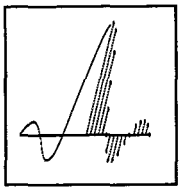

TECNOLOGÍAS DEL HABLA: CONVERSIÓN DE TEXTO A VOZ

Antonio Bonafonte Cávez

Profesor Titular en el Grupo de Procesado de Señal.
Departament de Teoria del Senyal i Comunicacions, UPC.
e-mail: antonio@gps.tsc.upc.es

TECNOLOGÍAS DEL HABLA

Aunque el procesado de voz ha constituido un área de intensa investigación durante varias décadas, ha sido durante ésta última cuando las tecnologías han alcanzado el grado de desarrollo suficiente para afrontar un amplio espectro de aplicaciones. En esta comunicación se pretende describir qué es, cuál es la aplicación y cómo funcionan los sistemas que convierten texto en voz. Sin embargo, esta área se complementa con otras áreas del procesado de voz, por lo que en primer lugar describiremos brevemente las distintas áreas tratadas por el procesado de voz.

Codificación de voz y audio

El objetivo de la codificación de fuente (bien sea voz, imagen o datos en general) es comprimir la información de tal forma que pueda ser transmitida o almacenada de forma eficiente. Para la señal de voz, los codificadores más eficientes utilizan un modelo de producción de voz consistente, por una parte, en una excitación que modela el aire que fluye desde los pulmones y la vibración de las cuerdas vocales y, por otra, un filtro que representa las cavidad bucal y nasal. Por otra parte, tanto para la señal de voz como para la señal de audio, se utiliza un modelo del sistema de percepción auditiva que indica que componentes de la señal no se han de preservar puesto que no pueden oírse, al quedar enmascaradas por otras componentes de la señal de mayor energía situadas próximas, bien temporalmente, bien frecuencialmente. Los codificadores actuales son capaces de lograr tasas de compresión de 8 para señales de voz y de 5 para señales de audio sin que se aprecie, en la mayoría de los casos, una pérdida en la calidad.

Reconocimiento del habla

Los sistemas de reconocimiento del habla extraen la información del mensaje de la señal de voz. Es un área en la que ya han aparecido un gran número de aplicaciones que se multiplicarán en los próximos años. Para comparar los distintos sistemas se han de analizar sus especificaciones o prestaciones según distintos criterios:

· Locutores que se reconocen: los sistemas puede estar diseñados para uno o un grupo pequeño de usuarios, puede necesitar de un periodo de adaptación al usuario o bien pueden funcionar con cualquier usuario.

· Características de la señal, dependiendo del ruido, de la distorsión, etc. Aplicaciones típicas que requieren un buen comportamiento en estas situaciones son las relacionadas con comunicaciones móviles, línea telefónica, lugares públicos, etc.

· El tipo de habla: palabras aisladas, en donde sólo se pronuncia una palabra o se introducen pausas significativas entre palabras; habla continua; habla espontánea, donde se permiten las dudas, errores gramaticales, reformulación de frases, etc. propios del habla natural.

· Características del léxico: tamaño del vocabulario (decenas, centenas, miles o hasta decenas de miles de palabras), parecido entre palabras.

· Dominio semántico de la aplicación: dominios restringidos (como transacciones bancarias, etc.), dominios extensos (acceso a grandes bases de datos, dictado de informes médicos, etc.)

· Coste de la aplicación: ninguno de los sistemas existentes han sido diseñados para alcanzar las máximas prestaciones en todos los criterios anteriores, sino que, dependiendo de la aplicación, se centran en alguno de ellos a costa de limitar los objetivos en los otros. Por otra parte, se ha de tener en cuenta que muchas aplicaciones quedan limitadas por el coste del producto, valorado tanto en el coste del desarrollo como en la capacidad de cómputo y en la memoria requerida por cada sistema. Por tanto, en ocasiones es necesario utilizar sistemas sencillos y por tanto viables, aunque ello exija cierta rigidez y simplicidad a los diálogos con el usuario.

