SISTEMA PARA LA CONVERSIÓN DE TEXTO A VOZ EN ESPAÑOL EN TIEMPO REAL

Andrés Santos, J.C. Olabe, Elías Muñoz, C. López Barrio
(Esc.Téc.Sup. Ingenieros de Telecomunicación.Madrid)

Y

Antonio Quilis, Miguel Martínez
(Cons.Sup.Inv.Cient.Madrid)

La síntesis del habla de una forma automática es un tema que siempre ha suscitado gran interés. Pero, hasta muy recientemente, no se ha podido abordar su realización práctica. Han contribuido a ello dos hechos - principalmente: por un lado, la elaboración de una teoría acústica de la producción del habla, y por otro, la posibilidad de realizar gran cantidad de cálculos y de manejo de datos, que presentan los medios informáticos actuales.

La aplicación principal de la voz sintética radica en su utilización como medio para transmitir información. El habla es el principal medio de comunicación entre los seres humanos, por tanto, en el intento de hacer ordenadores lo más accesibles a la mayoría de la población y acomodarse a sus necesidades, surge lógicamente la necesidad de interaccionar con ellos mediante la voz. Una segunda aplicación de un sistema de este tipo se encuentra en los lectores ópticos, que permiten a una persona - ciega "leer" cualquier tipo de texto. Por último, entre sus aplicaciones se puede mencionar también la posibilidad de investigar más profundamente los mecanismos de producción y percepción del habla.

Este artículo se centra en la descripción del sistema automático de conversión de texto en voz, en español, desarrollado en el Dpto. de Electrónica de la ETSI de Telecomunicación de Madrid, en colaboración con - miembros del Instituto Miguel de Cervantes del CSIC.

SISTEMAS DE SÍNTESIS DE VOZ

Los sistemas que producen una salida hablada se pueden clasificar en tres grupos, por orden de complejidad creciente:

1) mensajes pregrabados: sistemas que graban y reproducen mensajes (correo electrónico, contestadores etc.)

2) representación paramétrica: sistemas que producen un número limitado de mensajes mediante distintas técnicas (síntesis por concatenación de sílabas, demisílabas, etc.), y

3) síntesis por regla: sistemas de síntesis ilimitada, basados en reglas.

Mensajes Pregrabados

La forma más elemental de producir voz por medios electrónicos es simplemente almacenar, en forma digital o analógica, el mensaje requerido. Es un procedimiento en el que no se necesita ningún conocimiento de la naturaleza de la señal de voz ni de la estructura del aparato vocal humano. Pero su principal limitación es la necesidad de una memoria importante para almacenar un vocabulario incluso muy limitado.

Se puede conseguir una mejor relación entre la cantidad de memoria requerida y la duración del mensaje, aprovechando las propiedades conocidas de la señal de voz. Se han logrado condificaciones eficientes de esta señal que optimizan la cantidad de memoria requerida, pero si se desea sintetizar un número grande de mensajes, o sobre todo, si los mensajes son cambiantes, se debe acudir a otro tipo de técnicas.

Representación Paramétrica

Para conseguir la flexibilidad suficiente a la hora de generar distintos mensajes, además de disminuir la cantidad de memoria necesaria, es preciso emplear unidades o mensajes elementales (palabras, sílabas, fonemas...) que se unen o concatenan para formar las frases. Con objeto de efectuar correctamente las transiciones entre unidades, éstas se deben representar por un conjunto de parámetros que determinan sus características más sobresalientes, bien en el dominio del tiempo o en el de la frecuencia. De esta forma, en el momento de sintetizrar la señal de voz, se necesitará realizar un proceso más complicado, pero se habrá ganado capacidad de almacenamiento y se podrán manipular los parámetros para efectuar las transiciones entre mensajes elementales de una manera más natural. Así se pueden construir frases sin pérdida de naturalidad.

