INTRODUCCIÓN A LA CONVERSIÓN TEXTO-VOZ

Miguel Ángel Rodríguez Crespo

En este artículo se presenta una de las áreas de la Tecnología del Habla: la conversión texto-voz. Sin profundizar demasiado en los fundamentos matemáticos de los sistemas de conversión texto-voz, se intenta ofrecer una visión genérica de los mismos (tanto en problemas como en soluciones), si bien se hará partiendo de la descripción del sistema de conversión texto-voz desarrollado en la División de Servicios de Tratamiento del Habla de Telefónica I+D.

1. Introducción

La voz es uno de los principales recursos comunicativos del ser humano. A partir del desarrollo de medios de registro, transmisión y reproducción de la voz, ha constituido además un campo de estudio y desarrollo tecnológico. Con la evolución de los ordenadores digitales ha aumentado el interés en hacer de ella un medio de comunicación entre el ser humano y la máquina.

La Tecnología del Habla es un conjunto de conocimientos y técnicas de procesamiento de la señal de voz, cuyo objetivo es aumentar la capacidad de comunicación de la misma: permitir el control y acceso a la información de ordenadores, la comunicación entre personas que no tienen un idioma común, mejorar la comunicación en ambientes muy ruidosos, etc.

Destaca su carácter multidisciplinar. Se recurre a conocimientos y técnicas propios de disciplinas como Acústica, Fisiología, Lingüística, Procesado de Señal, Inteligencia Artificial, Teoría de la Comunicación y de la Información, Informática, ...

Además de la conversión texto-voz, otras áreas de trabajo en Tecnología del Habla son:

- Codificación de voz.
- Reconocimiento de voz.
- Reconocimiento de locutores.
- Otras técnicas relacionadas: reducción de ruidos, análisis de la voz con fines médicos y patológicos, desarrollo de prótesis auditivas y de ayuda en la locución, etc.

Desde su creación en 1988, en Telefónica I+D existe un grupo de trabajo en Tecnología del Habla, dedicado fundamentalmente a técnicas de reconocimiento de voz y de

I.S.S.N. 1132-0265

conversión texto-voz, y a la integración de las mismas en aplicaciones destinadas a ofrecer nuevos servicios en la red telefónica.

La estructura de este artículo consta de tres bloques principales:

- La voz: Se presentará una breve descripción de las características de la señal de voz, desde un punto de vista fisiológico y acústico. También se hará una pequeña introducción a los fundamentos de la Codificación de Voz, técnica básica para todas las de tratamiento del habla, y particularmente importante para la conversión texto-voz.
- Conversión texto-voz: Se describirá el problema de la conversión texto-voz, y se presentará la estructura general de un sistema de este tipo.
- Aplicaciones: Se presentarán ejemplos de aplicaciones de las Tecnologías del Habla (fundamentalmente del reconocimiento de voz y de la conversión texto-voz). Se hará especial énfasis en las aplicaciones basadas en la red telefónica.

2. La voz

Desde el punto de vista acústico, la señal de voz consiste en una onda sonora (condensaciones y rarefacciones del aire) que se propaga en la misma dirección de la

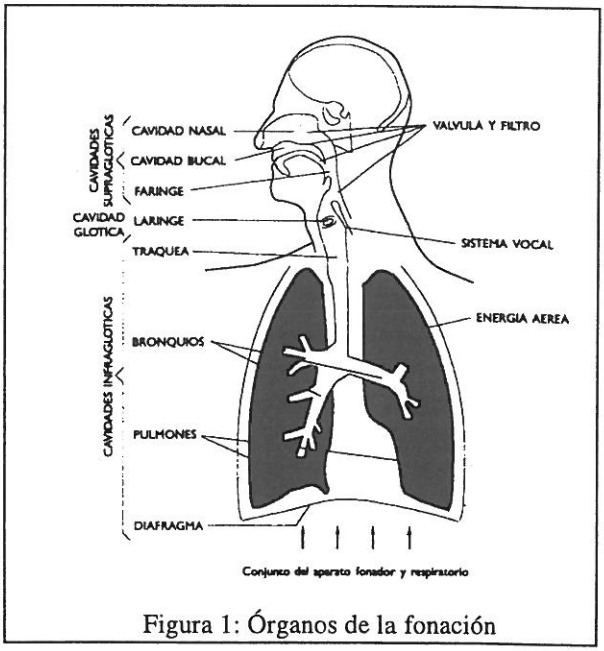

Figura 1: Órganos de la fonación

vibración. El origen de esta onda está en una corriente de aire, procedente de los pulmones, y modulada por los órganos de la laringe y del tracto vocal.

En la parte superior de la laringe hay dos membranas, llamadas cuerdas vocales. Éstas 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 emitir 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. Entonces vuelve a caer la presión del aire, y la glotis se cierra de nuevo. De esta manera vibran las cuerdas vocales, y se forma una variación cuasi-periódica en el volumen del aire espirado, que es el origen del sonido. Esta corriente pasa al tracto vocal, donde adquiere muchas de las características diferenciadoras de los fonemas, a través de un filtrado espacial.

