas propiedades en una tabla. En el momento de la síntesis, dicha tabla proporcionaría a la fuente y el filtro del sintetizador la información necesaria.
Sin embargo, con el conocimiento de las propiedades de cada sonido no se puede lograr todavía que

un sintetizador genere enunciados completos. Para ello faltan como mínimo dos aspectos esenciales: en primer lugar, la información sobre el modo como los sonidos del habla se enlazan unos con otros o, expresado en terminología de los sistemas de síntesis, se concatenan; y en segundo lugar, la información sobre las propiedades más globales del enunciado, agrupadas generalmente bajo la denominación de prosodia o elementos suprasegmentales. A continuación haremos un breve repaso a ambos aspectos.

Las unidades de la síntesis

En cualquier representación acústica del habla puede observarse que los sonidos no se producen de manera aislada, sino que se encadenan unos con otros, solapándose muchas veces las propiedades de un sonido con las de otro. Este fenómeno se conoce en fonética como coarticulación y se relaciona con la velocidad y la coordinación de los movimientos del tracto vocal. Es habitual que, por ejemplo, en una vocal producida entre dos consonantes de las denominadas nasales (en las que como en [m] el aire sale también por la cavidad nasal), se aprecie una cierta nasalización como resultado de la influencia de las consonantes adyacentes.

De forma análoga, en el paso de una consonante oclusiva como [p], [t] o [k] –articulada con una obstrucción de la salida del aire seguida de una liberación del aire acumulado durante el cierre del tracto vocal– a una vocal, pueden observarse en un espectrograma los cambios de configuración de la cavidad bucal en el cambio de un sonido a otro, en forma de las llamadas transiciones de los formantes.

Conseguir habla sintetizada de calidad a base de concatenar sonidos aislados e intentando imitar el resultado acústico de los movimientos del tracto vocal en las transiciones entre sonidos es una operación muy difícil. Por ello, las unidades a partir de las cuales se construye un sistema de síntesis no suelen ser sonidos aislados, sino combinaciones de sonidos.

Son habituales los denominados *difonemas*, que consisten en una combinación entre la mitad del primer sonido y la mitad del segundo, o las semisílabas, formadas por el primer sonido completo y la mitad del segundo. Con ello se persigue que al concatenar las unidades la unión se produzca por las partes en las que existe una menor variación acústica (típicamente en el «centro» de un sonido) y no por aquellas en las que se encuentra la transición de un sonido a otro. Para sintetizar la palabra *casa* mediante difonemas se recurriría a juntar [ka] con [as] y [as] con [sa], de modo que la unión de realizaría entre dos mitades de [a] y entre dos mitades de [s], zonas en las que la cavidad bucal se mantiene en una posición relativamente estable en comparación con el momento de cambio de [k] a [a] o de [s] a [a].

En un sistema de síntesis real, se dispondrá pues de la grabación de todos los difonemas existentes en la lengua para la que se desarrolla el sistema (el conjunto que constituye el diccionario de unidades de síntesis), y, para cada una de ellos se guardará debidamente codificada la información correspondiente a las propiedades acústicas que necesitan la fuente y el filtro del sintetizador para su reproducción.

La prosodia

Sin embargo, aún concatenando el conjunto de difonemas necesario para producir un enunciado, es preciso abordar otro aspecto fundamental, la prosodia, para que el enunciado se aproxime al máximo al habla natural. Los elementos que constituyen los rasgos prosódicos de un enunciado son, esencialmente, la duración, la intensidad y la melodía.

Por una parte, la duración de cada uno de los sonidos no es siempre la misma, pues se ve influida por diversos factores, como la velocidad a la que se habla, los sonidos anteriores o posteriores, el hecho de que aparezcan en una sílaba acentuada o no, la realización de una pausa, etc. Por ello, los sistemas de síntesis disponen de reglas que permiten alterar la duración de cada uno de los sonidos originalmente recogido en el diccionario de unidades para adaptarla en propiedad al habla natural en un determinado contexto o en un determinado tipo de frase.

Algo similar puede decirse respecto a la intensidad, responsable de que un sonido se perciba con mayor fuerza que otro. Puede observarse fácilmente que un sonido no posee la misma intensidad al principio de un enunciado que al final, por lo que es preciso realizar algún ajuste si se desea una síntesis natural.