Síntesis de voz

Es la tecnología complementaria a la anterior para alcanzar una comunicación oral hombre máquina puesto que es el proceso de crear una réplica sintética de una señal vocal de forma que una máquina pueda transmitir información a una persona. Es el objeto de esta comunicación y será tratado en el siguiente apartado.

Identificación del locutor

Consiste en identificar al locutor a partir de una señal de voz. La mayoría de las aplicaciones están relacionadas con aspectos de seguridad: comprobación de identidad en operaciones *telebanco*, tarjetas de crédito inteligentes que necesitan una contraseña oral para identificar al propieta-

rio, etc. También encuentra un rango de aplicaciones como soporte a investigaciones policiales, etc.

Reconocimiento del idioma

Esta incipiente tecnología está encaminada a identificar el idioma en el que se expresa el locutor a partir de unos pocos segundos de habla. El objetivo es activar un sistema de reconocimiento automático del habla que permita al usuario comunicarse con la máquina en su propia lengua.

Traducción oral automática

El objetivo es transformar una señal de voz expresada en un idioma en señal de voz en otro idioma. Se han desarrollados varios prototipos que ofrecen buenas prestaciones en aplicaciones muy específicas. Estos sistemas suelen constar de un sistema de reconocimiento, con lo que se obtiene una representación textual de la señal de entrada, un traductor basado en técnicas de procesado del lenguaje natural y un sistema de conversión texto a voz. Simultáneamente, se están desarrollando algunos sistemas que integran el sistema de reconocimiento con el de traducción sin necesidad de disponer de la representación textual en el lenguaje de la señal de entrada.

Aunque en las áreas anteriores ha influido de forma muy significativa los procedimientos derivados del procesado de señal, la tecnología que se ha desarrollado es multidisciplinar: acústica, inteligencia artificial, procesado del lenguaje natural, estadística, fonética y en general lingüística, por citar sólo algunas, son disciplinas que participan en las áreas anteriores.

En todas las áreas anteriores, con excepción de la traducción –por el momento–, centra su actividad investigadora el *Grup de Tractament de la Parla*, del *Departament de Teoria de la Senyal i Comunicacions* de la *Universitat Politècnica de Catalunya*.

LA SÍNTESIS DEL HABLA

Utilidad de la síntesis del habla

Un sintetizador de voz es un dispositivo capaz de producir una réplica sintética de una señal vocal humana. El objetivo fundamental es posibilitar que los sistemas que han de proporcionar a las personas información de cierta complejidad, lo puedan realizar oralmente. En esta amplia definición quedan englobados muchos de los codificadores de voz, en los que, en base a ciertos parámetros obtenidos a partir de una señal de voz humana, se regenera –se sintetiza– una replica de dicha señal.

La comunicación oral presenta distintas ventajas. Al ser la forma más natural de comunicación entre personas resulta atractiva y es un valor añadido en las aplicaciones. Además, cuando la aplicación se desarrolla sobre un ordenador personal, la incorporación de dicha prestación representa un coste muy pequeño en comparación al producto. En este sentido, cada vez es más frecuente encontrar la respuesta oral en multitud de productos: expendedoras de tabaco, básculas, muñecos, etc. Y sobre ordenadores personales encontramos guías de museos, enciclopedias multimedia, maestros de ajedrez, etc. Incluso aparecen nuevas aplicaciones, como el aprendizaje de idiomas, en el que la señal vocal es el componente fundamental del propio producto. Otra de las ventajas de la respuesta oral es que libera al usuario de prestar atención a una pantalla o a un mensaje impreso pudiendo concentrarse en otras tareas. Por ejemplo, un operario puede recibir instrucciones o ayuda sin desviar la atención de su tarea.

Además, existen algunas aplicaciones en las que la información se ha de transmitir por un canal de comunicación apto para la voz pero no para el lenguaje escrito. El caso de mayor importancia es la red telefónica, donde mediante la respuesta oral, cualquier usuario con un teléfono es capaz de acceder a multitud de servicios proporcionados por el sistema proveedor del servicio. Por ejemplo, accesos a bases de datos o informaciones, lectores de facsímil o de correo electrónico, compras por teléfono, etc. requieren que el sistema utilice señal de voz. Análogamente, la respuesta oral es de gran utilidad para personas con deficiencias en la vista (invidentes, personas mayores, etc.). En estos casos, en algunos de los productos como básculas o termómetros, la voz pasa de ser una prestación complementaria a ser la prestación que posibilita el acceso al producto. Dispositivos más complejos, como lectores de documentos, pueden mejorar de forma significativa la calidad de vida de dichas personas, proporcionando fácil acceso a gran cantidad de información.