El tracto vocal se abre desde la laringe hasta los labios y las ventanas nasales. Se puede modelar como un tubo acústico, con una serie de resonancias, llamadas formantes, y de antirresonancias. Mantiene la energía de la señal glotal en frecuencias próximas a las de los formantes, y la atenúa en las antirresonancias. Este filtrado realizado por el tracto vocal es lo que se asocia al timbre de los sonidos.

La posición de los formantes queda determinada principalmente por la forma de articulación, sobre todo el movimiento de la lengua y de los labios.

Las antirresonancias se producen al bajarse el velo del paladar. El aire sale por las fosas nasales, y la cavidad bucal actúa como una cavidad acoplada.

Esta es la forma de generación de los sonidos sonoros. En el caso particular de las vocales, se permite el paso del aire sin restricciones por la cavidad bucal. Para las consonantes sonoras se restringe el flujo de aire en algún punto.

En las consonantes sonoras fricativas, esta restricción es lo bastante estrecha para crear turbulencias en la corriente de aire, que son el origen del sonido.

En las consonantes sordas, el aire pasa entre las cuerdas vocales sin hacerlas vibrar. En algún punto se produce un estrechamiento que provoca una turbulencia (para las fricativas), o una oclusión total y una súbita liberación de la presión (oclusivas).

La disposición de todos los órganos se mantiene aproximadamente estable durante la realización de cada sonido. Además se produce un proceso de coarticulación o acomodo hacia la articulación del siguiente sonido. A pesar de eso, la señal mantiene unas ciertas características durante los intervalos correspondientes a cada sonido (unas decenas de milisegundos).

La señal de excitación (la que pasa a través del tracto vocal) va a determinar principalmente las características personales de la voz, así como la entonación (tono), que transmite información sobre el estado de ánimo y las intenciones del locutor, y sobre la estructura del mensaje.

La estructura de formantes (envolvente espectral) fundamentalmente va a transmitir información sobre la naturaleza de los alófonos. Los alófonos son las realizaciones de las

idealizaciones de los sonidos (fonemas) compartidas por los hablantes de una misma leng
y que tienen significación lingüística y son relevantes para la comunicación (p. ej.,

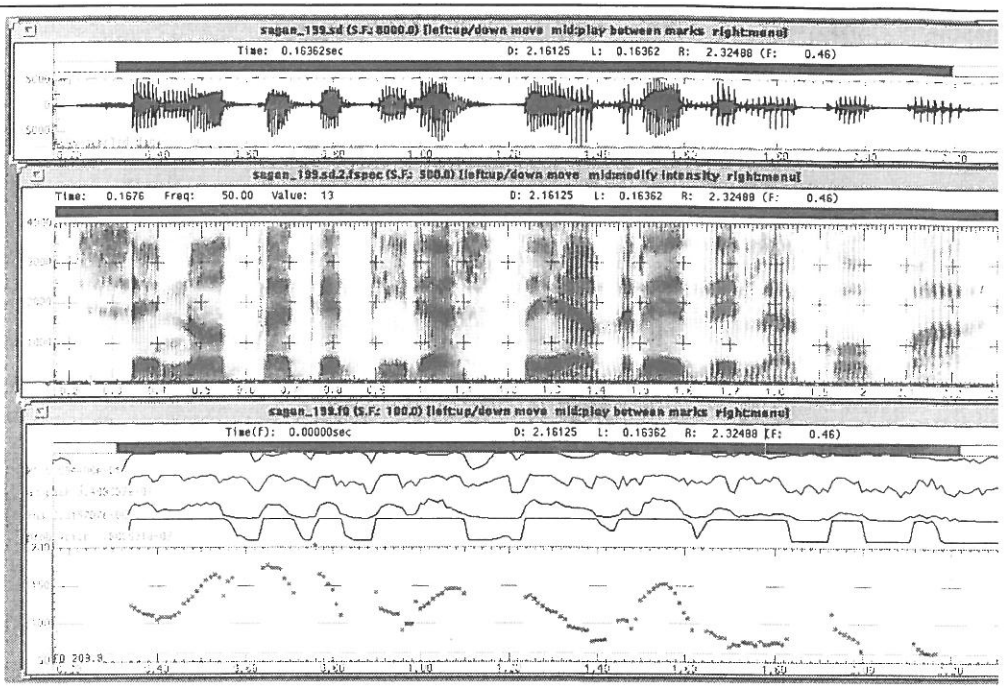

Figura 2: Forma de onda, espectrograma y contorno de frecuencia fundamental de la fra

Se nota que tiraste adrede la copa.

castellano una vocal puede nasalizarse cuando se encuentra entre sonidos nasales, pero es
información no es relevante para el oyente, como podría serlo en francés).

2.1. *Codificación de voz*

La codificación es un campo básico dentro de la Tecnología del Habla. Desde la
técnicas más sencillas, que permiten el almacenamiento y acceso de los ordenadores a l
señal de voz, hasta las más complejas que no sólo permiten un importante ahorro d
recursos, sino que también ofrecen representaciones y modelos útiles a todas las áreas d
trabajo de la Tecnología del Habla.