En tercer lugar, los enunciados que producimos poseen una melodía específica. La melodía es

responsable de la variación en la frecuencia de vibración de las cuerdas vocales o, en términos acústicos, la variación a lo largo del tiempo de la frecuencia fundamental, que puede ser importante, por ejemplo, para distinguir *Ha llegado* de *¿Ha llegado?* y de *¡Ha llegado!* (fig. 6).
En la síntesis tiene que ser posible «imitar» los cambios melódicos que dotan de diferente significado a los enunciados. Para ello es preciso aplicar lo que se denomina un *patrón melódico*, obtenido a partir del análisis de la entonación en el habla natural, estudiada a partir de representaciones acústicas conocidas como curvas melódicas.
El patrón se aplica a cada una de las oraciones, reproduciendo los cambios en la melodía que se darían en ese mismo enunciado pronunciado por un hablante humano. Ésta es uno de las áreas a los que actualmente se dedican más esfuerzos en el campo de la síntesis. La melodía de un enunciado no sólo se relaciona con su modalidad (enunciativa, interrogativa y exclamativa en el caso de los ejemplos anteriores), sino también con otros aspectos de su significado, como la intención del hablante de comunicar determinados matices o determinados estados de ánimo.

El tratamiento del texto

Una vez que se dispone de los elementos necesarios debería ser posible abordar la producción de mensajes verbales por parte de un ordenador. Simplificando mucho, estos elementos son un diccionario de unidades de síntesis con la información acústica correspondiente a cada unidad debidamente recogida; un conjunto de reglas prosódicas que asignen correctamente la duración e intensidad de cada sonido y que proporcionen la melodía adecuada a todo un enunciado; y un programa que haga las funciones de fuente y filtro de un modo análogo al aparato fonador humano.

La transcripción fonética automática
Sin embargo, falta un elemento importante: el mensaje que se desea sintetizar. En la mayoría de las aplicaciones el mensaje se presenta en forma de un texto en soporte digital (un documento elaborado con un procesador de textos, una página web, un mensaje de correo electrónico) y debe convertirse en su equivalente sonoro.
Puesto que las unidades almacenadas en el diccionario de síntesis son de naturaleza fonética es preciso transformar de algún modo el texto, que estará escrito en la ortografía convencional de cada lengua, en una representación más cercana a los sonidos incluidos en el diccionario.
Para ello se utilizan programas que realizan lo que se conoce como una transcripción fonética automática. Es decir, que convierten las grafías del texto en símbolos fonéticos correspondientes a los sonidos guardados en el diccionario de unidades. En el caso de lenguas como el castellano o el catalán, esta transcripción se realiza mediante un conjunto de reglas que asocian cada grafía al modo como se pronuncia en un determinado contexto: se establece, por ejemplo, que la grafía *g* se pronuncia como el sonido que se transcribe [x] (el sonido inicial de *jamón*) ante [e, i] pero se pronuncia [g] (el sonido inicial de *gato*) cuando va seguida de la grafía *u*. De este modo, a la hora de realizar la síntesis, se elige correctamente el difonema [xe] cuando aparece la grafía *je* o el difonema [ge] cuando aparece la grafía *gue*.

La preparación del texto
Si se observa un texto real, como el que publica un periódico, por ejemplo, se verá que contiene fenómenos como abreviaturas (Sr.), siglas (PSC), números, fechas y símbolos como el dólar o el euro, entre otros muchos. Todos ellos deben ser «deletreados», por así decirlo, de modo que después se pueda realizar la correspondiente transcripción fonética. Esto se consigue con la creación de diccionarios en los que se especifica cuál es la forma escrita completa de cada elemento: por ejemplo, *Sr.* se asociaría a *señor* y *$* a *dólar*.
Aunque en principio pueda parecer un proceso trivial esconde ciertos problemas. Por ejemplo, en catalán *2* debe asociarse a *dos* para el masculino y a *dues* para el femenino, y el sistema de transcripción debe elegir la forma correcta utilizando información sobre el género del nombre que sigue.
También es preciso tomar decisiones sobre la pronunciación de nombres extranjeros, adoptando una versión más o menos cercana a la nativa según las tradiciones de cada lengua, con el problema añadido de que los sonidos que no formen parte de la lengua no se encontrarán en el diccionario de

unidades de síntesis si no se ha previsto anteriormente (un sistema de síntesis del catalán, por mencionar un caso concreto, debería incluir el sonido [x] no existente en esta lengua para pronunciar los nombres castellanos que lo contienen). Estas cuestiones y otras de naturaleza similar se abordan en la fase que suele denominarse procesamiento previo del texto.