Los parámetros elegidos suelen ser de dos clases según el tipo de síntesis que se vaya a efectuar:

- Síntesis por Formantes: Se modelan las resonancias naturales del aparato fonador, caracterizando su frecuencia y su ancho de banda. Se deben caracterizar también las fuentes de excitación tanto sonora como de ruido de turbulencia. El sintetizador produce una señal que presenta las características acústicas más relevantes del habla.

- Síntesis por Predicción Lineal (LPC): Al igual que la técnica anterior, ésta caracteriza también a la señal en el dominio de la frecuencia, pero está basada en un modelo matemático que determina un conjunto de coeficientes, que en cada momento dependen de las condiciones previas de la señal. La calidad de la síntesis obtenida está relacionada con el número de coeficientes que definen la señal.

Por otra parte, la elección de las unidades o mensajes elementales determinan en gran medida la flexibilidad y la calidad del sistema. Cuanto menor es la longitud de estas unidades, se necesitan menos, pero el proceso de cálculo para concatenarlas es más laborioso.

Existen diferentes sistemas, basados en esta técnica, que utilizan como unidades las sílabas, demisílabas o demifonemas. Su principal inconveniente es la dificultad en parametrizar correctamente cada una de las unidades de síntesis, permitiendo además su adapatación a los diferentes entornos cambiantes. Cuantas más unidades se dispongan, mejor se podrá simular cada uno de los sonidos encontrados en el habla natural, pero más difícil será su caracterización y la de las transiciones entre ellas.

Nos resta por último considerar otro método de síntesis, que es el basado en la regulación de los diferentes parámetors a sintetizar, conocido como síntesis por regla.

Síntesis por Regla

Los fonemas se han considerado como las unidades básicas de síntesis en muchos sistemas, debido a su reducido número y al estudio detallado de que han sido objeto en numerosos trabajos. Existen diversas realizaciones en lengua inglesa que emplean síntesis por formantes, y que permiten obtener la evolución de los parámetros de los sonidos y reglas para determinarla automáticamente.

Un sistema completo de conversión de texto escrito a voz debe obtener también la secuencia de fonemas a partir del texto de entrada. En el español, el número de reglas requerido para establecer estea conversión es en cierta medida reducido, si se compara con el conjunto de reglas necesario en inglés.

Nuestro interés se centra en este tercer apartado: en sistemas que pueden sintetizar cualquier texto, utilizando como unidades fundamentales los fonemas, y mediante un gran número de reglas para su tratamiento.

SISTEMA DE CONVERSION DE TEXTO EN VOZ

Seguidamente se describe el sistema de conversión de texto en voz que hemos desarrollado. Consta básicamente de dos procesos bien diferenciados: la conversión del texto de entrada en parámetros de control del sintetizador, y el sintetizador propiamente dicho.

Respecto a la primera parte, la obtención de los parámetros se realiza en cuatro pasos distintos: preproceso del texto de entrada, conversión de letras a sonidos, cálculo de duración y entonación de cada sonido y conversión de sonidos a parámetros del sintetizador.

Preproceso del Texto de Entrada

El primer tratamiento que se debe hacer con el texto de entrada es una normalización. En cualquier texto que nos encontremos existen cifras, abreviaturas, diversos signos de puntuación, etc... Antes de hacer la conversión texto-voz debemos aplicar una uniformidad al texto.

Este tratamiento consta de las siguientes fases:

1) Conversión de Números en Texto: Las cifras que se encuentren,se convierten en el texto correspondiente. Se aceptan los siguientes casos:

   - Números cardinales de hasta 12 dígitos, separados o no por puntos.
   - Números cardinales seguidos por una abreviatura, en cuyo caso se efectúa la concordancia en género y número con la abreviatura, que, a su vez, se expande. También se introduce la preposición "de" cuando corresponde.
   - Números decimales, separados por una coma, o por un punto, cuando esto no causa confusión con un único número carinal.
   - Números ordinales comprendidos entre 1 y 19.
   - Fechas, escritas en la forma X-Y-Z, siendo X uno o dos dígitos que representan el día, Y tres letras que representan un mes, y Z dos o cuatro dígitos,que representan el año.

