SERVICIOS DE VOZ PARA LA GESTIÓN DE LA INFORMACIÓN

Carmen Díez Carrera* y Gregorio Escalada Sardina**

Resumen: Los servicios de voz constituyen uno de los medios más novedosos de acceso a los sistemas de información, basados en el tratamiento informático de la misma para conseguir sistemas de reconocimiento y síntesis de voz. El presente artículo la analiza en sus diversos aspectos —fisiológico, físico y lingüístico—, su tratamiento informático y concluye con un muestrario de aplicaciones.

Palabras claves: Tecnología del habla. Tratamiento automático de la lengua natural. Lingüística y Documentación. Industrias de la lengua.

Abstract: Voice facilities are one of the most innovative means of access to information systems, based on computer processing of it, in order to achieve systems of speech recognition and synthesis. This article analyses it in its physiological, physical and linguistic aspects as well as computer processing. It finishes with a collection of application examples.

Palabras clave: Voices Services for Information Management.

Introducción

La búsqueda de una comunicación entre el hombre y la máquina ha dado lugar a diversas soluciones, así en el hardware (del teclado al ratón, o a las tablas digitalizadoras...) como en el software (windows, gráficos...). Algunas investigaciones, ubicadas en el ámbito de la tecnología del habla y del procesamiento automático de la lengua humana, se orientan hacia el empleo de la lengua oral y escrita, tal como la usamos diariamente. Estos trabajos, en principio limitados al tratamiento del lenguaje escrito, se ocupan de dotar a la máquina de la capacidad de tratar y entender el lenguaje humano, en vez de imponer a los seres humanos el lenguaje de la máquina.

Recientemente, con los avances en la capacidad de proceso de los ordenadores y en las técnicas de procesado de voz, van apareciendo tímidamente nuevos servicios y productos que abren otra vía de comunicación para acceder a la información con unos comandos y solicitudes propios del lenguaje de las personas. Las técnicas empleadas en estos desarrollos van desde las más simples, aquéllas en las que la información se graba en un disco, hasta las más sofisticadas, que aúnan el uso de la voz como soporte de la comunicación con todos los conocimientos orientados a que las máquinas simulen la capacidad humana del lenguaje.

1 La voz

La *voz* es definida por el Diccionario de la Real Academia Española como «sonido que el aire expelido por los pulmones produce al salir por la laringe, haciendo que vi-

* Universidad Carlos III de Madrid. Departamento de Biblioteconomía y Documentación. e-mail: cdcil.@bib.uc3m.es

** Telefónica Investigación y Desarrollo. División de Servicios de Tratamiento del Habla.

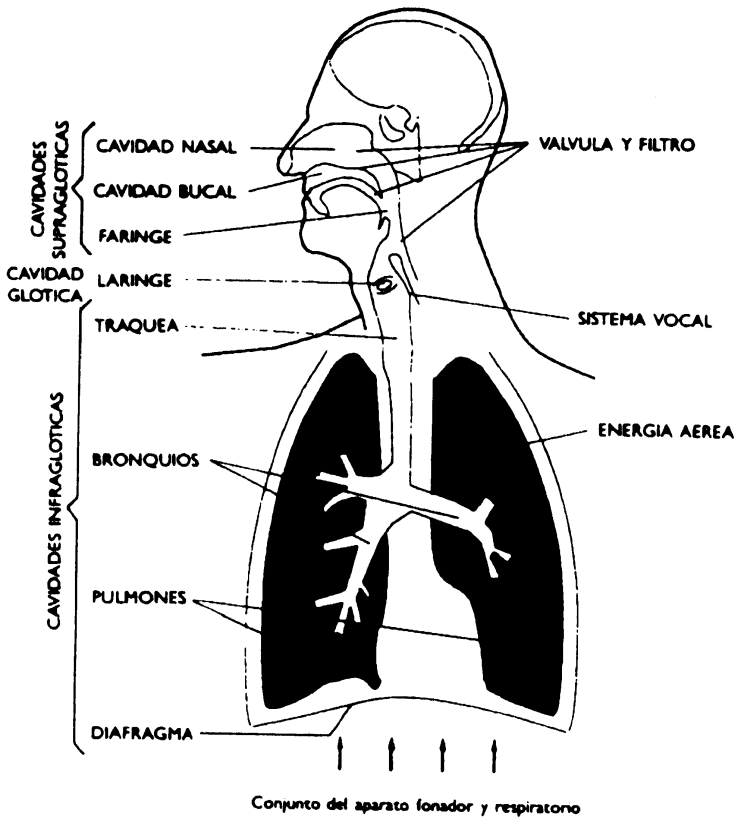
Recibido: 14-5-96.

bren las cuerdas vocales». Así se pueden estudiar sus aspectos fisiológico, físico y lingüístico, cuando se integra en el sistema articulado del lenguaje.

1.1 Aspecto fisiológico

La mayor parte de los sonidos que constituyen la voz tienen su origen en una corriente de aire, procedente de los pulmones, y modulada por los órganos de la *laringe* y el *tracto vocal.*

Figura 1
Esquema del aparato fonador humano



Conjunto del aparato fonador y respiratorio

En la parte superior de la *laringe* hay dos membranas, llamadas cuerdas vocales, que se oponen a manera de labios. La abertura que dejan entre sí es la glotis, por ella entra y sale el aire inspirado y espirado: cuando respiramos sin voz, la glotis está abierta; cuando emitimos voz, las cuerdas vocales se juntan por contracción de los músculos insertos en los cartílagos de la laringe, y la glotis se cierra. La presión del aire espirado aumenta y abre la glotis; tras caer la presión del aire, la glotis se cierra de nuevo. De esta manera vibran las cuerdas vocales, generando sonidos sonoros (p. e. las vocales). Para los sonidos sordos, las cuerdas vocales no vibran, y el origen del sonido es una turbulencia de aire producida en algún punto del *tracto vocal* y es en esta cavidad donde los sonidos adquieren muchas de sus características diferenciadoras (1):

— Según la posición elevada o caída del velo del paladar, los sonidos resultan respectivamente orales o nasales en función de si el aire sale por la boca o por la nariz; también puede salir simultáneamente por ambas cavidades y en este caso resultan oronasales.
— Según el punto donde se produzca la obstrucción del tracto vocal, o en función de los órganos bucales que intervienen se habla de sonidos bilabiales, labiodentales, linguodentales, linguointerdentales, linguoalveolares, linguopalatales y linguovelares.
— Según el tipo de obstrucción, las consonantes se clasifican en oclusivas, fricativas, africadas, laterales y vibrantes.