La información lingüística
Otras informaciones presentes en el texto se utilizan en fases posteriores de la síntesis. Los signos de puntuación, por ejemplo, se emplean para asignar las pausas (distinguiendo enunciados como «los sistemas de síntesis, que son eficientes, permiten leer un texto» de «los sistemas de síntesis que son eficientes permiten leer un texto») o, en el caso de la interrogación y la admiración, para determinar el patrón melódico del enunciado.
En el caso de las pausas, cabe considerar que un buen lector introduce en ocasiones silencios sin que aparezca un signo de puntuación, basándose en su interpretación del texto. Lo mismo puede decirse de la melodía, puesto que leer un texto con las inflexiones melódicas adecuadas requiere necesariamente su comprensión. Por ello se está trabajando para integrar en los conversores de texto en habla analizadores morfológicos que determinen la parte de la oración a la que pertenece cada palabra (de modo que por ejemplo no aparezca una pausa entre un artículo y un nombre); sintácticos, para mejorar la entonación con la que se sintetiza el texto; y a más largo plazo, analizadores semánticos y pragmáticos que ayuden a determinar aspectos como el foco o la intención comunicativa del enunciado.

La estructura del sistema

A pesar de que hemos simplificado mucho su descripción, puede apreciarse que la transformación automática de un texto escrito a su manifestación sonora (denominada conversión de texto en habla, como se indicaba al principio) es un proceso complejo.
El texto debe ser preprocesado para realizar su transcripción fonética sin dificultades y transcrito para localizar las unidades fonéticas en el diccionario de síntesis.
Estas unidades fonéticas deben concatenarse y modificarse a fin de adaptar la duración y la intensidad de los sonidos a cada enunciado, y debe aplicarse también el patrón melódico adecuado al significado y a la forma del enunciado.
Por último, esta información debe convertirse en un conjunto de parámetros acústicos, encargados de que la fuente y el filtro del sintetizador produzcan la onda sonora que llegará el receptor del enunciado y que, de forma ideal, debería parecerse todo lo posible a la lectura que realiza un hablante humano (fig. 7).
Una tarea de tal naturaleza es de difícil abordaje si no se lleva a cabo por un equipo interdisciplinario, que incorpore a especialistas en fonética, en análisis lingüístico, en tratamiento de la señal sonora y en informática. Como en otras áreas de las tecnologías lingüísticas, la conjunción de disciplinas es indispensable para conseguir que un ordenador realice de forma automática algo aparentemente tan simple para una persona como leer un texto en voz alta.

Bibliografía

La producción del habla
DENES, P.B.; PINSON, E.N.: T*he Speech Chain; the Physics and Biology of Spoken Language*, Garden City, Nueva York, Anchor Press / Doubleday (Anchor Science Study Series), 1963. [Existe una 2ª edición en: Nueva York, W.H. Freeman, 1993.]
GIL FERNÁNDEZ, J.: *Los sonidos del lenguaje*, Madrid, Síntesis, 1988 (Textos de apoyo, Lingüística 3, 1993).
LADEFOGED, P.: *Elements of Acoustic Phonetics* (2ª ed.), Chicago – Londres, University of Chicago Press, 1996.
LLISTERRI BOIX, J.: «Los sonidos del habla», en: MARTÍN VIDE, C. (ed.), *Elementos de lingüística*, Barcelona, Octaedro (Octaedro Universidad, Textos), 1996: 67-128.

Presentaciones generales de la síntesis del habla
CARLSON, R., GRANSTRÖM, B.: «Speech Synthesis», en: HARDCASTLE, W.J.; LAVER, J. (eds.), *The Handbook of Phonetic Sciences*, Oxford, Blackwell Publishers (Blackwell Handbooks in Linguistics, 5), 1997: 768-788.
FLANAGAN, J.L.: «The synthesis of Speech», *Scientific American* 1972; 226, 2: 45-58.

JAVKIN, H.R.: «Speech analysis and synthesis», en: LASS, N.J (ed.), *Principles of Experimental Phonetics*, St Louis, Mosby, 1996: 245-276.

Presentaciones generales de la conversión de texto a habla

OLIVE, J.P.: «The Talking Computer': Text to Speech Synthesis», en STORK, D.G. (ed.), *Hal's Legacy: 2001's Computer as Dream and Reality*, Cambridge, Mass., The MIT Press, 1998 [http://mitpress.mit.edu/e-books/Hal/chap6/six1.html].
RODRÍGUEZ CRESPO, M.A.: «Introducción a la conversión texto-voz», *Philologia Hispalensis* 1997; 11 (2): 177-192.