2) Expansión de Abreviaturas y Siglas: Los textos incluyen con mucha frecuencia palabras abreviadas que se deben expandir. Para ello se dispone de una tabla de abreviaturas y siglas con su texto correspondiente.

3) Localización del Acento Fonético: Un paso previo a la obtención de la entonación correcta del texto es la localización del acento fonético en cada palabra.

   Si existe acento ortográfico, el acento fonético coincide con éste. Si no existe, se debe determinar su posición de acuerdo con ciertas reglas bien conocidas. Las reglas en nuestro sistema exceptúan unas 60 palabras como artículos, preposiciones, conjunciones, etc., que no llevan ningún acento.

4) Tratamiento de los Signos de Puntuación: Los signos de puntuación en un texto nos dan información principalmente de los lugares donde se deben hacer las pausas y de la duración de éstas. Tienen además otras funciones como indicar el tipo de frase: enunciativa,exclamativa, interrogativa,etc., determinando el tipo de inflexión de la entoncación.

   Reciben un tratamiento especial los signos siguientes: punto (.), coma (,), signo de interrogación (?), paréntesis (()), dos puntos (:) y punto y coma (;).

Conversión de Letras a Sonidos

Una vez se tiene el texto formado únicamente por palabras escritas (letras)y los correspondientes signos de puntuación, el siguiente proceso es la conversión de letras en sonidos o fonemas.

El sistema desarrolado emplea 42 "unidades" prestablecidas, según se recogen en la tabla 1. En estas "unidades" están incluidos los 14 diptongos posibles en español.

A diferencia de otras lenguas, en español existe una relación sencilla entre ortografía y pronunciación por lo que se pueden definir reglas para efectuar la conversión de letras a sonidos.

Tabla 1.Unidades utilizadas para la Conversión Texto-Voz.

Unidad de Síntesis	Ejemplo	Alófono normativo	Unidad de Síntesis	Ejemplo	Alófono normativo
A	palabra	[a]	R	rama	[ɾ]
E	sello	[e]	L	luna	[l]
I	hilo	[i]	l	calle	[ʎ]
O	hola	[o]	M	mamá	[m]
U	lunes	[u]	N	nana	[n]
ai	aire	[ai]	n	leño	[ɲ]
ei	seis	[ei]	F	feo	[f]
ia	hacia	[ja]	S	casa	[s]
ie	tiene	[je]	X	caja	[x]
io	labio	[jo]	Z	caza	[θ]
iu	ciudad	[ju]	b	haba	[β]
ua	agua	[wa]	d	hada	[ð]
oi	hoy	[oi]	g	haga	[ɣ]
ue	suelo	[we]	y	haya	[ǰ]
ui	ruido	[wi]	p	papa	[p]
uo	arduo	[wo]	B	barco	[b]
ou	lo usó	[ou]	T	tela	[t]
au	causa	[au]	D	dama	[d]
eu	europa	[eu]	K	casa	[k]
C	charo	[ʧ]	G	gasa	[g]
r	muro	[r]	Y	yate	[ǰ]

Nuestro sistema para las letras A,E,I,O,U,F,J,K,M,ñ,P,S.T y Z hace corresponder un único sonido, en cualquier posición. Para las demás letras hay definido un conjunto de reglas muy sencillas que, en cada caso, tiene en cuenta, como mucho, el sonido anterior y el siguiente al procesado en cada momento.

Previendo el uso común de palabras extranjeras, se ha creado un diccionario que las recoja, junto con su correspondiente transcripción fonética.

Así se logran resultados correctos incluso para palabras que no cumplen las reglas definidas.

Funciones Prosódicas

Una vez que se tiene la secuencia de fonemas correspondientes al texto, se les debe asignar a cada uno su duración y su frecuencia fundamental. Estos dos factores, duración correcta de los sonidos, acompañada también de la apropiada frecuencia fundamental, proporcionan no solamente naturalidad, sino también mejor inteligibilidad al habla producida.