Hasta aquí hemos comentado las posibilidades de la síntesis de voz en cuanto a la comunicación hombre máquina. Sin embargo, la síntesis de voz también puede facilitar la comunicación entre personas, cuando alguna de ellas tiene afectada su aparato productor del habla. Es el caso, por ejemplo, del famoso físico S. Hawking . En esta dirección, el *Grup de Tractament de la Parla* está participando, aportando el sistema de síntesis, en un proyecto dirigido por *L'Institut de Educació Municipal de Barcelona* (concretamente por el Institut Municipal Pont del Dragó) y financiado por Inserso, para posibilitar la comunicación oral de personas que sufren parálisis cerebral y que actualmente basan su comunicación bien en el lenguaje escrito, bien en métodos alternativos (Bliss, Spc, etc.).

SISTEMAS DE PRODUCCIÓN DEL HABLA

Para dotar a un sistema de la capacidad de hablar se pueden utilizar distintos métodos. El primero de ellos, el más sencillo, consiste simplemente en registrar la señal de voz deseada y reproducirla en el momento adecuado. Anteriormente éste era el único modo y aparecieron aplicaciones como el contestador automático o los cursos de idiomas tradicionales. Actualmente la señal suele digitalizarse con lo que se consigue una mayor facilidad de acceso y un mayor rango de aplicaciones. El soporte puede ser magnético u óptico, necesario cuando el número de mensajes a reproducir es muy grande, o bien directamente en memorias de estados sólido, como es el caso algunos juguetes y de algunos contestadores telefónicos actuales. En

el caso de que la señal se almacene digitalizada, el uso de técnicas de codificación puede reducir enormemente la memoria necesaria.

Algunas aplicaciones de gran simplicidad requieren reproducir un número elevado de mensajes pero que pueden formarse fácilmente a partir de unas cuantas palabras o frases básicas. Este es el caso, por ejemplo, de los mensajes de información de algunas líneas de metro y ferrocarril en las que el mensaje es del tipo: "Pròxima estació: Palau Reial", o el método adoptado por el Servicio de Información 003 de Telefónica, para acoger al usuario y dar el resultado de las consultas. Los segmentos base ("pròxima estació", los dígitos, etc.) se concatenan de forma adecuada para formar el mensaje. En algunos casos, para disminuir el efecto de discontinuidad en las fronteras de los segmentos, o para mejorar la entonación, se combinan distintos locutores o se dispone de distintas versiones de cada palabra, por ejemplo, los dígitos en posición inicial, media o final.

Finalmente, el último método para dotar a un sistema de la capacidad de hablar es la conversión de texto en voz. Estos sistemas, para reproducir un mensaje, no parten de una señal vocal del mismo mensaje, sino que lo realizan encadenando pequeñas unidades acústicas (por ejemplo pronunciaciones de fonemas). Este método, que se desarrolla en el próximo apartado, es el indicado cuando el número de posibles mensajes a producir es ilimitado o desmesuradamente grande (por ejemplo el lector de documentos, de facsímil o de correo electrónico, la utilización por parte de personas con incapacidad en su aparato fonador, consultas a grandes bases de datos, etc.), o también cuando la información cambia frecuentemente en el tiempo. Incluso aunque los mensajes se mantengan relativamente estables en el tiempo y no sean muy numerosos, si se dispone de la información en forma textual, puede ser más económico y práctico disponer de un sistema de conversión de texto en voz. Por ejemplo, supongamos que un centro docente desea proporcionar información general del centro, fechas de matrícula, calificaciones de los estudiantes, etc; puede ser preferible desarrollar una aplicación que utilice la información escrita del centro, información que puede actualizarse sin más que modificar ficheros de texto, a registrar varias horas de señal de voz.

