Situación actual de las tecnologías del habla

Josep Martí Roca
Enginyeria La Salle
Universitat RAMON LLULL

PACS 43.72. Ja

Resumen

En la actualidad las tecnologías del habla han alcanzado un importe desarrollo, tanto en el ámbito de la síntesis como en el del reconocimiento. Este progreso está muy relacionado con el conocimiento de la fisiología de la voz y de la audición, a los que la tecnología trata de imitar.

El objeto de este artículo es presentar la situación actual de estas tecnologías y de los últimos avances relacionados con el mundo de las telecomunicaciones.

El texto analiza diferentes formas de codifición de voz, que permite un alto grado de comprensión de datos, con una pérdida de calidad aceptable. Específicamente el artículo analiza las técnicas más comunes de los convertidos texto-habla, del reconocimiento de palabras aisladas y del reconocimiento de habla continua. En este mismo sentido, en el trabajo se examinan algunos sistemas existentes en el mercado como aplicación de estas técnicas.

Summary

At the present time, speech technologies have achieved a remarkable development in synthesis aspects, as well as those of recognition. This progress is strictly related to the knowledge of the voice and hearing physiology, to which technology tries to emulate.

The object of this article is to present the current situation of these technologies and the last advances related to the world of telecommunications.

The text analyses different forms of voice coding, that keep a high degree of data compression, with acceptable quality losses. Specifically this article analyzes the most common techniques in text-to-speech converters, recognition of isolated words and continuous speech. In the same way, this paper examines some products on the market, as applications of these techniques.

I. Introducción

Los recientes progresos en el ámbito de las tecnologías del habla se han caracterizado por la interdisciplinariedad de los factores e investigadores que las están impulsando. Éste no es un caso único y aislado en el progreso de las ciencias. Precisamente muchas de las que denominamos tecnologías punta se hallan en esta situación de confluencia entre diferentes disciplinas, que encuentran una nueva posibilidad de desarrollo a partir de la intersección de sus conocimientos.

En el proceso de comunicación oral entre el hombre y la máquina confluyen muchos conocimientos del mundo de las Telecomunicaciones, de la Fisiología de la voz y del oído, de la Lingüística aplicada, de la Psicología, de la Informática, de la Acústica, etc. Cada una de estas ciencias, con sus múltiples derivaciones, aportan interesantes conocimientos con una aplicación muy directa a los problemas de la generación del habla, su transmisión, comprensión e interpretación.

La aportación de la Acústica se realiza básicamente a nivel del soporte físico del proceso: la generación, la propagación, el tratamiento auditivo de la señal, la captación y el análisis de sus características. De alguna manera, ésta puede parecer una aportación un tanto humilde, a modo de soporte básico sobre el cual las demás ciencias levantarían el edificio de nuevas elucubraciones y de fascinantes aplicaciones; pero siempre habrá que volver sobre el soporte físico como canal de experimentación de todas las aplicaciones. Podríamos decir con un cierto "orgullo acústico", un tanto exagerado, que sin alguna de estas disciplinas las tecnologías del habla avanzarían más lentamente, pero sin el soporte de la señal acústica, simplemente, estas tecnologías no existirían.

El objetivo de la presente exposición es puntualizar cuáles son las aportaciones más importantes, desde el punto de vista de la Acústica y del Tratamiento de la Señal, en el progreso de los sistemas que facilitan la interacción oral hombre-máquina y al mismo tiempo describir y detallar el estado de la cuestión en la actualidad.

II. El modelo acústico de la generación de la voz

El conocimiento preciso de la fisiología de las cuerdas y del tracto bocal ha aportado mucha luz para la generación de

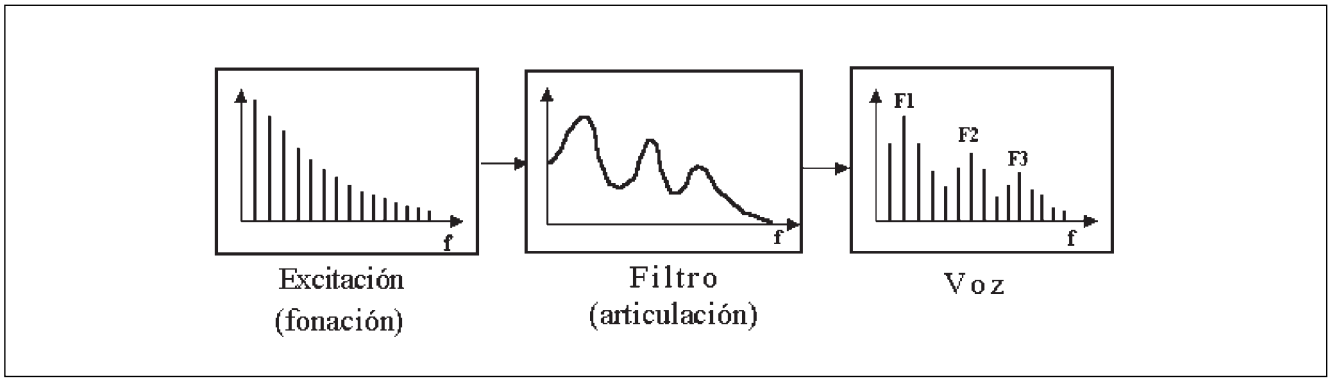

Figura 1. Modelo acústico de la generación del habla.

la voz sintetizada. De hecho, los circuitos sintetizadores o los sistemas implementados totalmente en "software" para la producción de la voz no hacen sino reproducir, sobre un soporte electrónico o sobre un modelo matemático, el mismo proceso de la fisiología del habla humana.