Ambos parámetros se determinan según una serie de reglas. Antes de aplicar estas reglas, efectuamos dos procesos sobre la secuencia de fonemas obtenida, como son la formación de diptongos y la homologación de sonidos adyacentes.

1) Reglas de Duración: La duración de un sonido se determina dependiendo tanto de sus propias características acústicas y articulatorias, como del contexto en que se encuentra localizado (posición en la palabra, acentuación, etc.). La duración de las pausas, por el contrario, depende generalmente del signo de puntuación que las provoca.

   Las reglas de duración se determinaron haciendo un estudio de 1400 palabras, todas ellas con sentido y leídas por un locutor en una frase portadora. La medición de los sonidos se realizó sobre espectrogramas digitales. Las medidas así obtenidas se introdujeron en una base de datos. Con los resultados logrados se han definido cinco reglas de duración para las vocales, cada una de las cuales tiene en cuenta un efecto distinto; cuatro para las consonantes y una para los diptgongos, que aseguran a cada sonido su correcta duración prácticament en todos los casos.

2) Reglas de Entonación: La frecuencia fundamental de una frase depende de un gran número de factores, algunos difíciles de estudiar como puede ser su estructura sintáctica o incluso semántica.

   Analizando un gran número de casos con voz natural se ha obtenido también un algoritmo para determinar la evolución de la frecuencia fundamental en una frase. Dicho algoritmo tiene en cuenta variables como el tipo de frase (enunciativa o interrogativa), posición de los acentos y de las pausas, y niveles de referencia inicial y final.

Conversión de Sonidos a Parámetros del Sintetizador

Los sonidos obtenidos de los procesos anteriores se descomponen en una secuencia de uno o más segmentos o sonidos elementales. Cada segmento se corresponde con un conjunto de parámetros de control del sintetizador que se presenta en la sección siguiente.

Para efecturar correctamente la síntesis de los sonidos y la transición entre ellos, se debe producir un nuevo conjunto (o trama) de parámetros del sintetizador cada 10 ms..