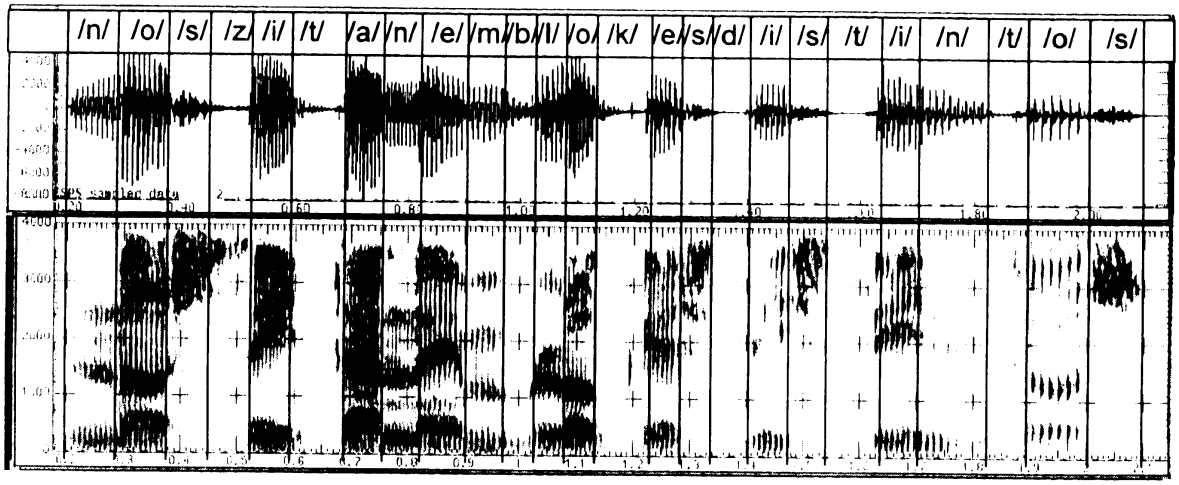
Asimismo, las vocales se clasifican en abiertas, semiabiertas y cerradas, y en anteriores o palatales, centrales y posteriores o velares, según la disposición de la lengua.

1.2 Aspecto físico

Tanto la formación como la propagación de la onda sonora pertenecen a la Física. La voz es una señal analógica, una onda continua que consiste en condensaciones y rarefacciones del aire. Por tanto, se pueden estudiar las distintas características físicas de la señal de voz, empleando para ello distintos modelos y formas de representación.

En el plano acústico, la representación más sencilla es la evolución en el tiempo de la amplitud de la señal. Este tipo de representación que corresponde a un espectrograma puede verse en la parte superior de la figura 2: en el eje horizontal se representa el tiempo (en este caso un par de segundos) y en el eje vertical el contenido de la frecuencia, la amplitud de la señal después de ser recogida por un micrófono y transformada en una señal eléctrica.

Figura 2
Forma de onda, espectrograma y transcripción fonética de la frase: «Nos citan en bloques distintos», pronunciada por un locutor masculino

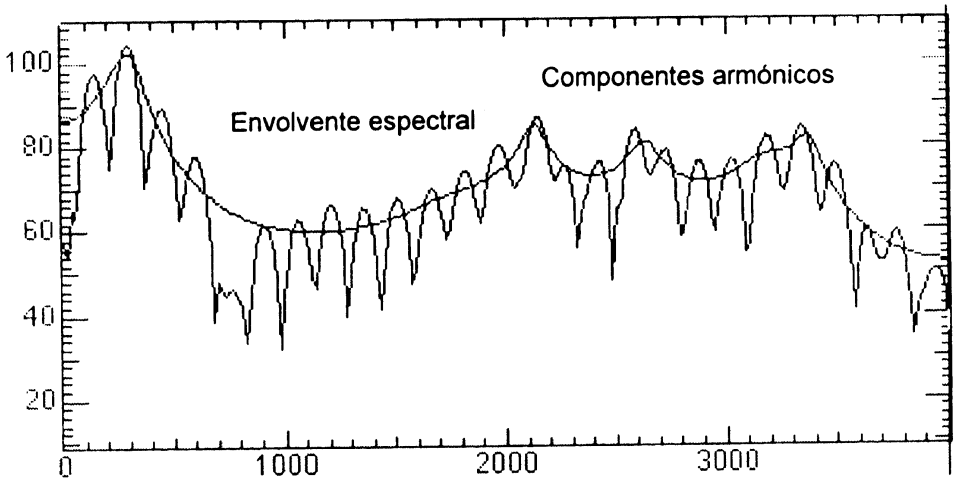


La cantidad de energía presente en un instante y para una frecuencia determinada se refleja en la densidad del tono gris que, a mayor energía, es más negro. La frecuencia es el número de veces que se repite una onda elemental o sonido puro por unidad de tiempo. Se corresponde con las notas de la escala musical. Un tono o nota grave tiene una frecuencia baja, mientras que un tono agudo tiene una frecuencia alta (vibra muchas veces por segundo). Cualquier sonido se puede descomponer en un conjunto de tonos puros de distintas frecuencias, y entonces se habla de sonidos graves o agudos según lo sean los tonos puros que los componen.

Otra representación consiste en elegir un instante de tiempo, y estudiar la distribución en frecuencias de los tonos que lo componen. La figura 3 corresponde a una representación de este estilo. En el eje horizontal se representa la frecuencia, y en el eje vertical la energía. Este espectro corresponde a un sonido vocálico (una /i/), y por tanto sonoro. La estructura periódica se aprecia en la presencia de tonos (armónicos), equiespaciados en frecuencia. La frecuencia de separación es la frecuencia a la que vibran las cuerdas vocales al generar el sonido, y se corresponde con el tono del mismo, como se verá más adelante.

Sin embargo, si lo que nos interesa es más la identidad del sonido que su tono, resulta más útil la representación que proporciona la envolvente espectral. Esta representación hace abstracción de los armónicos, y retiene las características globales. Estas características recogen el efecto de la transformación realizada por el tracto bucal sobre la corriente de aire procedente de los pulmones y, como se verá más adelante, se relaciona con el timbre del sonido y las características distintivas de los sonidos.