Básicamente, la producción de la voz natural se explica por un proceso de convolución entre la señal de excitación (señal periódica de las cuerdas vocales o señal de ruido producida por una constricción al paso del aire) y la respuesta impulsional del tracto bocal y/o nasal. Es el mismo modelo analógico de un generador de señal acoplado a un filtro de respuesta variable con el tiempo. La voz, como señal de salida del filtro, aporta a nuestro oído tanto la información sobre la excitación, como la posición de los órganos articulatorios que definen cada uno de los sonidos básicos de una lengua (Fig. 1).

El procesado digital de señales permite hoy en día la generación holgada de este proceso en tiempo real. De hecho, la señal de excitación se puede modelar a partir de unos pocos parámetros, que definen básicamente la periodicidad, la forma de onda, la energía o el ruido aleatorio. El modelo de tracto vocal se describe igualmente por unos pocos parámetros más que han de especificar la geometría del conducto vocal en cada momento y el grado de acoplamiento con el tracto nasal. La resolución temporal de nuestro oído para la percepción de los cambios en la onda acústica es tal que difícilmente puede apreciar fluctuaciones de muy corta duración. Normalmente se opta por una actualización de los parámetros de excitación y de articulación de una forma discreta, cada 10 o 20 ms. Nuestro oído, por su parte, percibe el resultado como una fluctuación continua de la onda acústica.

Tabla 1. Nivel de compresión

	Número de parámetros	Ritmo de actualización (ms)	Datos segundo	Grado de compresión
Voz digitalizada a 10 KHz			10000	1
Síntesis LPC	15	20	750	13,3
Síntesis articulatoria	9	50	180	55,5

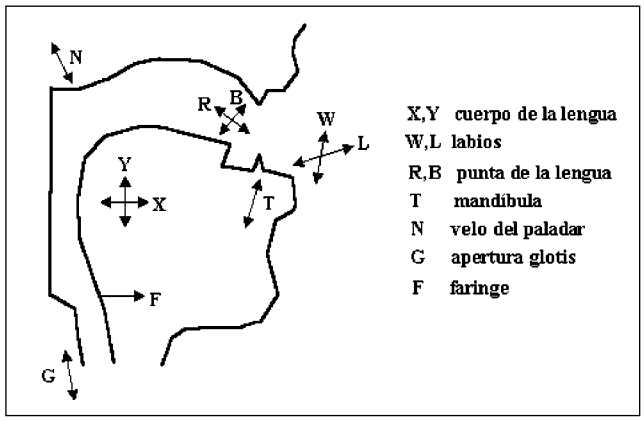

Figura 2. Modelo articulado

En total, los parámetros que definen el sistema fonatorio pueden ser del orden de 15, de modo que, para la generación de la voz se precisa un ritmo de información comprendido entre 1500 y 750 datos por segundo, que, comparados con un ritmo de muestreo de la señal a 10 KHz, representa un grado de compresión entre 6,6 y 13,3.

Ahora bien, el ritmo de producción de la voz viene marcado por la velocidad del movimiento de los órganos articulatorios (lengua, mandíbula, labios, velo del paladar, ...). La posición exacta de estos órganos se puede cuantificar mediante unos 9 datos numéricos (Fig. 2), actualizados cada 40 o 50 ms, si contamos con un sistema de interpolación que realice automáticamente las transiciones. El paso de estos parámetros a la descripción geométrica del tracto vocal resulta inmediata. Dicha geometría suele especificarse con la llamada función de área, que define la superficie de la sección del tracto en función de la distancia a los labios. Por tanto, a partir de los datos articulatorios, el ritmo de información puede ser aún más bajo: entre 225 y 180 datos por segundo, que corresponden a un grado de compresión entre 44,4 y 55,5 (Tabla 1).

A la hora de generar voz parametrizada necesitamos un sistema de actualización de la fonación y de los articuladores, que en ningún caso puede presentar problemas de velocidad trabajando sobre los sistemas informáticos actuales. La función de área del tracto acostumbra a darse en forma discreta, lo que equivale a un sistema de tubos conectados, cada uno de los cuá-

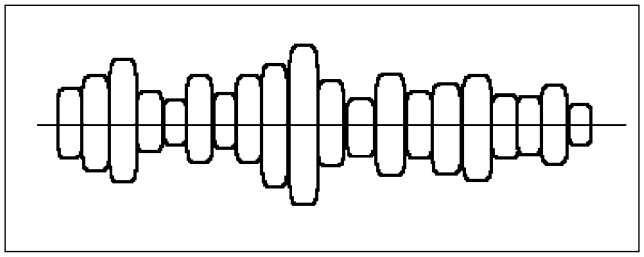

Figura 3. Modelo de tubos del tracto vocal.

les es de sección constante (Fig. 3). Acústicamente disponemos de herramientas rápidas de cálculo que permiten deducir la función de transferencia del conducto total, con sus resonancias y antiresonancias (polos y ceros), que caracteriza a cada uno de los sonidos. Este sistema de generación resulta totalmente coherente con el proceso de coarticulación de los sonidos contiguos del habla humana; el cual es totalmente ajeno a una simple yuxtaposición de sonidos estacionarios y discretos.