La señal de voz es una señal limitada en banda, aproximadamente entre 20 Hz y 2
KHz. Sin embargo, la mayor parte de la energía se concentra por debajo de 2 KHz, y s
asegura gran parte de la inteligibilidad con un ancho de banda entre 300 y 3400 Hz (salv

en el caso de algunas consonantes fricativas, con rasgos distintivos hasta 6 ó 7 KHz). El margen dinámico de la voz es muy amplio, tanto entre locutores (unos 20 dB) como para un mismo locutor (hasta 40 dB de diferencia entre zonas sonoras y sordas).

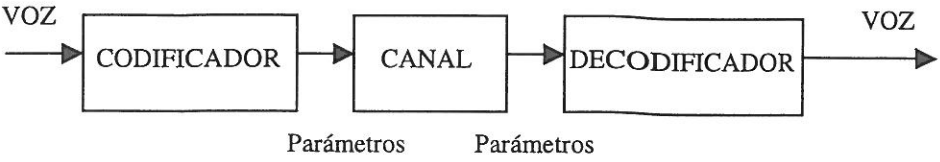

Figura 3: Esquema de un sistema de codificación-descodificación.

El estándar de codificación para calidad telefónica es de 64 KBit/s (PCM). Esto quiere decir que para almacenar un segundo de voz son necesarios 8000 bytes (se muestrea la señal de voz 8000 veces por segundo, y cada muestra se codifica en un octeto). Otras técnicas explotan la gran redundancia de la señal de voz para reducir a 4 el número de bits por muestra, mediante adaptación del tamaño del escalón de cuantificación y predicción del valor de las muestras a partir de los valores anteriores (ADPCM a 32 Kbit/s).

Sin embargo, si renunciamos a recuperar la forma de onda, se puede reducir más el régimen binario. A título orientativo, cuando se lee un texto a una velocidad normal (unas 150 palabras/minuto), la información presente en el texto se transmite a una velocidad de unos 75 bit/s.

Ya se ha señalado que la señal de voz, aunque muy distinta para cada uno de los sonidos, mantiene sus características durante periodos relativamente grandes, correspondientes aproximadamente a los alófonos que se están pronunciando. Se puede intentar encontrar algunas de las características que definen estos sonidos, y transmitirlas sólo al ritmo que éstas cambian, cada pocas decenas de milisegundos.

Para definir y extraer esas características es preciso desarrollar un modelo de la señal. El modelo más empleado y aceptado está muy relacionado con la descripción que se ha hecho de la generación de la voz.

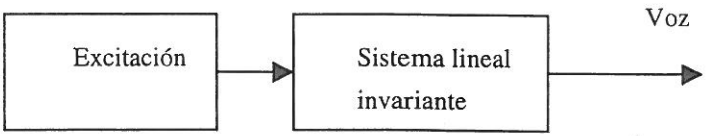

Figura 4: Esquema de un modelo de producción de la señal de voz

Se divide la señal de voz en intervalos, en los que mantiene sus características, y supone que las propiedades de la señal en cada intervalo se extienden indefinidamente en tiempo. En cada uno de estos intervalos se separan e independizan las características de excitación y las del tracto vocal y la radiación.

En el dominio de la frecuencia:

$$V(\omega) = H(\omega) \cdot G(\omega)$$

donde:

$V(\omega)$ es la transformada de la señal de voz

$G(\omega)$ es la transformada de la señal de excitación

$H(\omega)$ es la función de transferencia del tracto vocal (jumto con el efecto de radiaci de los labios), todo ello para un intervalo de tiempo dado y una configuraci concreta de los órganos fonadores.

Los métodos de codificación que utilizan propiedades de la señal de voz en vez intentar regenerar la forma de onda, suelen denominarse genéricamente VOCODERS. L parámetros con los que trabajan suelen estar relacionados con una descripción del espectr

3. Conversión texto-voz

3.1. *Definición*

La conversión texto-voz es la generación, por medios automáticos, de la secuencia sonidos que produciría una persona al leer un texto cualquiera en voz alta.

Algunos aspectos destacables de esta definición son:

- La generación debe hacerse de forma automática, sin mediar correcciones ajustes "a mano" por parte de un operador en ninguna de las etapas d proceso.
- La meta de la conversión texto-voz es producir habla emulando, en lo posibl el modo en que un ser humano lee. No bastará que se pueda entender lo que conversor dice (inteligibilidad), sino que además debe ser apreciado p oyentes humanos como semejante a un hablante humano (naturalidad). Est último aspecto es el gran reto de la conversión texto-voz.
- Aunque depende del tipo de aplicación, en el caso más general el convers sólo tendrá como entrada los datos que se encuentren en un texto arbitrari Además, debe ser capaz de tratar todos los fenómenos (abreviaturas, número vocablos extranjeros, etc) que aparecen en un texto corriente.