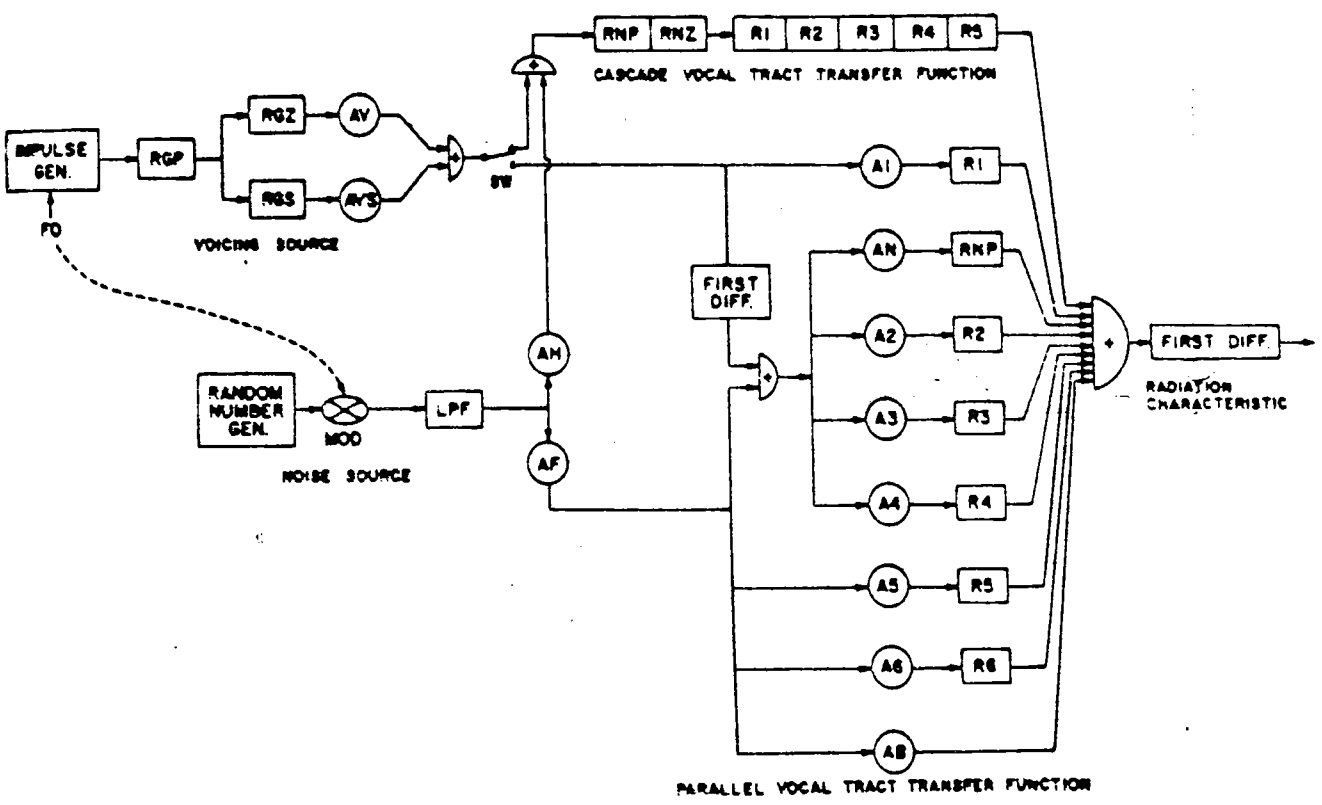

Fig. 1. Sintetizador de Klatt

La parametrización de los distintos sonidos ha sido un esfuerzo muy laborioso, pues especialmente la frecuencia de los tres primeros formantes es muy sensible a los sonidos adyacentes.

El proceso de cálculo por regla de los coeficientes es bastante complejo, porque, además de determinar los valores estacionarios para cada sonido, se deben calcular los valores intermedios correspondientes a las transiciones entre ellos. El proceso se encuentra descrito en 1 .

Sintetizador de voz

Consideramos, finalmente, la última parte de la conversión, que es la producción de voz a partir de la información obtenida en los procesos anteriores. Esta conversión se realiza a través de un filtro digital, complejo en su funcionamiento, que se conoce como el sintetizador de Klatt. 2 . Este sintetizador (véase fig. 1) tiene dos fuentes de excitación, una fuente sonora y una fuente de ruido blanco; y reproduce las características acústicas del aparato fonador humano, principalmente en cuanto a sus resonancias naturales. El resultado es una señal de salida similar a la onda sonora obtenida en el tracto vocal de un hablante.

Se ha elegido este sintetizador, primero por la calidad de la voz producida, y segundo por la facilidad con que se pueden reproducir los sonidos de la voz humana. El gran número de controles que posee permite aproximar las características de energía, formantes, etc., de cualquier segmento de voz. Además, a diferencia de otros tipos de sintetizadores, como los LPC, existe una analogía inmediata entre los datos que se pueden obtener del sintetizador, con lo que la parametrización de los sonidos se puede realizar más fácilmente.

CONCLUSIONES

Se ha presentado un sistema que permite producir voz a partir de un texto escrito con vocabulario ilimitado. El sistema emplea un método de síntesis por regla que tiene en cuenta gran número de factores para producir una voz de calidad buena, tanto en cuanto a su naturalidad, como a su inteligbilidad.

Este sistema descrito se ha realizado mediante una arquitectura digital que permite la ejecución de todo el proceso formando un sistema autónomo y con respuesta en tiempo real.

BIBLIOGRAFÍA

1 . Olabe, J.C., et al. Real Time Text-to-Speech Conversion System for Spanish, Proc. of the IEEE int. Conf. on Acoust., Speech and Signal Processing. San Diego EE.UU. 1984.

2 . Klatt,D.H., Software for a Cascade/Parallel Formant Synthesizer, J.Acoust.Soc, Am., Vol 65, Nº 3, pp. 971-995, Mar. 1980.