Figura 3
Espectro de un segmento de /i/ pronunciada por un locutor masculino. También se representa la envolvente espectral.



Las *características acústicas* del sonido son:

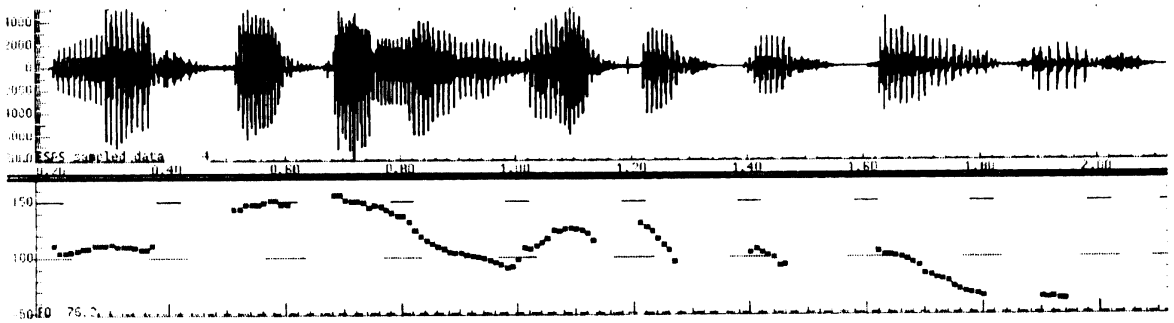
• **Tono**: configura la línea melódica de la voz, la entonación. Depende de la frecuencia de vibración de las cuerdas vocales, y en virtud de este parámetro los sonidos son graves o agudos. El tono se mide en herzios (Hz), número de veces que se repite la

unidad de muestra o ciclo de una onda en un segundo (una señal de 10 Hz produce 10 ciclos en un segundo). Corresponde a la altura de las notas musicales.

El tono depende de las características físicas del hablante, pues la frecuencia de vibración de las cuerdas vocales depende de la masa y grosor de las mismas. Generalmente, la masa de las cuerdas vocales de las mujeres es menor que la de los hombres, por lo que el tono de voz de las mujeres suele ser más agudo, al vibrar las cuerdas vocales con mayor frecuencia. El registro promedio de los hombres (el tono habitual) suele rondar los 100 Hz, mientras que el de las mujeres se aproxima a los 200 Hz.

Además, el hablante puede controlar la tensión de los músculos de las cuerdas vocales y por tanto la frecuencia de vibración de las mismas. Así el hablante modifica a voluntad el tono de su voz, y puede añadir información o reforzar el mensaje que está transmitiendo mediante la configuración de la curva melódica o entonación, que no es más que la evolución en el tiempo del tono. En la figura 4 se presenta de nuevo la forma de onda, junto con la evolución en el tiempo del tono (para los sonidos sonoros). El tono marca el acento de las palabras, la característica enunciativa o interrogativa de las frases, y cualquier énfasis que el hablante quiera dar al mensaje.

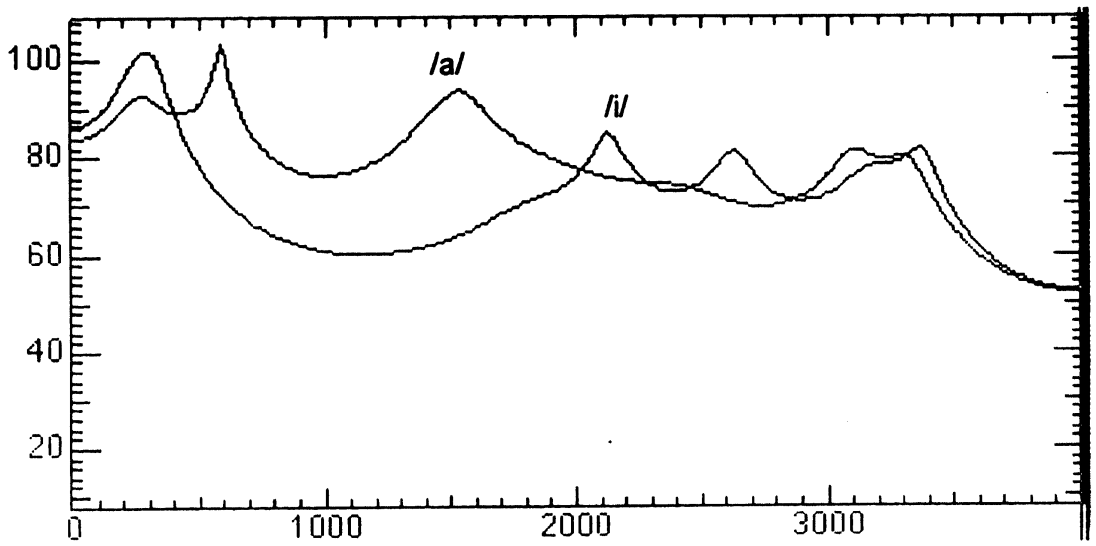
Figura 4
Forma de onda y curva melódica de la frase: «Nos citan en bloques distintos», pronunciada por un locutor masculino



• **Timbre:** «conformación que depende del volumen y abertura de las cavidades de resonancia donde se produce» (2). Es lo que distingue una misma nota musical (que sería el tono), producida por un violín o por una trompeta. En el caso de la voz, el timbre es lo que diferencia un sonido de otro, y es lo que permite reconocer dos sonidos como iguales, pese a ser pronunciados con distinto tono, e incluso por distintas personas. Depende fudamentalmente de la disposición y evolución de los formantes, que a su vez dependen de la configuración de las cavidades del tracto vocal, y por tanto del punto y modo de articulación.

En el espectro se puede apreciar el distinto timbre de dos sonidos, aunque se aprecia mucho mejor si se hace abstracción de la estructura del tono y se estudia la envolvente espectral. En la figura 5 se compara la envolvente espectral de dos vocales pronunciadas por un mismo locutor, una /a/ y una /i/. La /a/ presenta su primer y segundo formante bastante próximos y centrados (600 y 1500 hz), mientras que la /i/ tiene el primer formante muy bajo (unos 300 Hz) y el segundo muy alto (unos 2100 Hz).