Una vez diferenciados los dos aspectos de fonación y articulación podemos observar que, básicamente, la articulación hace referencia a la sucesión de sonidos que se pueden transcribir mediante una cadena fonética de alófonos. Por el contrario, la excitación, y concretamente la fluctuación de la frecuencia fundamental y de la energía aportan una información de tipo suprasegmental, que conjuntamente con el ritmo, constituye la prosodia del lenguaje. Ambos procesos se pueden considerar independientes el uno del otro. Por ejemplo, con una misma cadena de sonidos y con diferentes curvas melódicas se puede expresar una afirmación, una interrogación o una duda.

En cuanto a la caracterización a nivel segmental de una cadena de sonidos, básicamente hay que acudir a la evolución espectral, que no es sino otra forma de aportar la información articulatoria. Para evitar informaciones redundantes, también aquí se puede acudir a sistemas de compresión que permitan igualmente la reconstrucción de los datos fundamentales del espectro. Así, por ejemplo, existen sintetizadores de voz que utilizan las resonancias espectrales o formantes (Fig. 4), como parámetros que permiten modelar la evolución del espectro. Otros sintetizadores codifican la distribución de la energía espectral mediante bandas yuxtapuestas, más o menos estrechas, que cubren la totalidad del espectro de la voz.

Actualmente, uno de los métodos más utilizados para la codificación del espectro es el que utiliza la transformada Cepstrum (Fig. 5), definida como la antitransformada de Fourier del logaritmo del módulo del espectro de la señal de voz:

$$s(t) = F^{-1} [\log | F (s (t)) |]$$

en donde:

s(t) es el Cepstrum de s(t)

F i F^{-1} representan los operadores Transformada y antitransformada de Fourier.

De alguna manera, podríamos decir que esta transformada detecta las periodicidades que se observan en un espectro logarítmico (expresado en dB). Dichas periodicidades, para un fragmento sonoro, son de dos tipos: unas periodicidades rápidas, debidas a la estructura armónica del espectro, que se repiten en los múltiplos de la frecuencia fundamental Fo, y unas fluctuaciones mucho más lentas, no periódicas, que nos dan la envolvente espectral. Estas fluctuaciones lentas se manifiestan en la parte baja del Cepstrum y caracterizan la forma del tracto vocal. De hecho, también se reducen a unos pocos datos numéricos (del orden de 15 o 20), que parametrizan la información articulatoria. La información correspondiente a la fuente excitadora se sitúa, por el contrario, en la parte alta del Cepstrum, y corresponde básicamente a la periodicidad de Fo (detección de periodicidad y período correspondiente) y a la energía global de la señal. Esta separación de las dos informaciones permite un proceso de descon-