Conversión de texto a habla en castellano y en catalán

BONAFONTE, A.; ESQUERRA, I.; FEBRER, A.; FONOLLOSA, J.A.; VALLVERDÚ, F.: «The UPC Text-to-Speech System for Spanish and Catalan», en: *Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP'98*, Sydney, Australia, 30th November-4th December 1998 [http://gps-tsc.upc.es/veu/research/pubs/download/Bon98c.pdf], 1999.
CASTEJÓN LAPEYRA, F.; ESCALADA SARDINA, G.; MONZÓN SERRANO, L.; RODRÍGUEZ CRESPO, M.A.; SANZ VELASCO, P.: «Un conversor texto-voz para el español», *Comunicaciones de Telefónica I+D* 1994; 5 (2): 114-131. http://www.tid.es/presencia/publicaciones/comsid/esp/articulos/vol52/artic8/8.html
MONTERO, J.M.; GUTIÉRREZ ARRIOLA, J.; COLÁS, J.; MACÍAS GUARASA, J.; ENRÍQUEZ, E.; PARDO, J.M.: «Development of an emotional speech synthesiser in Spanish», en: *Eurospeech99, 6th European Conference on Speech Communication and Technology*, September 5-9, 1999, Budapest, Hungría, pp. 2099-2102 [http://www-gth.die.upm.es/~macias/doc/pubs/eurosp99/submitted/m058.pdf], 1999.

Trabajos especializados

ALLEN, J.; HUNNICUTT, M.S.; KLATT, D.H. (con R.C. ARMSTRONG y D. PISONI): *From Text to Speech: The MITalk System*, Cambridge, Cambridge University Press (Cambridge Studies in Speech Science and Communication), 1987.
BAILLY, G.; BENOÎT, C. (eds.): *Talking Machines. Theories, Models and Designs*, Amsterdam, North-Holland / Elsevier Science Publishers, 1992.
COLE, R.: «Spoken Output Technologies», en: COLE, R.A.; MARIANI, J.; USZKOREIT, H.; ZAENEN, A.; ZUE, V. (eds.), *Survey of the State of the Art in Human Language Technology*, Cambridge, Cambridge University Press [http://cslu.cse.ogi.edu/HLTsurvey/ch5node2.html#Chapter5], 1997.
DUTOIT, T.: *An Introduction to Text-to-Speech Synthesis*, Dordrecht, Kluwer Academic Publishers (Text, Speech and Language Technology, 3), 1997.
LLISTERRI, J.; AGUILAR, L.; GARRIDO, J.M.; MACHUCA, M.J.; MARÍN, R.; DE LA MOTA, C.; RÍOS, A.: «Fonética y tecnologías del habla», en: BLECUA, J.M.; CLAVERÍA, G.; SÁNCHEZ, C.; TORRUELLA, J. (eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos*, Barcelona, Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona, Editorial Milenio, 1999: 449-479.
SPROAT, R. (ed.): *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Dordrecht, Kluwer Academic Publishers, 1997.
VAN SANTEN, J.H.P.; SPROAT, R.; OLIVE, J.; HIRSCHBERG, J. (eds.): *Progress in Speech Synthesis*, Nueva York, Springer-Verlag, 1996.

Desarrollo histórico de la síntesis del habla

DUDLEY, H.; TARNOCZY, T.H.: «The Speaking Machine of Wolfgang von Kempelen», *Journal of the Acoustical Society of America* 1950; 22 (2): 151-166.
KLATT, D.H.: «Review of Tex-to-Speech Conversion for English», J*ournal of the Acoustical Society of America* 1987; 82 (3): 737-793; en: ATAL, B.S.; MILLER, L.J.; KENT, R.D. (eds.), *Papers in Speech Communication: Speech Processing*, Nueva York, Acoustical Society of America, 1991: 57-114. [Ejemplos sonoros en: http://www.icsi.berkeley.edu/eecs225d/klatt.html; http://www-uilots.let.uu.nl/~audiufon/data/e_milestones_in_text-to-speech_conversion.html]
LIENARD, J.S.: «La synthèse de la parole. Historique et réalisations actuelles», *Revue d'Acoustique* 1970 ; 3 (11): 204-213.
LIENARD, J.S.: «From speaking machines to speech synthesis», en *Actes du XIIème Congrès International des Sciences Phonétiques*. 19-24 août 1991, Aix-en-Provence, France. Aix-en-Provence, Université de Provence, Service des Publications, 1991 ; vol. 1: 18-27.
METTAS, O.: «Aperçu historique sur les appareils de synthèse de la parole», *Travaux de Linguistique et Littérature* 1965 ; 3 (1): 185-200.