LA CONVERSIÓN DE TEXTO EN VOZ

Como se ha establecido en el apartado anterior, un sistema de conversión de texto en voz es aquel que es capaz de transformar un mensaje escrito, habitualmente un fichero de texto, en una señal de voz. Básicamente, el sistema requiere unos registros de señales orales relacionadas con unas unidades básicas (por ejemplo fonemas), que ha de concatenar siguiendo el texto de entrada. Para que el sistema realice una conversión de calidad, con la máxima inteligibilidad y naturalidad posible, es preciso modificar las unidades básicas, de forma que reflejen una entonación natural y apropiada al mensaje que se sintetiza, y suavizar las transiciones entre unidades, de forma que no se perciban discontinuidades entre ellas. Tanto en la definición de los sonidos elementales, en el análisis de la influencia de unos sonidos sobre sonidos adyacentes y en la definición de patrones prosódicos juegan un papel fundamental los estudios que realizan fonetistas y lingüistas. Los principales módulos de dichos sistemas pueden apreciarse en la figura 1 y serán descritos a continuación.

Normalización del texto

El objetivo de este módulo es proporcionar al sistema un texto escrito ortográficamente, sin caracteres de control o de formato, números, acrónimos o abreviaturas. Aunque puede parecer un módulo sencillo es extraordinariamente complejo.

Las siglas, son las letras, generalmente iniciales, de nombres de personas, empresas, asociaciones, publicaciones, etc., que se utiliza por comodidad en vez del nombre entero. Los acrónimos son las siglas que, debido a que siguen la estructura fonética de la lengua se pronuncian como palabras. Sin embargo, es frecuente leer las partes de los acrónimos que se adaptan a la lengua como palabras y deletrear el resto: CD ROM se suele leer como ce de rom, y no como ce de erre o eme; PSOE como pe soe. Algunas siglas se han de sustituir por la palabra que representan y no leer como palabra ni deletrear: CCOO, EEUU o EUA JJOO, etc. Por otra parte, no es fácil de distinguir una sigla de una palabra que aparezca en mayúsculas y menos si esta es extranjera. En ocasiones aparecen en las siglas cifras o símbolos: TV3, UpC, o Canal +. Además, la variedad de acrónimos y siglas que pueden aparecer en un texto escrito es enorme y el uso hace que aparezcan y desaparezcan con un gran dinamismo.

El tratamiento de las abreviaturas también es complejo. En primer lugar una misma abreviatura puede tener distintos significados dependiendo del entorno semántico y además, una misma abreviatura puede escribirse distinto dependiendo del autor o de la situación. Por ejemplo, para abreviar la palabra teléfono, en una misma página de información, se ha encontrado Tf., Tel., Telf. y Tfno.

El tratamiento de los número tampoco es trivial. Obviamente se necesita de un módulo específico que traduzca de la representación numérica a la ortográfica, pero este módulo no depende únicamente del propio número. Se ha de establecer concordancia en genero con la palabra a la que se refiere. Los números ordinales, indicados de forma diversa, requieren un tratamiento específico. Los números romanos también suelen aparecer en el texto escrito y en ocasiones son difíciles de detectar. Por ejemplo, Carlos I, no debe leerse como carlos i. Otro campo de dificultad son las fechas y las horas, presentes en gran variedad de formatos y que precisan establecer la concordancia en genero y número entre los numeros y hora/s, minuto/s.

Finalmente otros casos que presentan ambigüedad son los signos de puntuación. El punto puede indicar fin de frase, pero también marcar millares en los números, o incluso inicio de parte decimal. Además indica el fin de una abreviatura. Algo similar ocurre con la coma, con los dos puntos, los guiones, etc.