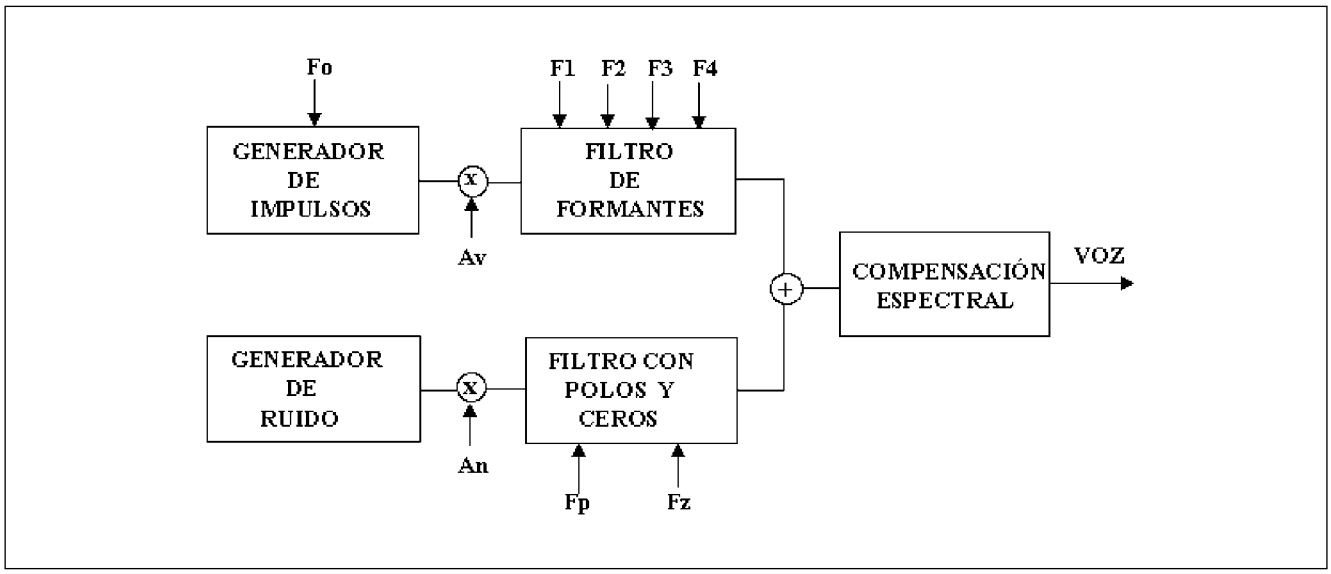

Figura 4. Sintetizador por formantes

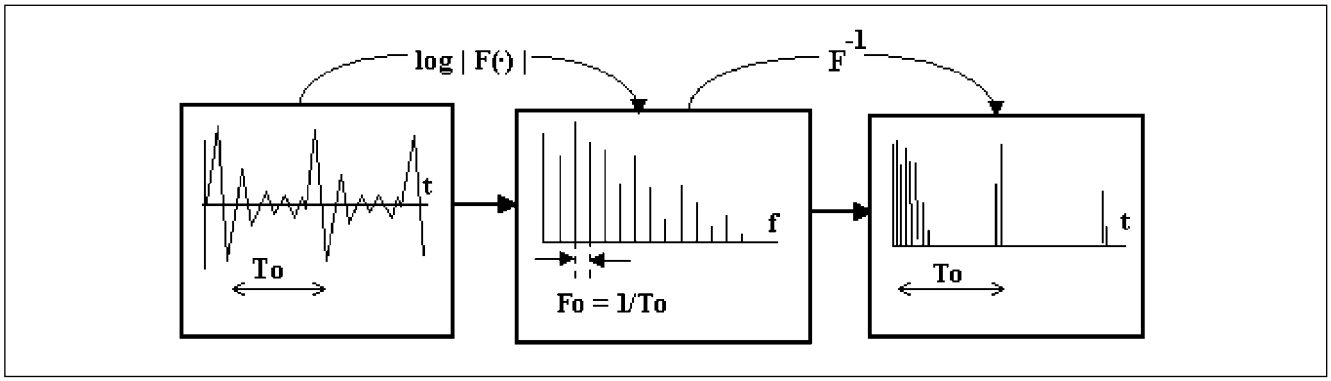

Figura 5. Cepstrum.

volución de la señal de voz, para recuperar separadamente la excitación y la respuesta impulsional del tracto vocal.

III. El modelo de percepción auditiva

Los estudios del sistema auditivo humano han ofrecido aportaciones muy interesantes en el campo del reconocimiento del habla. Hoy en día, se conoce con bastante precisión el funcionamiento del oído humano como detector de frecuencias, a modo de analizador espectral. Concretamente, el órgano de Corti realiza un análisis espectral por bandas discretas yuxtapuestas denominadas bandas críticas. El tema ha sido ampliamente estudiado (Zwiker, 1980) y traducido a normativas internacionales.

Los equipos clásicos de análisis espectral trabajan según un sistema de filtros de banda constante o de porcentaje constante (bandas de octava, de 1/3 de octava, de 1/2 octava, etc.). Nuestro oído, sin embargo, utiliza un sistema de bandas que no sigue estas reglas. La anchura de las bandas críticas tiene un crecimiento de forma aproximadamente lineal hasta la frecuencia de 1 KHz y aproximadamente logarítmico por encima de ella. El paso de la primera a la segunda zona no es brusco sino progresivo (Fig. 6). Para codificar la señal de voz es muy adecuado un sistema que analice el contenido energético dentro de cada una de estas bandas y que permita una reproducción posterior que esté también de acuerdo con el grado de resolución que presentan dichas bandas. Por esta vía, el reconocimiento acústico de los sonidos se realiza exactamente según nuestra fisiología, lo que ha conducido a resultados mucho más adecuados de los sistemas de reconocimiento.

Actualmente está muy en boga un sistema de codificación del habla a través de un sistema denominado mel-Cepstrum (Davis, 1980), que combina el análisis Cepstrum y una ponderación de los coeficientes, de acuerdo con las bandas críticas.

En algunas aplicaciones de reconocimiento del habla se va más allá en lo referente a la emulación del sistema auditivo humano; teniendo en cuenta otros aspectos como el llamado efecto de enmascaramiento de unas bandas críticas por otras, y la variación de la sensibilidad auditiva en función de la frecuencia y del nivel. Para dicho proceso se aplican curvas de ponderación de sonoridad ("Loudness") constante, que son variables según el nivel global del sonido. Finalmente, una vez aplicadas estas transformaciones sobre la señal temporal, ésta se parametriza según alguno de los métodos ya conocidos. Mediante este desarrollo se han conseguido resultados apreciablemente más satisfactorios en los procesos de reconocimiento (Hermansky, 1990).

IV. El entorno acústico y las tecnologías del habla.

La voz humana para poder satisfacer el objetivo de una buena comunicación hombre-máquina - al igual que sucede en la comunicación hombre-hombre - precisa de un medio de comunicación que sirva de soporte para la onda acústica y para la señal eléctrica, que pueda transportar la información a más larga distancia. La interacción de la voz con este medio supone la consideración de otros muchos problemas acústicos que puedan interferir en el proceso de transmisión y propagación.

Mencionemos particularmente la problemática asociada a los recintos donde se realiza la propagación del sonido. Dichos recintos tienen un comportamiento acústico específico que se puede caracterizar por su respuesta impulsional, que se convoluciona con la señal de voz dando lugar a una señal acústica diferente. Así, por ejemplo, un sistema de reconoci-

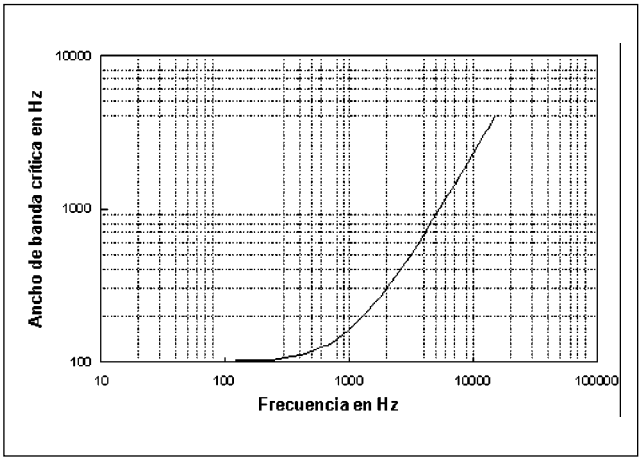

Figura 6. Anchura de las bandas críticas en función de la frecuencia

miento que funciona perfectamente en un determinado entorno, puede resultar totalmente inadecuado cuando nos trasladamos a un espacio distinto. Para minimizar este problema cabe la solución de utilizar un nivel de señal directa muy por encima del nivel reverberante del recinto.