Joaquim Llisterri

Profesor titular de lingüística general en la Universitat Autònoma de Barcelona. Su investigación se ha centrado en la aplicación de la fonética experimental a las tecnologías del habla, especialmente en el campo de la síntesis, y a la caracterización de la interferencia fonética en la adquisición de segundas y terceras lenguas. También se ha ocupado de temas relacionados con la evaluación del habla sintetizada y la constitución y anotación de bases de datos y corpus orales, así como del uso de nuevas tecnologías en la enseñanza de lenguas y la investigación en lingüística.

Joaquim.Llisterri@uab.es
http://liceu.uab.es/~joaquim/home.html

Figura 1 El reconocimiento del habla convierte una señal sonora en su representación escrita, mientras que la conversión de texto en habla transforma un texto escrito en su equivalente hablado

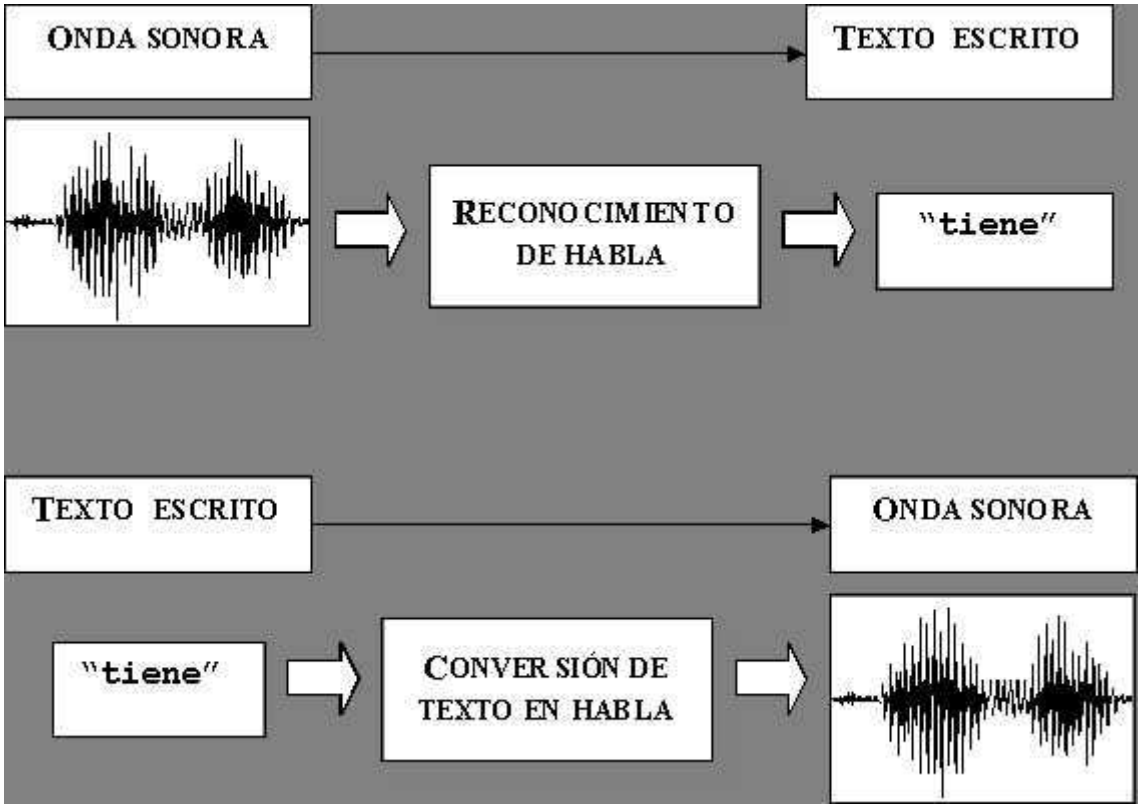


Figura 2 La fuente periódica de un sintetizador modela la producción de los sonidos sonoros (producidos con vibración de las cuerdas vocales) y la fuente aperiódica la de los sonidos sordos (producidos con las cuerdas vocales abiertas)