Conviene aclarar que la conversión texto-voz no es síntesis de voz a partir de concept Es decir, la conversión texto-voz siempre trabaja a partir de un texto previamente escrito, no incluye la capacidad de generar el texto respondiendo a condiciones variables n previsibles de antemano, a diferencia de como hacen los ordenadores parlantes qu

aparecen en las películas de ciencia-ficción. Actualmente, algunas técnicas de proceso de lenguaje natural y de inteligencia artificial trabajan en este sentido.

3.2. *Estructura general*

Uno de los problemas de la lectura de texto es que no es un proceso fácilmente divisible en tareas totalmente separables de manera secuencial en el tiempo. Por ejemplo, el género que debe tener la expansión de un número no se puede determinar hasta que no se haya hecho algún tipo de análisis sintáctico para determinar las relaciones de dependencia. Sin embargo, el módulo que realice este análisis seguramente no podrá manejar la variedad de un texto arbitrario, y necesita que se haya realizado previamente un cierto preproceso, una de cuyas tareas típicas es la expansión de abreviaturas y números.

Sin embargo, las estructuras no secuenciales son muy difíciles de realizar y aumentarían la complejidad de un sistema ya de por sí bastante complejo. Por ello, la mayoría de los sistemas de conversión texto-voz (entre ellos el de Telefónica I+D) adoptan una estructura secuencial, en el que las restricciones impuestas por la secuencialidad se intentan suplir con un uso inteligente de la información que comparten los distintos bloques y módulos componentes.

La estructura del conversor texto-voz de Telefónica I+D es la que se encuentra en la figura 5.

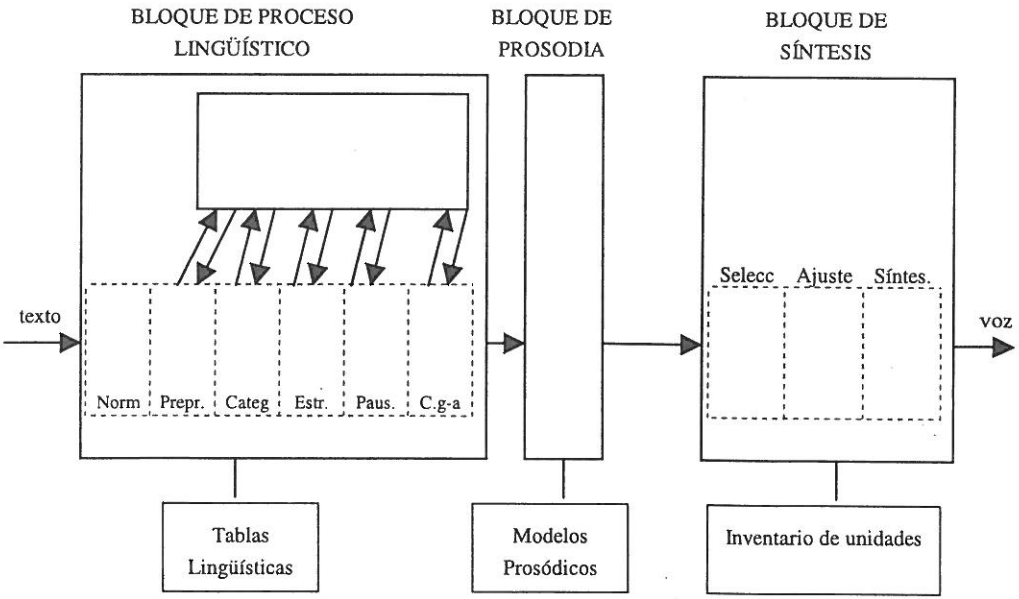

Figura 5: Esquema de un sistema de conversión texto-voz con una estructura secuencial

La mayor parte de los sistemas de conversión texto-voz responden a la estructur general de esta figura. Las mayores diferencias respecto a otros sistemas de conversió texto-voz de otros idiomas aparecerán en el bloque de proceso lingüístico, debido entr otras razones a las particularidades del español respecto a otros idiomas.

3.3. *Bloque de proceso lingüístico*

Para leer un texto es preciso saber qué sonidos hay que producir, y cómo producirlo: Los objetivos de este bloque se traducen en obtener la cadena de alófonos correspondient al texto de entrada e información para el correcto modelado del bloque generador d prosodia.

La tarea de generar toda esa información para un texto sin restricciones es demasiad compleja para abordarla globalmente. Por eso es necesario fraccionar el problema en un serie de problemas más abordables.

A continuación, describiremos someramente cada uno de los módulos que componen e bloque de proceso lingüístico del conversor texto-voz, presentando las tareas que realizan.

3.3.1. Módulo normalizador

Su tarea principal es detectar y reunir un conjunto de caracteres en el texto de entrada Este conjunto de caracteres forma la unidad de trabajo que se tomará como común a todo los módulos del proceso lingüístico. Además normaliza la escritura de las palabra detectadas, para simplificar comparaciones y consultas en el resto de los módulos.

La unidad de trabajo que se ha definido es la frase. La frase abarcará una serie d palabras, hasta que se encuentre algún carácter o combinación de caracteres que marque e fin de frase (punto de fin de frase, cierre de interrogación, dos puntos, etc).