Figura 5
Comparación de las envolventes espectrales de dos sonidos vocálicos, una /a/ y una /i/



• **Intensidad**: «cualidad por la que se oye a mayor o menor distancia; depende de la mayor o menor amplitud de las ondas sonoras» (DRAE). Los sonidos sonoros suelen tener mayor amplitud que los sordos, como ya se vio en la figura 2. La intensidad es tanto una característica propia de cada tipo de sonido, como también un parámetro que controla el hablante a voluntad. La intensidad se usa así para marcar el acento y el énfasis.

• **Cantidad**: es el tiempo que se invierte en la pronunciación, esto es, la duración del sonido. También se utiliza para marcar el acento.

Estas cualidades físicas del sonido identifican a cada hablante porque configuran su voz —la cual es individual— y le da su impronta personal. El habla sería como una huella digital: cada persona tiene unos parámetros vocales propios —un tono, un timbre, una intensidad y una cantidad, variables sólo en determinadas circunstancias físicas o psicológicas— que la diferencian de los demás. Además, los parámetros de tono, cantidad e intensidad son utilizados por el hablante para modificar o completar el sentido del mensaje. Estos parámetros constituyen lo que se denomina prosodia.

1.3 Aspecto lingüístico

Pero la voz (el habla) no es solamente un fenómeno físico y fisiológico. Se trata fundamentalmente de un fenómeno comunicativo, mediante el cual los hablantes que comparten un léxico, una gramática y un conocimiento del mundo similar, son capaces de intercambiar información. Los sistemas que pretenden utilizar eficazmente la voz como vía de comunicación no pueden quedarse en tratar su «forma», sino que deben adentrarse en los distintos niveles de conocimiento lingüístico y extralingüístico que soporta la misma (y el lenguaje en general): léxico, estructura morfológica, sintáctica, se-

mántica y pragmática (que recoge la información que tácitamente comparten hablante y oyente sobre el discurso y la realidad a la que se refiere).

Estos aspectos lingüísticos, compartidos con los sistemas de tratamiento del lenguaje escrito, presentan en el caso de la voz algunas peculiaridades que dificultan aún más la tarea: el lenguaje escrito suele ser mucho más estable y «normativo» que el hablado, en el que aparecen frecuentemente frases sin completar, estructuras no gramaticales, palabras nuevas, etc. También hay que considerar la variabilidad del lenguaje hablado por razones dialectales y sociales del hablante (sexo, edad, zona geográfica, ...), variabilidad que tiende a reducirse en el caso del lenguaje escrito.

2 Tratamiento automático de la voz

Si se pretende utilizar la voz como medio de comunicación entre el hombre y las máquinas, es necesario desarrollar técnicas que permitan el tratamiento automático de la misma, normalmente en un ordenador.

El primer paso consiste en transformar la señal de voz (una variación continua de la presión del aire o de la corriente en el cable de un micrófono) a un formato que permita su manejo por el ordenador (símbolos discretos o números). Esto se consigue con técnicas de cuantificación y codificación, que transforman la voz en una secuencia de números, y permiten al ordenador representarlos en pantalla, almacenarlos para luego reproducirlos con las mismas técnicas, o intentar descubrir la frase que pronunció el hablante, por ejemplo.

La señal de voz, convertida en un impulso eléctrico por medio de un micrófono, llega a un convertidor analógico digital. Este dispositivo mide la amplitud de la señal a intervalos de tiempo fijos, y transmite la secuencia de números a la memoria del ordenador. La frecuencia (el número de veces por segundo) con que se repite este proceso depende de la fiabilidad con que se quiera representar la señal de voz. La calidad de un Compact Disc precisa 44000 valores por segundo. Sin embargo, para las técnicas del procesado automático de voz es suficiente con 8000 valores por segundo. Las distintas representaciones que se han utilizado en las figuras corresponden a esta frecuencia de muestreo. Con este valor se pueden recoger las características de la voz hasta una frecuencia máxima de 4000 Hz. Es precisamente entre 0 y 4000 Hz donde se concentra la mayor parte de la energía de la señal de voz, y la estructura de formantes identificadora de los sonidos.

Pese a que existen aparatos diseñados específicamente para el estudio de la señal de voz utilizando estas técnicas (como los espectrógrafos digitales), la evolución de los ordenadores personales hacia "sistemas multimedia" les dota de unas tarjetas de sonido que realizan esta misma labor. Sin embargo, algunos sistemas de procesamiento de voz (sobre todo los de reconocimiento), no sólo requieren esta capacidad de grabar y reproducir voz en el ordenador, sino que además precisan tal potencia de cálculo que tienen que recurrir a tarjetas especiales de procesado de señal diseñadas específicamente para esta tarea.

Estas técnicas constituyen la base de todas las tecnologías del habla: reconocimiento de voz, síntesis de voz, reconocimiento y verificación de locutor, reconocimiento del idioma, etc.

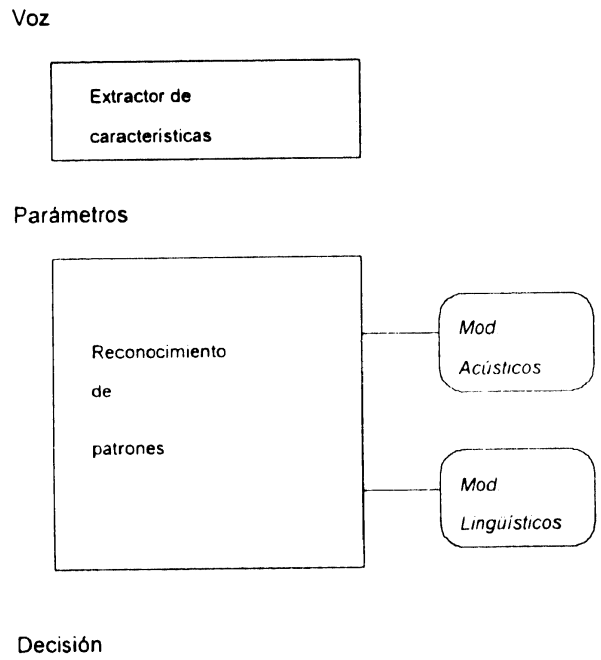
2.1 Reconocimiento de voz

Es el proceso mediante el cual la máquina reconoce y «entiende» lo que un locutor pronuncia y lo traslada a un texto o realiza una función determinada. La dificultad de la tarea se suele medir en tres dimensiones:

— **Estilo del habla**: se refiere a la manera de pronunciar las palabras que constituyen el mensaje. Respecto a este parámetro, los sistemas se clasifican desde el habla aislada (el usuario debe pronuciar una sola palabra o comando de los recogidos en la aplicación; cuando se encadenan varias palabras, cada una ha de ir precedida y seguida de un silencio), hasta sistemas de habla continua (en los que el usuario pronuncia las frases de una manera natural).
— **Dependencia del locutor**: el sistema puede estar entrenado para un usuario específico (con lo que se obtienen las mejores prestaciones), o bien ser capaz de funcionar con prestaciones similares para cualquier locutor (necesario en aplicaciones de servicios al público).
— **Tamaño y dificultad de la tarea**: este parámetro se refiere al número de palabras recogidas en el vocabulario de la aplicación (desde vocabularios pequeños con unas pocas decenas hasta vocabularios muy grandes de varias decenas de miles de palabras). Además, hay que considerar el grado de similitud fonética entre las palabras del vocabulario, y también las restricciones que limitan las combinaciones entre palabras del vocabulario y facilitan así la decisión.

La mayoría de los sistemas de reconocimiento presentan una estructura como la de la figura 6.

Figura 6
Estructura general de un sistema de reconocimiento de voz



En el nivel acústico, después de ser introducida la información mediante un micrófono, un conversor análogico-digital traduce la señal analógica de la voz en señal digital. A partir de este momento el ordenador ya está en condiciones de interpretarla y se dan los siguientes pasos:

• *Parametrización:* sirve para analizar los parámetros básicos que distinguen y caracterizan los sonidos. Es la fase de extracción de las características. Con la parametrización se segmenta el continuum de la señal acústica y se extraen los rasgos pertinentes. Para ello se divide la señal de voz en intervalos de entre 10 y 30 milisegundos, durante los cuales se supone que la voz mantiene unas características estables. De cada uno de estos intervalos se extrae la información relevante para el reconocimiento, que suele ser algún tipo de representación de la envolvente espectral (el timbre del sonido).

• *Reconocimiento de patrones:* a partir de la representación paramétrica de los distintos intervalos de la señal de voz, este módulo selecciona la unidad (o unidades) cuyo patrón mejor se ajusta a la pronunicación que se pretende reconocer. La peculiaridad más característica de este proceso, que marca su dificultad, es la variabilidad de la señal de voz. Variabilidad espectral y temporal entre distintas realizaciones de una misma palabra, incluso por un mismo locutor. Actualmente, los sistemas que mejor responden a esta variabilidad son los basados en la generación automática de los patrones a partir de realizaciones de las unidades que se pretenden reconocer. En la fase de reconocimiento, se calcula una medida de similitud entre la nueva realización y el modelo. Para ello se cuenta con un gran banco de sonidos. Un variado conjunto de técnicas permite comparar los datos introducidos (sonidos, palabras, construcciones) con los patrones almacenados (modelos acústicos).

Las técnicas empleadas (3) para reconocer los patrones son varias:

— Alineamiento temporal basado en algoritmos de programación dinámica. Obtiene la alineación óptima entre la locución que se está tratando y los patrones almacenados. Fue la primera técnica utilizada con un cierto éxito en esta tarea, durante los años 70.
— Modelos ocultos de Markov. Constituyen en la actualidad el núcleo de los principales sistemas de reconocimiento. Se basan en procedimientos totalmente estadísticos, tanto en la generación de los patrones de referencia como en la propia tarea de reconocimiento, y cuentan por tanto con un elaborado aparato matemático.
— Sistemas basados en redes neuronales. Similares a los modelos ocultos de Markov, no sólo realizan la configuración de los patrones a partir de los datos de entrenamiento, sino también la propia estructura de la red. Su principal inconveniente es que precisan una gran capacidad de cálculo, por lo que todavía son inviables para aplicaciones reales.

Los modelos acústicos son uno de los elementos más importantes para la obtención de una buena calidad en el reconocedor. El proceso de generación de esos patrones comienza con la selección del tamaño de las unidades lingüísticas. En sistemas con vocabularios reducidos se suele optar por unidades lingüísticas que coinciden con las palabras del vocabulario. Sin embargo, cuando el vocabulario aumenta, o puede variar frecuentemente, es mucho más práctico elegir unidades similares a los sonidos elemen-

tales de la lengua. Así, si se eligen los fonemas, es suficiente con un conjunto de entre 30 y 50 unidades para representar cualquiera de las palabras del vocabulario.

Cuando se opta por utilizar modelos acústicos correspondientes a unidades inferiores a la palabra, es necesario recurrir al conocimiento fonológico y fonético para realizar la transcripción de las palabras del vocabulario a este tipo de unidades. Normalmente esta transcripción se genera una vez y se almacena junto con el vocabulario de la tarea, de manera que no se plantea la tarea inversa de deducir la forma escrita a partir de la pronunciación de la palabra (normalmente es mucho más difícil la tarea de escribir al dictado que la de leer).

Estas técnicas sirven para crear el modelo acústico a partir del cual se comparan los sonidos introducidos. En sistemas con vocabularios pequeños, se obtienen así unas prestaciones aceptables. Sin embargo, con el avance de los sistemas hacia grandes vocabularios y habla continua, y sobre todo con el uso de unidades inferiores a la palabra como patrón del reconocimiento, el modelo acústico pierde capacidad discriminativa. Se precisa más información para obtener la secuencia correcta de palabras. Esta información se puede obtener de los modelos lingüísticos, mecanismos que capturan la redundancia inherente al lenguaje natural. Por ejemplo, si el modelo acústico entrega como mejores opciones:

Ella *diente* lo sucedido
Ella *siente* lo sucedido

Un adecuado modelo lingüístico puede elegir la segunda opción frente a la primera, pues tras un pronombre personal en modo nominativo es más probable que aparezca un verbo que un sustantivo.

La informatización de las unidades y reglas de los niveles morfológico, sintáctico, semántico y pragmático de la lengua sirve para crear el modelo lingüístico. Un conocimiento extralingüístico —el saber general que toda persona tiene para mantener una conversación- es necesario cuando se trata de sistemas que «comprenden» el habla, es decir, entienden lo que se les dice.