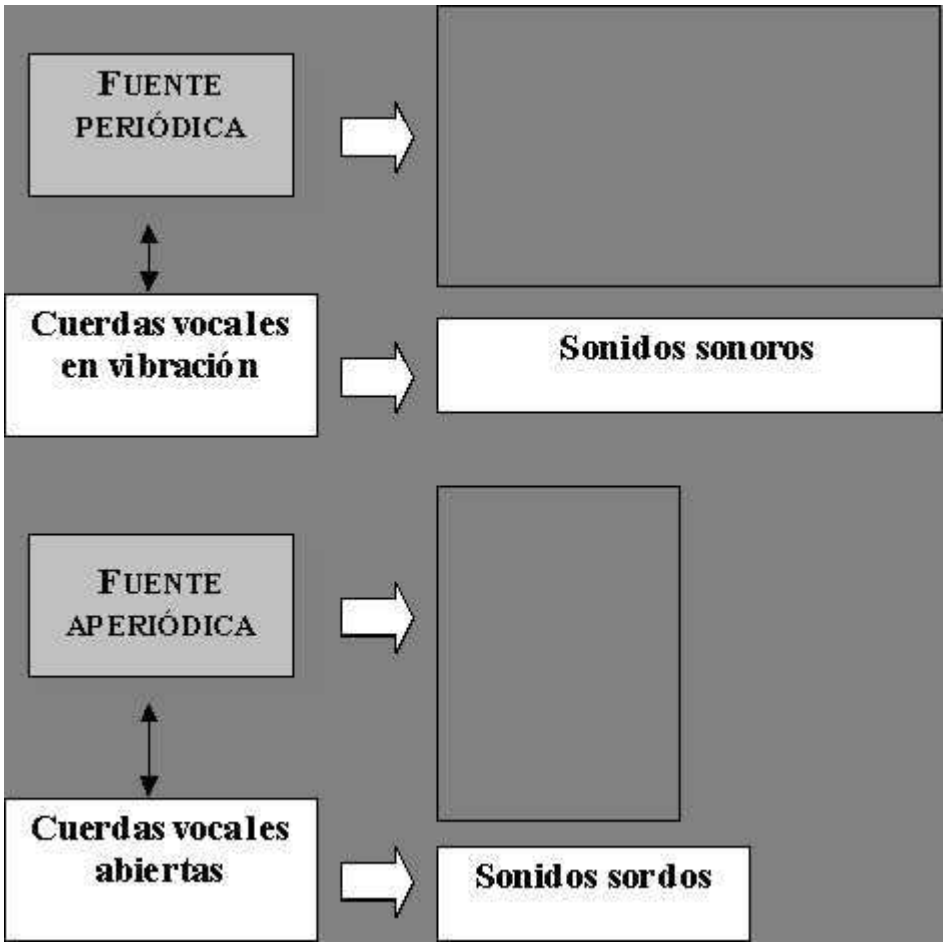


Figura 3 Representación esquematizada de la articulación de las vocales [i] (anterior cerrada), [u] (posterior cerrada) y [a] (media abierta) y de las modificaciones que la resonancia en el tracto vocal produce en los armónicos de la onda sonora creada en las cuerdas vocales. En la representación de la onda sonora el eje vertical indica la amplitud y el horizontal la frecuencia de cada uno de los armónicos. F1 y F2 señalan el primer y el segundo formante