3.3.2. Módulo de preproceso

Aunque el módulo normalizador ya ha decidido y agrupado lo que es una frase separado los elementos que la constituyen y realizado una primera normalización de su escritura, la variabilidad del texto de entrada es todavía demasiado grande. El resto de lo módulos sólo van a poder manejar palabras y signos ortográficos, y hay que reducir a est representación cualquier otra cosa (por ejemplo: abreviaturas, números, representaciones d fechas, representaciones horarias, ...).

Por tanto la principal tarea de este módulo es la de reducir la complejidad (variabilidad del texto. Además se realizan otra serie de tareas, como son la silabicación, acentuació fonética y tratamiento de los acrónimos y secuencias impronunciables.

3.3.3. Análisis del texto.

En nuestro caso el análisis de texto se compone de un módulo categorizador y un módulo estructurador (análisis sintáctico).

3.3.3.1. Módulo categorizador

La principal tarea del módulo es la de asignar categorías a las palabras. Las categorías que se asignan no son exactamente categorías gramaticales, sino un conjunto de códigos que en muchos casos se corresponden con verdaderas categorías gramaticales, pero que en otros se han escogido por criterios prácticos.

3.3.3.2. Módulo estructurador

Para la generación de una prosodia verdaderamente natural sería necesario que el sistema tuviese información del significado de la frase. Actualmente no es posible realizar análisis semántico y pragmático de texto fuera de entornos muy restringidos. Por eso se utiliza la estructura sintáctica como un pálido reflejo de la estructura semántica de la frase. Se necesita un módulo capaz de generar esa estructura sintáctica.

Aunque el análisis sintáctico del texto en sentido estricto es un problema abordable, el diseño de este módulo presenta algunas peculiaridades: ha de ser capaz de tratar de alguna manera cualquier frase que reciba a su entrada, aunque no sea gramatical (bien porque la gramática que se ha construido no es capaz de recoger todos los fenómenos del idioma, o bien porque la frase de entrada no responde a las normas del lenguaje); y ha de ser lo suficientemente simple como para funcionar en tiempo real dentro de la estructura del conversor.

3.3.4. Módulo pausador

Al leer un texto, es muy frecuente que los lectores hagan más pausas que las que vienen marcadas por los signos ortográficos. Este hecho normalmente viene motivado por la necesidad que tiene el hablante de recuperar el aliento. Además, la inserción de pausas, junto con la utilización de la prosodia, le permiten transmitir parte del contenido del mensaje, y facilitar así la comprensión para el oyente.

Evidentemente, un conversor texto-voz no necesita recuperar el aliento cuando procesa un texto, pero daría una desagradable sensación de ahogo al oyente si pronunciara un fragmento de texto largo sin ningún signo ortográfico. Hemos comprobado que fragmentos de texto de este tipo aparecen frecuentemente en noticias de periódicos, artículos de opinión, novelas, ... por lo que no puede decirse que siempre sean producto de un incorrecto (o al menos escaso) uso de los signos de puntuación.

La tarea principal del módulo pausador es insertar pausas automáticamente cuando el texto que se encuentra entre dos pausas marcadas ortográficamente produciría una secuencia de habla demasiado larga.

3.3.5. Módulo conversor de grafemas a alófonos

Este módulo se ocupa de obtener la secuencia de alófonos (de un alfabeto previamen definido) correspondiente a la secuencia de letras de una frase dada. Para ello emplea parte de la información lingüística obtenida en los módulos anteriores.

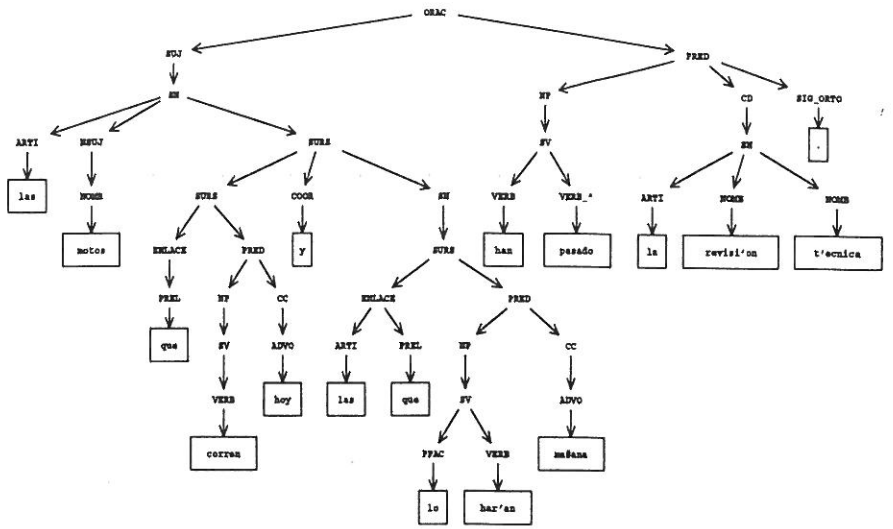

Figura 6: Resultado del análisis de la frase: ***Las motos que corren hoy y las que lo mañana han pasado la revisión técnica***. El analizador descubre correctamente la enumer de proposiciones de relativo.