En algunas ocasiones, con todo, esta predominancia del nivel directo no se puede conseguir fácilmente. Por ejemplo: cuando la distancia entre el locutor y el micrófono no se puede reducir suficientemente. En estos casos es preciso utilizar alguna técnica que permita eliminar o atenuar el efecto de la sala. Esto se puede lograr mediante procesos de desconvolución que eliminan la contribución del recinto, o por "arrays" de micrófonos que focalizan su directividad sobre una zona concreta de la sala, desde la que se emite el mensaje oral. Un caso particular de interferencia del recinto es la presencia de eco o repetición del mensaje por reflexión, con un retardo superior a los 50 ms y unos niveles de señal retardada no inferiores a - 10 dB respecto de la señal directa. La cancelación de eco es un tema suficientemente estudiado que puede resolverse por desconvolución cepstral.

Otro handicap inherente al medio de transmisión es la presencia de ruido aleatorio que oscurece la señal de voz, dificultando el reconocimiento del habla o la identificación del hablante. Las soluciones al problema del ruido se pueden atacar sencillamente a partir de aislamientos adecuados o también mediante filtros adaptativos que cancelan los ruidos después de una identificación de sus características espectrales.

Los problemas de ruido y eco se pueden dar también con el soporte eléctrico a través de las líneas de transmisión, que no siempre presentan la calidad que cabría esperar. El tratamiento para la cancelación de eco o de ruido eléctrico en las líneas telefónicas se realiza por procesos similares a los que se dan en el caso del ruido acústico, con filtros adaptativos y canceladores de eco.

V. Áreas de desarrollo actual de las tecnologías del habla.

Comentamos a continuación algunos rasgos sobre el desarrollo y la situación actual de las tecnologías del habla, como respuesta a los retos y a las necesidades de los nuevos sistemas de telecomunicación y de la interacción oral hombre-máquina.

Una necesidad fundamental de las comunicaciones y del almacenamiento de voz es la reducción del "bit rate" que permite utilizar los mismos canales para transmitir muchas más comunicaciones de voz simultáneamente, o reducir el espacio de almacenamiento de mensajes orales. Para la compresión de voz telefónica digitalizada con un ancho de banda comprendido entre 200 y 3400 Hz se utilizan técnicas que van desde la modulación PCM con 64 Kbps (Kbits por segundo), hasta la codificación LPC-10 que consigue los 2,4 Kbps (Tabla II).

Naturalmente, la reducción del "bit rate" lleva siempre consigo una reducción en la calidad del mensaje transmitido, pero no tan elevada como podría parecer en un principio, ya que estos sistemas aprovechan el alto grado de redundancia que lleva implícita, normalmente, la señal de voz. Una adecuada codificación de esta redundancia permite su regeneración al otro extremo de la línea de comunicación, sin necesidad de transmitir formas de onda repetitivas.

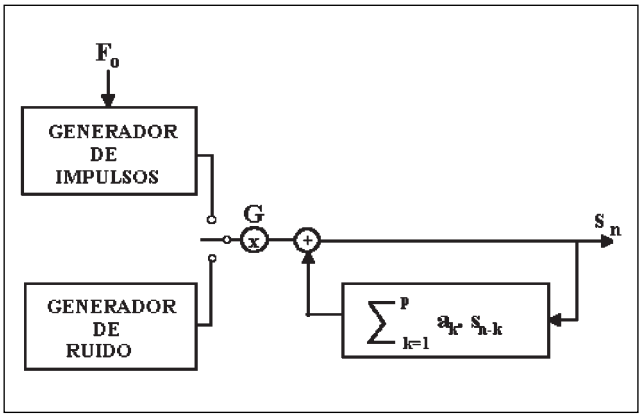

Figura 7. Sintetizador LPC.

Tabla II. "Bit rate" de los diferentes sistemas de compresión de voz (Rabiner, 1995)

Nombre	Standard	Bit rate (Kbps)
μlaw pulse code modulation (PCM)	G.711	64
Adaptive differential PCM (ADPCM)	G.721	32
Low delay code excited linear prediction (LD-CELP)	G.728	16
Regular pulse linear prediction	GSM	13
Vector storage excitation (VSELP)	IS-54	8
Code excited linear prediction	FS-1016	4,8
Linear predictive coding (LPC-10E)	FS-1015	2,4

Hoy en día, aparecen otras formas de transmisión de sonido que exigen anchos de banda más elevados, como es el caso de la videoconferencia, que posee una calidad de transmisión de radio en AM (50-7000 Hz) o FM (20-15000 Hz), o incluso con calidad CD audio (10-20000 Hz). Estas tecnologías precisan factores de compresión muy altos para poder garantizar unos costes no excesivamente elevados. Así, por ejemplo, para poder asegurar la calidad CD con un muestreo de 44,1 KHz, dos canales estéreo y una resolución de 16 bits por muestra, se precisa un "bit rate" de 44,1 x 2 x 16 = 1,41 Mbps. El sistema "Perceptual audio code" (PAC) desarrollado por los laboratorios AT&T Bell Labs ha logrado codificar dos canales de audio a 128 Kbps sin pérdida apreciable de calidad, lo que representa un factor de compresión de 11.