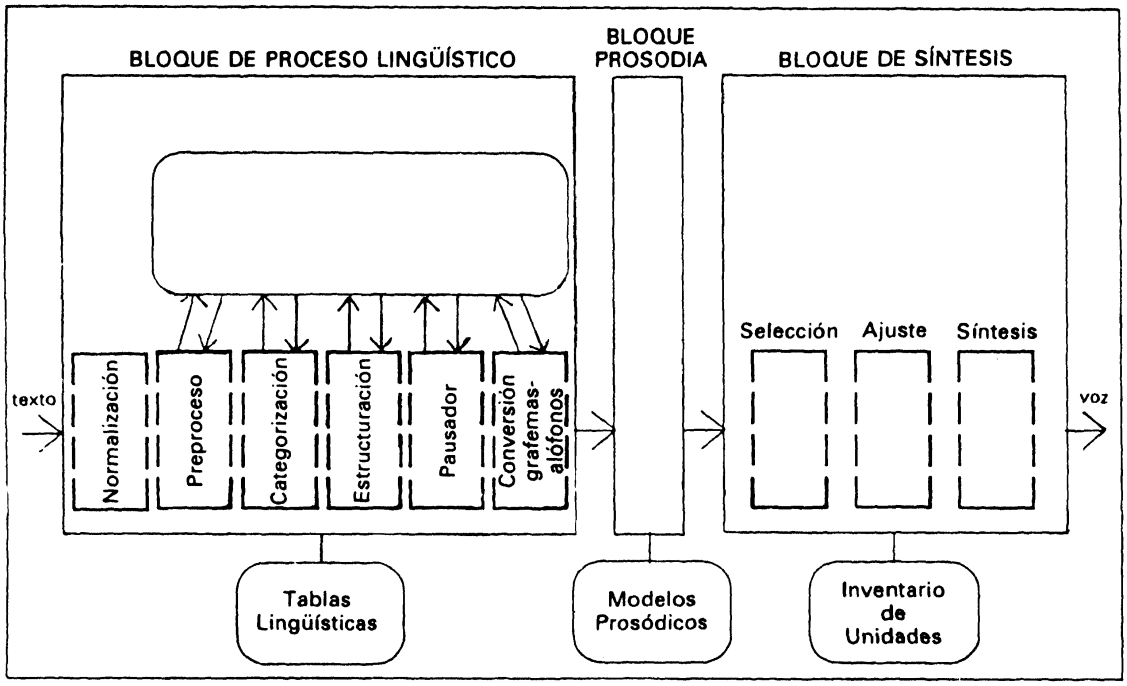
Es preciso considerar también otros factores que incrementan la complejidad de esta investigación: la variedad de locutores por razones sociológicas (sexo, edad, zona geográfica a la que pertenecen...), la coarticulación de los sonidos en la cadena hablada, la amplitud del vocabulario, la consideración de un gran volumen de datos para hallar las invariantes, la exigencia de conocimientos de disciplinas muy diversas —fisiología, psicología, pedagogía, lingüística, informática, telecomunicaciones—, el ruido del entorno físico en el que funcionen, por citar algunos de ellos.

2.2 Sistemas de síntesis de voz

Se utiliza indistintamente la expresión «síntesis de voz» como «conversión texto-voz» para referirse a estos sistemas. Su objetivo es leer «en voz alta» un texto, con una inteligibilidad y calidad similar a la de la voz humana. La capacidad de hablar (síntesis de voz a partir de concepto) todavía es un tema de investigación exploratoria.

La estructura típica de un sistema de conversión texto-voz es la que se presenta en la figura 7.

Figura 7
Estructura general de un sistema de conversión texto-voz



• **Bloque de proceso lingüístico:** su misión es obtener la secuencia de fonemas que se corresponde a lo que aparece escrito en el texto. Para ello es necesario normalizar y expandir la presentación del texto de entrada (transcribir números, expandir abreviaturas, etc.). También es necesario realizar un análisis lingüístico suficiente para otras funciones del sistema (como elegir los puntos del discurso más adecuados para la realización de pausas y otros elementos prosódicos, o determinar la acentuación de las palabras). En los sistemas más sencillos, este análisis lingüístico no existe, mientras que en los más complejos consiste en un análisis morfológico y sintáctico e incluso de nivel superior (por ejemplo, para detectar la presencia de nuevos elementos de información en el discurso). Este tipo de sistemas puede hacer uso de esta información adicional para mejorar la naturalidad de la lectura, mediante la inserción de pausas adicionales a las marcadas con signos ortográficos o la utilización de esa información en la generación de una prosodia más rica y variada. Finalmente, con la información generada se realiza una división en sílabas de las palabras, se determina la acentuación fonética de las mismas, y se transcribe la secuencia de letras en la secuencia de fonemas correspondientes.

• **Bloque de generación prosódica**: se encarga de asignar duración (cantidad) a cada uno de los fonemas, y generar un contorno entonativo (tono) para toda la frase, que facilite la comprensión del mensaje y mejore la naturalidad. Normalmente no se suele marcar de manera especial la intensidad de los sonidos, sino que se conserva su amplitud característica.

• **Bloque de síntesis**: con la información de la secuencia de fonemas y las características de duración y tono, se encarga de generar los sonidos correspondientes, de ma-

nera que la voz resultante tenga una calidad próxima a la de la voz humana. Según el modelo de síntesis empleado se tienen sistemas (4) con:

— Modelos articulatorios: intentan reproducir el comportamiento del sistema humano de la fonación. Utilizan conocimientos muy próximos al dominio de la fonología y la fisiología. Aunque teóricamente pueden alcanzar la calidad y naturalidad de la voz humana, en la práctica no se dispone de conocimientos suficientes para realizar modelos lo suficientemente complejos ni para suministrarles los parámetros de control necesarios. Su utilización es muy reducida, y se localiza en grupos de investigación más interesados en el estudio y caracterización del lenguaje que en sistemas de conversión texto-voz.
— Modelos de formantes: intentan reproducir las características espectrales de los sonidos, fundamentalmente la evolución de los formantes. Utilizan conocimientos de la acústica de la señal de voz, aunque también de la fonología y la fonética. Hasta hace unos años, han sido los modelos de síntesis de mayor calidad y los más utilizados, pues entre otras ventajas presentan unas necesidades de cálculo y memoria muy reducidas.
— Modelos de concatenación de unidades: intentan simplemente copiar la señal de voz. Para ello se almacenan los sonidos representativos del idioma y posteriormente se genera la voz uniendo unos sonidos con otros. Sin embargo, es preciso utilizar algún tipo de modelo de la señal de voz (procedente generalmente de las técnicas de codificación de voz) para poder controlar y modificar la prosodia de los trocitos de voz, fundamentalmente su tono. Utilizan técnicas de la acústica y la codificación de voz. Su utilización es relativamente reciente, pues sus necesidades de cálculo y memoria son superiores a las de los modelos de formantes. Sin embargo, permiten alcanzar en poco tiempo una calidad similar (e incluso superior) a la de los mejores sistemas de formantes, que precisan grandes conocimientos de fonética y decenas de años de refinamiento.