3.4. *Bloque generador de los parámetros prosódicos*

Entre las definiciones del Diccionario de la Real Academia Española para el término prosodia se encuentra la siguiente:

> Prosodia: Parte de la fonología dedicada al estudio de los rasgos fónicos que afecta a unidades inferiores al fonema, como las moras, o superiores a él, como las sílaba u otras secuencias de la palabra u oración.

En Tecnología del Habla, se suele emplear este término para referirse al conjunto de parámetros suprasegmentales de la señal de voz; es decir, los que no quedan determinados por la identidad de los sonidos que hay que pronunciar. Habitualmente se trata de la frecuencia fundamental (entonación), la duración (cantidad) y la energía (intensidad).

La información prosódica refleja elementos lingüísticos (interrogación, exclamación, pausas, acentos, ...) y elementos no lingüísticos (características del locutor, estado de ánimo y actitud del mismo ante lo que se está leyendo, ...). Sólo los primeros pueden ser controlados directamente por el conversor texto-voz. Para los segundos, lo único que se puede hacer es generar distintos modelos para cada situación, y que sea el operador el que indique qué modelo utilizar en cada caso.

Aunque los tres parámetros considerados habitualmente para la prosodia (duración, frecuencia fundamental y energía) interactúan para producir una dicción natural, en una primera aproximación se pueden obtener de manera separada, suponiendo que son independientes.

3.4.1. Duración

Por duración nos referimos al intervalo de tiempo durante el que se mantiene cada alófono que hay que pronunciar (entre los alófonos se incluyen también los silencios correspondientes a las pausas).

La duración contribuye a marcar la acentuación fonética y la caracterización de los grupos fónicos (y a la naturalidad en general). Va a repercutir sobre la velocidad del discurso.

Este parámetro depende no sólo de la estructura de la frase (suprasegmental), sino también de la naturaleza de cada uno de los alófonos (segmental). Esto dificulta su análisis y modelado.

3.4.2. Frecuencia fundamental

La generación del contorno de evolución temporal de la frecuencia fundamental (F0) es quizá el factor que más influye en proporcionar naturalidad a la voz sintética.

La frecuencia fundamental se emplea para marcar los acentos fonéticos y los diferentes tipos de frases (enunciativas, interrogativas, ...). Pero además de marcar este tipo de efectos que, de algún modo, podemos decir que se encuentran reflejados en el texto, el contorno de F0 también refleja otros efectos que, de ningún modo, están en el texto, como pueden ser las características individuales del locutor, y la sensación más o menos "cantarina" del discurso.

3.4.3. Energía

La energía es el parámetro de la prosodia menos considerado en los sistemas de conversión texto-voz. Aunque es innegable la presencia de la energía en el discurso, y su

variación a lo largo del mismo, y aunque algunos lingüistas la describen como el factor má importante para marcar el acento, se ha comprobado que su repercusión en la prosodia n es relevante, al menos desde el punto de vista psicoacústico.

3.5. *Bloque de síntesis de voz*

La misión de este bloque es generar sonidos tan similares a la voz como sea posible presentando un alto grado de flexibilidad en cuanto a su capacidad para ser controlado, d modo que se pueda variar la realización de los sonidos.

La información de entrada a este módulo incluye la secuencia de alófonos que hay qu generar, y los datos de prosodia (típicamente, duración de los alófonos, contorno d frecuencia fundamental, y contorno de energía o amplitud).

Hay dos enfoques que, en cierto modo, determinan el tipo de sintetizador que se emplee

El primero de ellos intenta modelar el mecanismo de producción de la voz, con mayor menor detalle. Este enfoque ha dado origen a dos tipos de sintetizadores:

- Sintetizadores articulatorios: su idea básica consiste en controlar un modelo de aparato fonador, de un modo semejante a como lo hace el cerebro. Lo parámetros internos de control de este tipo de modelos son la posición de lo distintos órganos articulatorios y las leyes que rigen su movimiento. Po ejemplo, para producir una "u" hay que redondear y proyectar los labios retraer la lengua, tensar las cuerdas vocales y expulsar aire a través de ellas co una cierta presión, para que vibren con el tono deseado. El problema principa de los modelos articulatorios es, por un lado, la enorme cantidad de parámetro internos de control que precisan y la dificultad de coordinarlos y derivarlos d los parámetros de control disponibles a la entrada del sintetizador; y, por otr lado, la gran cantidad de información que se necesita obtener analizando (en u espacio tridimensional) la posición y el movimiento de los órgano articulatorios de una persona que habla normalmente.
- Sintetizadores de formantes: en este tipo de sintetizadores se hace un simplificación del aparato fonador (el modelo de producción de voz presentad anteriormente). Consisten en realizar un filtrado de una señal de excitación introduciendo resonancias (y antirresonancias). Puesto que los alófonos s distinguen, entre otras cosas, por el tipo de excitación y por los valores d frecuencia central y ancho de banda de sus resonancias, un sintetizador po formantes asocia a cada alófono un determinado conjunto teórico de eso parámetros, y así se puede hacer la síntesis mediante varios filtros de segund orden, conectados en serie y/o en paralelo. Este modelo también necesit determinar la evolución del conjunto de parámetros de un alófono al siguiente de forma que el sonido siga resultando natural. Estos sintetizadores tambié reciben el nombre de sintetizadores "por reglas", pues tienen información sobr las características asociadas a cada alófono, y reglas que controlan cómo s alcanzan y realizan esas características. Los sintetizadores por formantes ha

sido y son muy utilizados en conversión texto-voz. Son modelos poco costosos computacionalmente y que no precisan mucha memoria.