En el entorno telefónico se utiliza voz codificada en lo referente a mensajería oral, sistemas de respuesta vocal y contestadores automáticos. En todos estos casos se exige una calidad de voz muy cercana a la voz natural. En otras aplicaciones, como la telefonía móvil y las comunicaciones orales por satélite, se toleran niveles de calidad más bajos, ya que el ambiente y los canales son mucho más ruidosos; de forma que no se apreciaría un mayor esfuerzo en la mejora de la calidad.

Es interesante observar también el proceso que se está siguiendo en la mejora de los sistemas de síntesis de voz y,

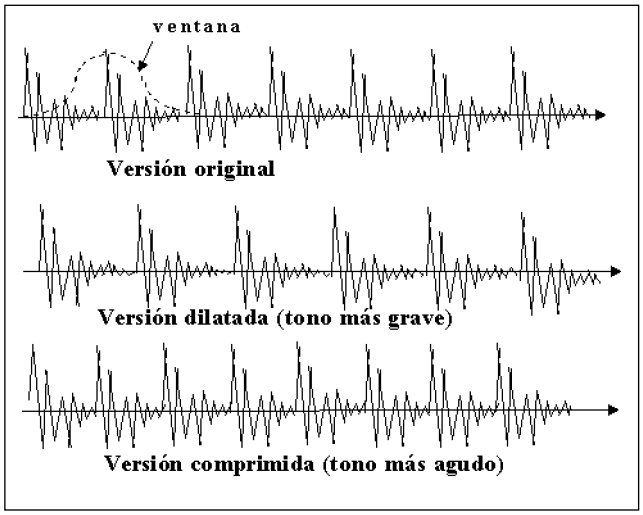

Figura 8. Principio de funcionamiento de la técnica PSOLA.

concretamente, en los conversores texto-habla. Actualmente se están obteniendo buenos resultados en la coarticulación de sonidos concatenados a partir de unidades elementales, como pueden ser los fonemas, los difonemas o las semisílabas. El avance de la investigación se centra en la obtención de modelos prosódicos en concordancia con el sentido del texto y que se apliquen de forma automática en la generación del habla.

La parametrización de unidades elementales se ha realizado, con frecuencia, a partir de la idea básica de la predicción lineal (LPC) (Fig. 7) y con diferentes complementos que intentan mejorar el sistema de excitación (excitación multipulso, RELP, CELP, ...); pero no siempre se alcanza una calidad de voz aceptable.

La introducción de modelos articulatorios para la generación del habla ha aportado una importante mejora cualitativa en la síntesis. El sistema implica un esfuerzo de cálculo considerable para la pronunciación en tiempo real, pero facilita de forma natural la coarticulación de sonidos contiguos (Meyer, 1989). Las dificultades estriban en el conocimiento preciso de los parámetros articulatorios correspondientes a cada uno de los sonidos que permitan una adecuada cuantificación del espacio articulatorio. (Larar, 1988).

Otra técnica muy distinta parte de un almacenamiento de formas de onda temporal que se pueden concatenar, asegurando igualmente un buen control de la prosodia. Es el sistema denominado Pitch - Synchronous Overlap - Add (PSOLA) (Charpantier, 1989) (Moulines, 1990) que realiza un solapado de señales enventanadas de una longitud exactamente igual a 2 o 4 períodos de la frecuencia fundamental (Fig. 8). El mayor o menor grado de solapamiento en el momento de la síntesis permite actuar sobre el período de Fo, con un efecto posterior de desenventanado que regenera una señal temporal con las características articulatorias de la señal de partida. El sistema, hoy por hoy, es el que ofrece unos mejores resultados de naturalidad, a costa de una gran extensión en la base de datos que contiene las unidades elementales. De todas formas, la disponibilidad de grandes extensiones de memoria y la facilidad de acceso rápido a la misma no es un problema para los actuales sistemas informáticos.

El impacto más importante de los conversores texto-habla se está manifestando en los sistemas de transmisión oral con acceso directo a bases de datos o informaciones textuales, sin necesidad del intermediario humano. Así, por ejemplo, se ofrecen sistemas de consulta de directorios telefónicos, direcciones, servicios..., acceso oral automático a informaciones bancarias, lectura automática de noticias y otras informaciones modificables diariamente sobre textos escritos, informaciones sobre medios de transporte y turismo, lectura oral para invidentes, etc.

Un campo mucho más extenso y a la vez menos logrado es el reconocimiento del habla humana o el control oral de máquinas, que supone un proceso inverso al de la síntesis. Los estadios más sencillos son aquellos que trabajan con vocabularios restringidos de palabras pronunciadas de forma discreta y entrenamiento previo con un locutor determinado. En estos casos, los resultados obtenidos hasta el momento son ya altamente convincentes.

Más difícil es el reconocimiento de palabras concatenadas, como sucede en los sistemas de numeración de cada lengua, donde resulta difícil detectar la separación entre las palabras pronunciadas sin pausa intermedia. El caso más complejo es el que corresponde al reconocimiento del habla continua, tal como se produce en una conversación normal. A este nivel no se han logrado aún resultados totalmente satisfactorios, aunque ya están apareciendo algunos productos de mercado en diferentes lenguas, incluso en español, que precisan de una fase previa de aprendizaje con el usuario. De momento, aún no se ha logrado un sistema de reconocimiento de habla continua con independencia total del locutor.