3 Recursos, sistemas y aplicaciones

En este último apartado se van a presentar los distintos sistemas y aplicaciones, basados en estas tecnologías, que se encuentran ya disponibles para los usuarios, tanto servicios públicos como aplicaciones profesionales y domésticas, comenzando por un pequeño apartado previo para mencionar los distintos recursos lingüísticos disponibles en la actualidad para el español.

Estas tecnologías necesitan una serie de *recursos* básicos, relacionados directamente con el idioma. Parte de ellos son comunes también a los sistemas de procesamiento del lenguaje escrito. Su obtención y desarrollo es un proceso lento y caro. Ésta es una de las razones que han venido limitando el desarrollo de estos sistemas para el español, dada la escasa disponibilidad de estos recursos, sobre todo comparados con los del inglés.

Desde su fundación, el Instituto Cervantes ha comenzado una labor para recoger y distribuir la información sobre los distintos recursos de este tipo que se están desarrollando y utilizando para el español. Uno de los recursos básicos es el corpus, una colección de muestras del lenguaje. Entre los corpus del texto escrito recogidos en el Informe (5) merece la pena destacar el Corpus Chileno de Referencia y el Corpus del Español de

la República Argentina, ambos de la Universidad Autónoma de Madrid; de fines específicos es TANGORA, corpus recogido por IBM para el desarrollo de su sistema de reconocimiento, con 120 millones de palabras. Entre los orales -básicos para el desarrollo de sistemas de reconocimineto y síntesis del habla- se encuentra el Corpus Oral de Referencia del Español Contemporáneo, también de la Autónoma de Madrid, EUROM1 y Albayzin —en fase de desarrollo.

Recientemente la Real Academia Española de la Lengua ha comenzado la construcción del Corpus de Referencia del Español Actual (CREA) que constará de 200 millones de palabras tomadas de textos escritos y orales, tanto del español de Europa como del español de América, del período comprendido entre 1975-2000. El Corpus Diacrónico del Español (CORDE), desarrollado por la misma institución, incluirá textos de la lengua española desde sus orígenes hasta 1975.

Otros recursos básicos para el desarrollo de estos sistemas son: diccionarios, gramáticas y estudios de la lengua, lematizadores, analizadores sintácticos (parsers), etc. (6).

A partir de los recursos y con una labor importante de investigación, algunas instituciones han estado durante los últimos 20 ó 30 años desarrollando *sistemas* basados en tecnologías del habla. Numerosas universidades se han dedicado a los sistemas de reconocimiento y síntesis del habla. De hecho, los mejores sistemas de reconocimiento en los últimos años han sido conseguidos en Cambridge University y Carnegie Mellon University. En España, numerosas universidades cuentan con grupos trabajando en procesado de voz, así como algunas empresas nacionales y multinacionales. Uno de los mejores foros para consultar la actividad en este campo en español son las publicaciones de la Sociedad Española para el Procesamiento del Lenguaje Natural.

A continuación se presenta un extracto (7) de las *empresas,* que desarrollan y comercializan sistemas de reconocimiento y síntesis del habla:

— Reconocimiento de voz: AT&T, BBN Hark, Daimler-Benz, Dragon Systems, Marconi, IBM, Philips, Responsive Systems, Telefónica de España, Voice Control, Voice Processing, Lemout & Hauspie. La mayor parte de estos sistemas están disponibles para múltiples idiomas.
— Conversión texto-voz: AT&T, Berkeley Speech Technologies, Centigram, CSELT, Elan Informatique, Telefónica de España, Infovox, First Byte, Lemout & Hauspie. La tendencia entre estos productos va también hacia los sistemas multilingües.

Mientras que los sistemas de reconocimiento suelen precisar una gran potencia de cálculo, y por tanto una tarjeta adicional de procesado de señal, los sistemas de conversión texto-voz pueden funcionar directamente sobre un ordenador personal sin otros requisitos que una tarjeta de sonido tipo Sound Blaster. Algunos de estos sistemas de conversión también están disponibles en tarjetas especiales de coste reducido, de manera que descargan al ordenador de la tarea.

Como se puede comprobar, ya hay disponibles numerosos sistemas (de hecho son mucho más numerosos que la lista referida en el citado informe; falta, por ejemplo, uno de los sistemas más famosos de conversión texto-voz: DecTalk, reputado como uno de los de mayor calidad: es el sistema que utiliza Stephen Hawkins para comunicarse). Sin embargo, hay muy pocas *aplicaciones* basadas en estas tecnologías; las más difundidas no pasan del uso de la voz pregrabada: máquinas expendedoras, juguetes, presentación

del número solicitado en el servicio de páginas blancas (003), o servicios de contestador automático centralizados. Recientemente han aparecido algunas basadas en reconocimiento y/o síntesis de voz: servicios de consulta de saldo bancario, encuestas, servicio de información sobre oferta de empleo público del MAP, servicio de información de pistas de esquí de ATUDEM, etc. Sin embargo, la difusión de estos servicios (8) es muy reducida, y la capacidad de diálogo con el sistema no pasa de la elección de una de las opciones posibles de un menú muy limitado.

En lo que se refiere a las aplicaciones del ámbito doméstico y profesional, la situación ha sido muy similar hasta muy recientemente, pese a que este tipo de aplicaciones se benefician de la dependencia de un único locutor, lo que mejora sus prestaciones. Algunas de las empresas antes mencionadas ofrecen sus productos para PC, normalmente con una placa adicional destinada a la captura y reproducción del sonido y al intensivo cálculo numérico (si se trata de reconocimiento de voz). Sin embargo, el coste de estos sistemas es muy elevado, y tan sólo han tenido éxito entre determinados grupos de profesionales, como los radiólogos y los abogados, y personas con algún tipo de minusvalía. El lanzamiento por IBM de su sistema de dictáfono (y otro similar de Dragon Systems) supuso un primer cambio de esta tendencia, con un producto destinado a un grupo mucho más amplio de usuarios, con un coste más reducido, y perfectamente integrado en una aplicación «completa», un editor de textos. Muy recientemente, en el entorno del sistema operativo Windows95, Microsoft ha lanzado una propuesta que facilitará el que realmente surjan aplicaciones que utilicen las posibilidades de la tecnología del habla.