El segundo enfoque es el que intenta modelar la señal de voz, en lugar de modelar el mecanismo de producción de la misma. Este último enfoque podría denominarse "modelo de señal", en contraste con el "modelo de sistema" del enfoque anterior.

- Sintetizadores por concatenación de unidades. Se basan en tener un conjunto de pequeños segmentos de voz tomados de un hablante, que se van concatenando para formar el discurso deseado. Naturalmente, la concatenación debe ser controlada, de manera que, por un lado, se eviten discontinuidades y sonidos anómalos y, por otro lado, se puedan variar los elementos prosódicos respecto a aquellos que originalmente tenían los segmentos de voz escogidos. Para decidir el tamaño y número de estas unidades hay un compromiso entre la calidad de la voz que se quiere sintetizar, y limitaciones de memoria de datos. La posibilidad más inmediata es tener grabados únicamente cada uno de los alófonos. Así sólo se necesitaría tener grabadas entre 30 y 50 unidades, según el idioma. Por contra, habría que realizar el "pegado" de las unidades en zonas acústicamente inestables y comprometidas, como son las transiciones entre alófonos. Por esto, la aproximación más común es trabajar con demialófonos, que son unidades que recogen parte de un alófono, la zona de transición, y parte del siguiente alófono. Así, el "pegado" se hace en una zona acústicamente más estable, pero el número de unidades aumenta a unas 300 ó 400 (no todas las combinaciones de alófonos son posibles en un idioma).

En cuanto a la forma de tener almacenadas las unidades de esa colección, cualquier sistema de codificación es apropiado, siempre que tenga la suficiente flexibilidad para permitir controlar los parámetros de la prosodia. Se pueden emplear técnicas de predicción lineal (LPC, MPLPC) o de solapamiento y suma síncronas con la frecuencia fundamental (FD-PSOLA, TD-PSOLA). Éstas han sido hasta ahora las técnicas más empleadas. Recientemente están apareciendo otras, como representación por wavelets o modelado sinusoidal.

4. Aplicaciones y sistemas

La Tecnología del Habla ha alcanzado el grado de madurez necesario para soportar una gran variedad de aplicaciones. A pesar de las limitaciones actuales, es posible su utilización para realizar aplicaciones destinadas a ofrecer nuevos servicios o mejorar servicios ya existentes.

Deben cuidarse los detalles que hagan cómodo y agradable el diálogo con los usuarios, dado que de este diálogo depende en gran medida la aceptación (o el rechazo) de una determinada aplicación, y por extensión, de toda la tecnología que involucra. En este sentido los estudios sobre Factores Humanos en los servicios de telecomunicación deben contribuir de forma importante a paliar las limitaciones actuales de la Tecnología del Habla.

Dado el amplio campo de aplicación de la Tecnología del Habla, se ha considerad necesario clasificar sus aplicaciones en tres grupos diferentes, cada uno de los cual impone una serie de requisitos comunes a la vez que da lugar a conjuntos de aplicacion semejantes. Estos grupos son:

- Aplicaciones locales.
- Respuesta vocal interactiva.
- Automatización de sistemas telefónicos.

4.1. Aplicaciones locales.

El sistema tradicional de interfaz hombre-máquina basado en la utilización de teclado pantalla impone serias restricciones a la movilidad de los usuarios, que de hecho s encuentran virtualmente atados a la máquina, además de exigir la dedicación prácticamen exclusiva de las facultades del tacto y de la vista. Haciendo uso de la Tecnología del Habl en particular del reconocimiento del habla y de la conversión texto-voz, es posible realiz interfaces utilizando el lenguaje hablado y, como consecuencia, eliminar las restriccione anteriores. Además, cuando sea posible la utilización del lenguaje natural, el usuario n necesitará recordar complicadas y engorrosas secuencias de operaciones.

Este grupo de aplicaciones son en general monousuario, con lo que es posible utiliz reconocimiento dependiente del locutor. Dado que la interacción con el usuario se realiz en local, no es necesario considerar las limitaciones que imponen las redes d comunicación, con lo que las dificultades para la realización de estos reconocedores se ve notablemente reducidas.

En el campo de la ayuda a discapacitados, la Tecnología del Habla puede ser la soluci para los problemas de comunicación y movilidad de este colectivo.

El reconocimiento de voz se emplea por personas con discapacidades motrices par controlar equipos en general (sillas de ruedas, electrodomésticos, ...).