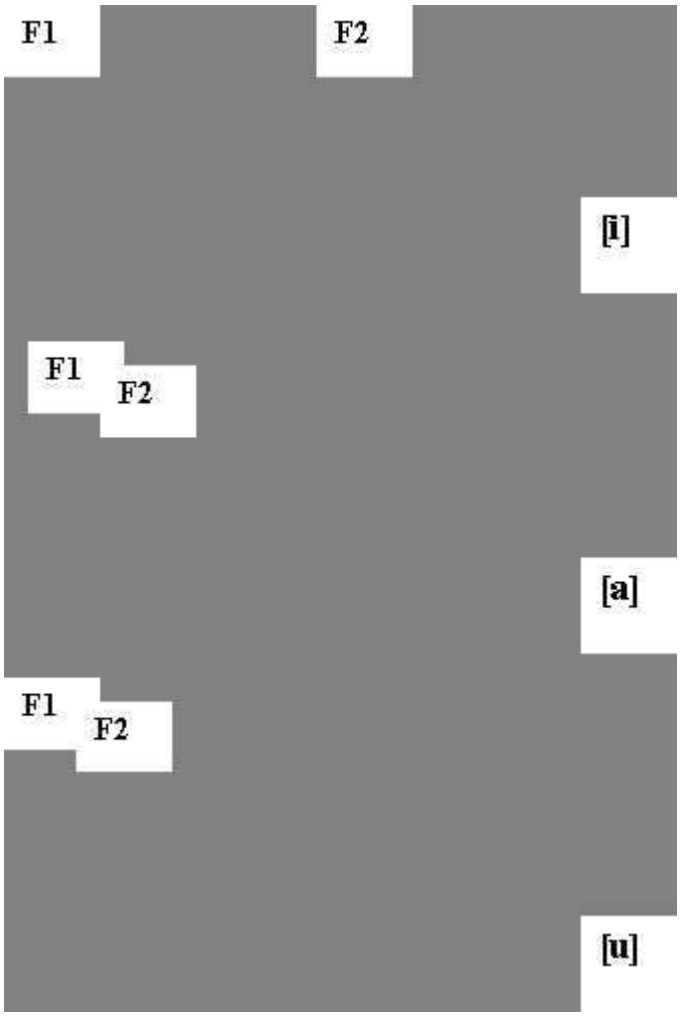


Figura 4 Onda sonora producida por la vibración de las cuerdas vocales (fuente) y resultado de la resonancia producida en el tracto vocal (filtro): la amplitud (en el eje vertical) de determinados armónicos se ve aumentada, mientras que la de otros se ve reducida, dotando así de un timbre diferente a cada sonido del habla