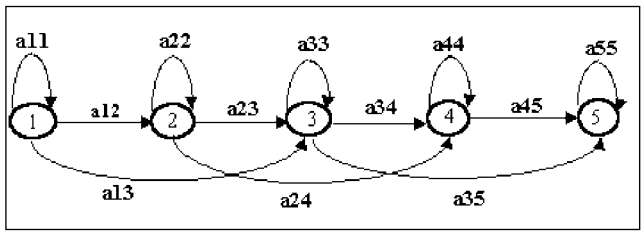

Figura 9. Modelo de Markov con cinco estados.

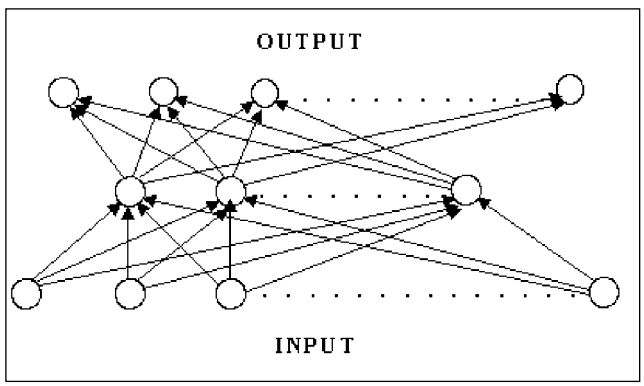

Figura 10. Red neuronal.

Uno de los problemas fundamentales que se dan en el reconocimiento del habla es la dificultad de alineamiento temporal entre los modelos y las palabras a reconocer; ya que la pronunciación de una misma palabra puede cambiar el ritmo de forma no lineal, según los efectos de la prosodia de un locutor concreto. Una vez logrado el alineamiento temporal óptimo se contabiliza el grado de semejanza entre la locución a reconocer y los diferentes modelos en términos de distancia. El sistema "Dynamic Time Worping" (DTW) es uno de los más utilizados para realizar este proceso; con la particularidad de que simultáneamente efectúa el alineamiento temporal óptimo y el cálculo de la distancia. El sistema se ideó, primariamente, para el reconocimiento de palabras aisladas, pero con algunas adaptaciones se aplica también al reconocimiento de palabras concatenadas, dentro de un vocabulario limitado. Uno de los sistemas más utilizados en este caso es el "Level building" que busca, entre todas las posibles concatenaciones de palabras, aquella que ofrece una distancia acumulada mínima.

El procedimiento denominado "Hidden Markov Model" (HMM) (Rabiner, 1989) es hoy en día uno de los más utilizados para el reconocimiento de palabras concatenadas. El método consiste en modelar estadísticamente una serie de estados y transiciones entre ellos, a partir de una base de datos que contiene todo el vocabulario a reconocer (Fig. 9).

El proceso de aprendizaje de un vocabulario o de unos sonidos elementales es fundamental cuando se pretende alcanzar el reconocimiento del habla continua. La implantación de las redes neuronales ha abierto nuevos horizontes para el entrenamiento automático de estos sistemas de reconocimiento (Fig. 10).

El impacto comercial de los sistemas de reconocimiento se ha acentuado últimamente en los equipos de dictado que realizan el reconocimiento de textos largos dictados por un locutor entrenado y, generalmente, en forma de palabras separadas por silencios. Igualmente se comercializan sistemas de "navegación informática" a partir de menús orales, propios de diferentes programas o con aplicaciones telefónicas que permiten acceder a informaciones precisas de acuerdo con los intereses del usuario. Otras aplicaciones tales como la distribución de llamadas telefónicas a partir de la vocalización del destinatario, la pronunciación oral de los dígitos telefónicos o la solicitud de informaciones concretas han conseguido también resultados satisfactorios.

Un aspecto particular del reconocimiento es la verificación o identificación de hablantes a partir de su voz, como sistema de control de acceso y seguridad. En estos casos, la atención del sistema debe centrarse más en las diferentes variantes de la pronunciación, y no tanto en el reconocimiento del vocabulario concreto utilizado. Se distinguen dos tipos de aplicaciones: los sistemas de verificación en los cuáles un usuario determinado reclama su reconocimiento y la máquina ha de verificar su identidad y los sistemas de identificación que tratan de reconocer al parlante entre una lista de candidatos previamente registrados.

VI. Temas avanzados

Mencionemos finalmente algunos de los temas de investigación más recientes en tecnologías del habla, que están empezando a ofrecer aplicaciones interesantes, o que están aún en fase de experimentación; de tal forma que podrían ofrecer resultados espectaculares dentro de pocos años.

En cuanto a los conversores texto-habla, la mayor parte de los esfuerzos se centran en la obtención de una mayor naturalidad en la locución, con modelos prosódicos automáticos deducidos del análisis sintáctico del texto a pronunciar. Actualmente, muchos sistemas se limitan a la introducción de unos pocos modelos entonativos que se van repitiendo continuamente, dando la sensación de una lectura monótona.

A parte de la lectura, existen también otros modelos de entonación correspondientes al habla espontánea que pueden tener aplicación en formas de expresión menos formales. Éste es un campo en el que se están iniciando estudios muy alentadores.

En lo que se refiere al reconocimiento del habla, se están obteniendo buenos resultados en la identificación de palabras determinadas dentro de contextos totalmente aleatorios ("word spotting"). El sistema consiste en la detección de unas palabras clave sin necesidad de proceder a un reconocimiento total de toda una cadena de sonidos. Este sistema puede aportar toda la información necesaria para provocar una actuación rápida en procesos de diálogo o de interacción.