La conversión texto-voz se utiliza para que los ciegos tengan acceso a la informaci textual residente en ordenador, y para que personas incapacitadas para hablar pueda sintetizar voz utilizando un teclado, que puede ser un teclado especial cuando l discapacidad afecta además a las facultades motrices.

4.2. *Respuesta vocal interactiva.*

Cuando se piensa en aplicaciones cuyo objetivo es difundir o capturar informació involucrando a un gran número de usuarios, es imprescindible contar con el concurso de la redes de telecomunicación como vehículo de acceso a la información. En particular, la re telefónica básica es el procedimiento de acceso ideal dada su gran difusión, aunque s capacidad para transportar datos es reducida, ya que maneja un ancho de banda limitad (300 a 3400 Hz), al haber sido concebida para transmitir voz.

Esta limitación se ha solventado utilizando terminales de datos equipados con modem, aunque su difusión dista mucho de la de los teléfonos convencionales. Es pues muy interesante disponer de recursos de reconocimiento de voz, o como alternativa, detección de tonos multifrecuencia, para la comunicación usuario-máquina, y de conversión texto-voz o mensajes digitalizados (codificación de voz) para la comunicación máquina-usuario, ya que estas señales pueden ser manejadas por la red telefónica básica.

Las aplicaciones de respuesta vocal interactiva pueden clasificarse en:

- Difusión de la información: Su objetivo es la difusión de información de interés general o particular (noticias generales, información meteorológica, horarios de trenes, aviones,...)
- Captura de información: Su objetivo es la captura de información de carácter general o particular para ser procesada con distintos fines (encuestas, notificaciones de avería, citas,...).
- Mensajería vocal: Su objetivo es permitir la comunicación entre los usuarios cuando no es posible la comunicación directa. Los mensajes vocales quedan almacenados en un buzón, y pueden ser recuperados posteriormente por la persona a la que van dirigidos.

4.3. *Automatización de sistemas telefónicos.*

Tienen como objetivo realizar la interfaz de usuario entre el abonado y la red telefónica empleando el lenguaje hablado. Las aplicaciones son soportadas por la propia red telefónica, con lo que los usuarios pueden seguir utilizando los mismo terminales telefónicos.

- Marcación vocal: Esta aplicación permite a los usuarios realizar la marcación pronunciando el número o el nombre de la persona llamada. La marcación por número llamado resulta ventajosa en los casos de usuarios con limitaciones motrices o para los usuarios de la telefonía móvil cuando conducen un vehículo. La marcación por nombre hace además innecesario el conocimiento del número telefónico llamado.
- Manejo de facilidades suplementarias: El creciente número de servicios suplementarios hace necesaria la utilización de un número cada vez mayor de códigos multifrecuencia necesarios para su manejo. Esta aplicación tiene por objetivo reemplazar los códigos multifrecuencia normalmente empleados por comandos vocales mucho más fáciles de recordar.
- Cobro alternativo automático: Esta aplicación tiene como objetivo automatizar el tratamiento de las llamadas a cobro revertido, a crédito o con cargo a terceros.

5. Para terminar

En el capítulo LXII de la segunda parte de "El Quijote" se refiere el siguiente episodio:

"Otro día le pareció a don Antonio ser bien hacer la experiencia de la cabeza encantada, y cc don Quijote, Sancho, y otros dos amigos ...

... El primero que se llegó al oído de la cabeza fue el mismo don Antonio, y díjole en vc sumisa, pero no tanto que de todos no fuese entendida: - Dime, cabeza, por la virtud que en ti s encierra: ¿qué pensamientos tengo yo agora? Y la cabeza respondió, sin mover los labios, con vc clara y distinta, de modo que fue de todos entendida, esta razón: - Yo no juzgo de pensamiento Oyendo lo cual todos quedaron atónitos, y más viendo que en todo el aposento ni al derredor de mesa no había persona humana que responder pudiese ...

... ¡Aquí sí que fue el admirarse de nuevo; aquí sí que fue el erizarse los cabellos a todos, c puro espanto! Y apartándose don Antonio de la cabeza, dijo: - Esto me basta para darme entender que no fui engañado del que te me vendió, ¡cabeza sabia, cabeza habladora, cabez respondona, y admirable cabeza! Llegue otro y pregúntele lo que quisiere ...

... El último preguntante fue Sancho, y lo que preguntó fue: - ¿Por ventura, cabeza, tendré otı gobierno? ¿Saldré de la estrecheza de escudero? ¿Volveré a ver a mi mujer y a mis hijos? A que le respondieron: - Gobernarás en tu casa, y si vuelves a ella, verás a tu mujer y a tus hijos, dejando de servir, dejarás de ser escudero. - ¡Bueno par Dios! -dijo Sancho Panza-. Esto yo me dijera. No dijera más el profeta Perogrullo. - Bestia -dijo don Quijote-, ¿qué quieres que respondan? ¿No basta que las respuestas que esta cabeza ha dado correspondan a lo que se pregunta? - Sí basta -respondió Sancho-; pero quisiera yo que se declarara más y me dijera más".