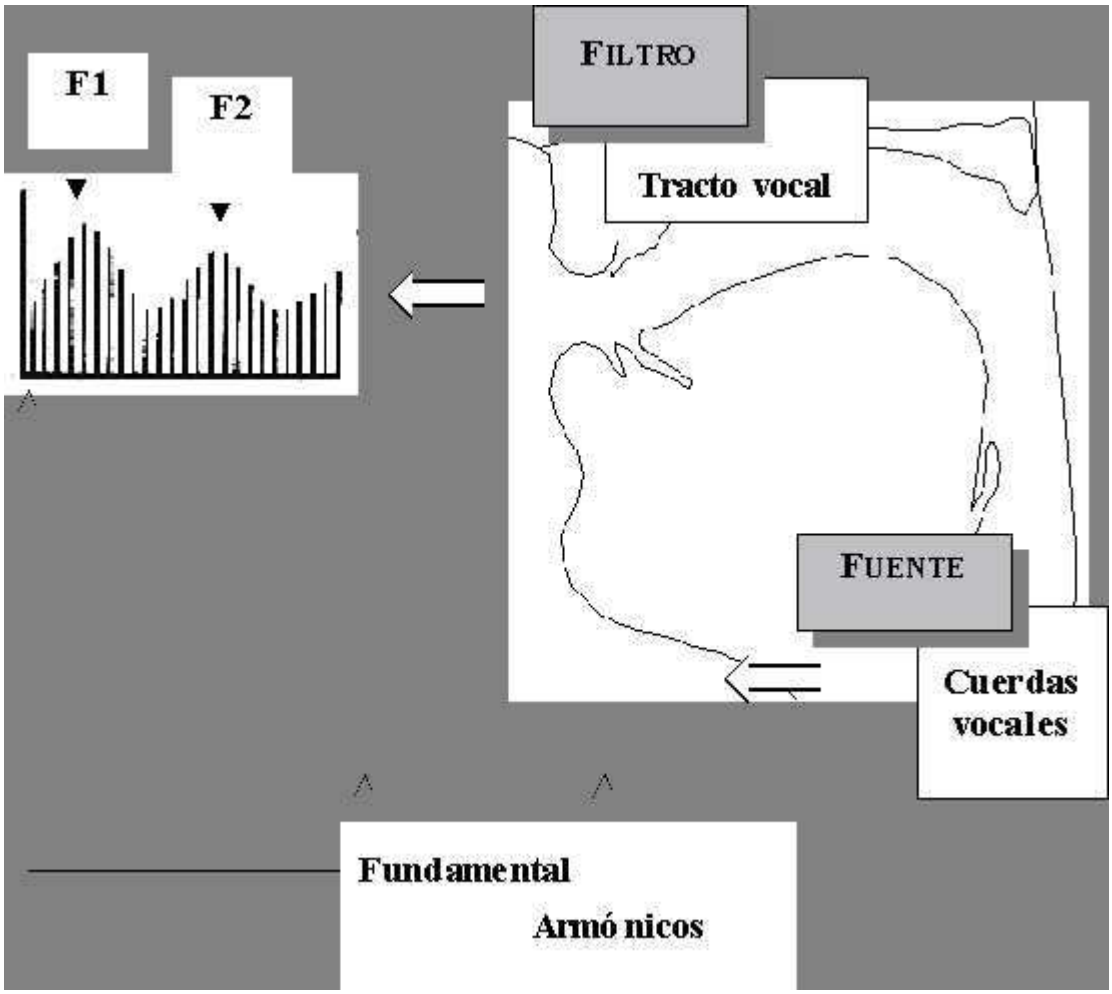
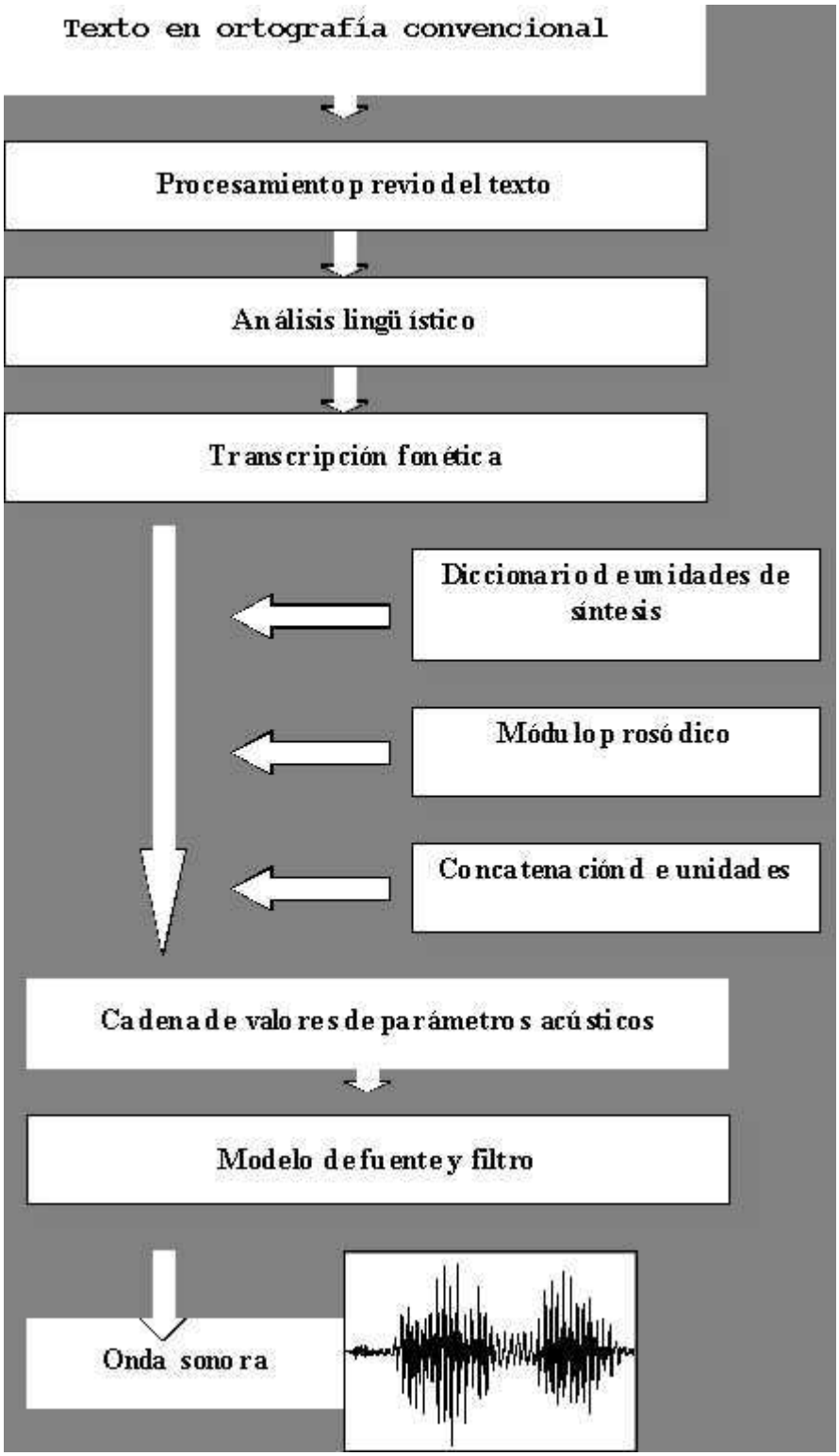


Figura 5 Esquema simplificado del proceso de conversión de texto en habla


Texto en ortografía convencional
Procesamiento previo del texto
Análisis lingüístico
Transcripción fonética
Diccionario de unidades de síntesis
Módulo prosódico
Concatenación de unidades
Cadena de valores de parámetros acústicos
Modelo de fuente y filtro
Onda sonora