Tanto en síntesis como en reconocimiento se están probando modelos de sincronismo entre la voz y la imagen en movimiento de los labios del hablante, que, en el caso de los conversores texto-habla, permiten una imagen visual del rostro del locutor. En el caso del reconocimiento del habla se puede integrar la información acústica con la información visual del movimiento de los labios (lectura labial), para conseguir unos mejores resultados de reconocimiento.

En sistemas completos de interacción hombre-máquina es muy importante considerar una fase intermedia entre el reco-

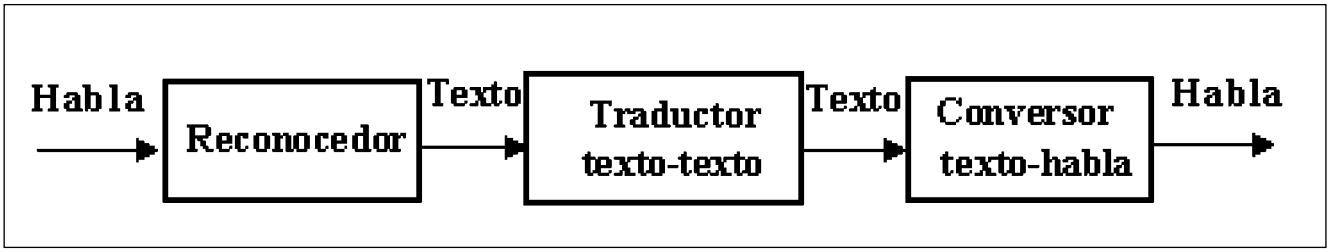

Figura 11. Traducción automática voz-voz.

nocimiento y la síntesis, que organice el diálogo de modo que la respuesta esté debidamente coordinada con los requerimientos expresados oralmente por el hablante. En algunas ocasiones, la máquina deberá solicitar informaciones complementarias al locutor, para poder ofrecer la información solicitada. Todo ello supone una aproximación cada vez más semejante a lo que sería un proceso de conversación entre dos interlocutores humanos.

En estos procesos de interacción oral se tiende a utilizar sistemas multilingües, capaces de trabajar en diferentes lenguas. Un servicio muy importante sería el reconocimiento automático de la lengua utilizada por el locutor humano, que en cualquier caso, deberá ser un proceso previo al inicio de cualquier sistema de diálogo. Actualmente se han iniciado ya experiencias de reconocimiento acústico de las características fonéticas de una determinada lengua, a partir de la forma de locución utilizada por el usuario.

Un paso más será la traducción automática voz-voz entre diferentes lenguas. Desde hace ya bastantes años se vienen comercializando productos de traducción automática texto-texto, con resultados menos satisfactorios de los que en un principio se esperaban. Muchos de ellos ofrecen traducciones excesivamente literales, sin un proceso de análisis sintáctico y semántico que facilite la comprensión profunda del sentido de cada frase. A partir de un traductor suficientemente inteligente, podríamos añadir un proceso inicial de reconocimiento de voz-texto en la primera lengua y un paso final de síntesis texto-voz en la segunda lengua; llegando así a la traducción automática voz-voz que se podría establecer en doble dirección, de forma que ya podríamos soñar con el entendimiento automático entre usuarios de distintas lenguas, utilizando cada uno la suya propia (Fig. 11). Hace unos años se presentó una experiencia de traducción oral, con vocabulario reducido (453 palabras), entre el inglés y el español. El sistema fue presentado por Telefónica en la Feria de Sevilla (1992) (Roe, 1992). De momento la traducción con lenguaje ilimitado no deja de ser un sueño, aunque no sería de extrañar que, en pocos años, lo pudiéramos ver convertido en realidad.

Referencias

F. Charpentier; E. Moulines. *Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones.* Eurospeech 89, vol. 2, pp 13-19. (1989).

S. Davis; P. Mermelstein. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* IEEE Trans. ASSP, 28, pp 357-366. (1980).

H. Hermensky. *Perceptual linear predictive (PLP) analysis of speech.* JASA, 87 (4), pp 1738-1752. (1990).

J. N. LARAR; J. SCHROETER; M. M. SONDHI. *Vector quantization of the articulatory space.* IEEE Trans. on ASSP vol. 36, nº 12, pp. 1812-1818, (1988).

P. Meyer; R. Wilheims; H. W. Strube. *A quasiarticulatory speech synthesizer.* JASA, 86 (2), pp 523-539 (1989).

E. Moulines; F. Charpentier. *Pitch sinchronous waveform processing techniques for text-to-speech synthesis using diphones.* Speech Communication, 9, pp 453-467, (1989).

L. R. Rabiner. *A tutorial on Hidden Markov Models and its applications to speech recognition.* Proc. IEEE, vol. 72, 2, pp 257-286, (1989).

L. R. Rabiner. *The impact of voice processing on modern telecommunications.* Speech Communication, 17, pp 217-226, (1959).

D. B. Roe; P. J. Moreno; R. W. Sproat; F. C. Pereira; M. D. Riley; A. Macarron. *A spoken language translator for restricted-domain context-free languages*". Speech Communication, vol. 11, pp 311-319, (1992).

E. Zwicker; E. Terhard. *Analytical expressions for critical-band rate and critical bandwidth as a function of frequency.* JASA, 58 (5), pp 1523-1525, (1980).