Las últimas generaciones de PCs presentan ya unas características que posibilitan la integración de las tecnologías del habla sin coste adicional: suficiente capacidad de proceso, memoria, y sobre todo placas de sonido de muy bajo coste. Teniendo en cuenta esta situación, la propuesta pretende implantar el desarrollo de aplicaciones, lo que ha sido el factor clave para el éxito del PC: la existencia de un entorno en el que desarrolladores de tarjetas y programas pueden ofrecer sus productos, sabedores de que la arquitectura abierta del sistema garantiza su funcionamiento (siempre que se cumplan unos determinados requisitos). A su vez, esta posibilidad facilita la competencia, la reducción de precios, y por tanto el crecimiento del mercado con lo que los fabricantes pueden amortizar sus costes de investigación y desarrollo.

Microsoft ha definido una interfaz de programación para aplicaciones que empleen reconocimiento y síntesis de voz, y ha facilitado un mecanismo para ofrecer estos servicios como si fueran un dispositivo más del PC, como puede ser la pantalla o el teclado. De esta manera, la empresa que desarrolla la tecnología del habla puede ofrecer sus sistemas como un servicio básico, y dejar a otras empresas más próximas al mercado el desarrollo de aplicaciones verdaderamente útiles. Incluso será posible que un usuario pueda cambiar el sistema de reconocimiento, por ejemplo, y cualquier aplicación que lo utilizase seguiría funcionando perfectamente. Exactamente igual que si se cambia el tipo de monitor. Esta nueva posibilidad, disponible desde principios de 1996 con el paquete TAZZ de Windows95, tiene ya disponibles numerosos sistemas de reconocimiento y conversión texto-voz. De la misma manera, según se vayan desarrollando nuevas aplicaciones para Windows95, dichas aplicaciones podrán incluir prestaciones de tratamiento de voz.

Aparte de estas aplicaciones ya disponibles en el mercado, en los laboratorios de empresas y sobre todo de universidades se están desarrollando prototipos, que aunque muy lejos todavía de la imagen de comunicación con el ordenador de *2001: Una odisea*

del espacio o *StarTrek,* ofrecen ya un entorno de comunicación mucho más confortable y potente. Se caracterizan por el uso de sistemas de reconocimiento de habla continua y muy grandes vocabularios, junto con técnicas de procesamiento de lenguaje natural. Estas aplicaciones reúnen las tecnologías del procesamiento de voz con las que tradicionalmente estaban restringidas al tratamiento del lenguaje escrito. Por ejemplo, numerosos centros tienen proyectos de traducción automática entre dos o más idiomas, en los que el usuario puede hablar en un idioma, y el sistema reconoce, traduce, y pronuncia en otro idioma el mensaje deseado.

Otras aplicaciones están centradas en el acceso del usuario a la información, sin tener que preocuparse de cómo está almacenada y organizada esa información, ni del lenguaje que utiliza el ordenador para acceder a la misma. Por ejemplo, en el Instituto Tecnológico de Massachusets (MIT) se está desarrollando un sistema genérico para el acceso a información almacenada en un ordenador, GALAXY. Este sistema utiliza reconocimiento de voz y conversión texto-voz, y junto con técnicas de comprensión y generación del lenguaje, permite el acceso a informaciόm sobre vuelos, guías de ciudades, tiempo meteorológico o anuncios de automóviles de segunda mano.

El sistema INFORMEDIA de Carnegie Mellon University recoge noticias de los informativos de televisión, e interpretando el contenido de las mismas por medio del lenguaje, las clasifica y posteriormente permite el acceso del usuario a las mismas mediante solicitudes empleando el lenguaje natural hablado, por ejemplo: «Dame todas las referencias del accidente de un camión en la Nacional II».

Terminamos esta exposición con dos proyectos europeos de suministro de información a través de servicios de voz, desarrollados para su aplicación en el entorno de las bibliotecas públicas y de gran repercusión para los discapacitados:

— SPRINTEL (Speedy Retrieval of Information on the Telephone). Iniciado en 1994, participan cuatro instituciones/bibliotecas públicas de Bélgica, Alemania, Irlanda y Holanda. Se propone examinar la eficacia y los costes/beneficios de la tecnología de reconocimiento de voz en la búsqueda de información.
— REACTIVE TELECOM (Residential Access to Information Via Everyday Telecommunication Tools). Iniciado en 1995, plantea un nuevo servicio de suministro de información a través de las tecnologías del teléfono y la televisión. Las demandas se realizan por teléfono y se procesan mediante un sistema de reconocimiento de voz que las traslada a la base de datos. Los resultados de dichas búsquedas son enviados a casa del cliente a través de la televisión por cable, habiéndosele informado previamente por vía telefónica, mediante un sistema de síntesis de voz, del número de páginas para su localización.

Bibliografía

1. GILI GAYA, S. *Elementos de fonética general.* Madrid: Gredos, 1971.
2. QUILIS, A. *Fonética acústica de la lengua española.* Madrid: Gredos, 1971.
3. Para una más completa explicación de las distintas técnicas de reconocimiento se puede consultar el número especial de la revista IEEE ASSP Magazine, *Decoding Methods for Continuous Speech Recognition,* julio de 1990.
4. Para una visión histórica de los sistemas de síntesis de voz se puede consultar el artículo

«Review of Text-to-Speech Conversion for English», de Dennis H. Klatt, en el *Journal of Acoustic Society of America,* vol. 82, número 3.

5. *Informe sobre recursos lingüísticos para el español (I).* Alcalá de Henares: Instituto Cervantes, 1994; una actualización del mismo ha salido en 1996.
6. DIEZ CARRERA, C. Las industrias de la lengua: panorámica para los gestores de información. Madrid: Biblioteca Nacional, 1994.
7. GRAY, M. Speech Recognition and Text-to Speech Survey. *Voice+.* Junio, 1995.
8. El número de septiembre de 1995 de la revista *Speech Communication* está dedicado a distintos desarrollos piloto que integran técnicas de procesado de voz en servicios para el público.