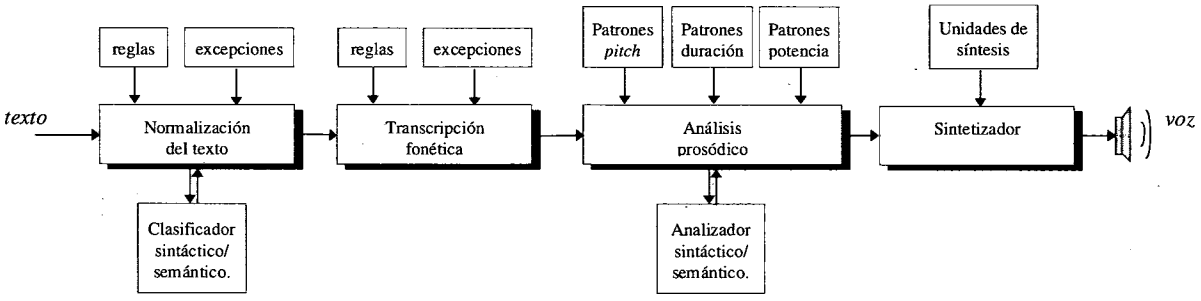

Figura 1. *Sistema de conversión de texto en voz.*

El método que se suele seguir para tratar toda esta casuística es el de tener un analizador que identifica cada uno de los componentes como palabra ordinaria, acrónimo, abreviatura, etc. Para cada caso, se dispone de unas reglas generales y de una lista de excepciones. Finalmente, se ha de aceptar un cierto error o una necesidad constante de actualizar los módulos anteriores adaptándolos a ámbitos específicos.

Silabificación y Transcripción fonética

Este módulo se encarga de representar el texto de entrada mediante un conjunto de sílabas y cada una de ellas mediante los fonemas (o alófonos) que lo componen. En castellano pueden definirse un conjunto de reglas que permiten determinar los alófonos que corresponden a las distintas letras. En catalán también es posible establecer estas reglas salvo algunas excepciones. Concretamente, no es posible distinguir cuando las vocales /e/ y /o/ tónicas son abiertas o cerradas, a no ser que éstas presenten tilde. En otras lenguas, como en la inglesa, es necesario un diccionario con la transcripción fonética de todas las palabras de la lengua, o al menos las más probables. Aún así, en ocasiones es preciso realizar un análisis sintáctico o semántico para asignar la transcripción adecuada a la palabra. Tal es el caso, en inglés, de read, (presente o pasado), o de lives (sustantivo o verbo).

Análisis prosódico

Una vez que se ha determinado cuál es la cadena de alófonos que componen el texto que se desea convertir en voz es necesario determinar cuál han de ser las características prosódicas de cada uno de forma que la señal de voz resultante tenga la continuidad y las características apropiadas al mensaje. La prosodia suele estar referido a tres características físicas: la frecuencia fundamental, la duración y la potencia.

La frecuencia fundamental, también llamada pitch, es la frecuencia de vibración de las cuerdas vocales, lógicamente cuando éstas vibran, en los sonidos sonoros. La evolución de la frecuencia fundamental con el discurso es la característica que influye de forma más notable y clara en la entonación. Es lo que se estudia, al hablar de entonación, en los libros básicos de lengua que se utilizan durante el bachillerato.

Para determinar el pitch que debe tener cada uno de los fonemas sonoros se recurre a unos patrones melódicos que dependen del tipo de frase sobre la que se aplica el patrón. Así, un patrón melódico de interrogación, por ejemplo, indica que a partir de la última sílaba acentuada se produce un fuerte incremento de la frecuencia fundamental. Estos patrones melódicos básicos a nivel de frase se pueden complementar con otros globales, que ajustan el nivel medio de la frase y su rango de variación dentro del párrafo de forma que las partes más importantes queden resaltadas. También se complementan con patrones de detalle que configuran la evolución en cada palabra, o incluso en cada fonema, sobre los valores básicos establecidos por los patrones de frase.

La duración establece el tiempo que ha de durar cada uno de los fonemas y cada una de las pausas que aparecen en el mensaje. Mediante coeficientes globales se regula la velocidad de articulación de la voz resultante. La duración de los fonema dependen del tipo de fonema, pero también del énfasis del fonema dentro del mensaje (acentos, inflexiones en la voz, etc.). Además, existe una tendencia de alargar la duración de los fonemas hacia el final de las frases.

La potencia asociada a la pronunciación de un fonema depende de cada fonema (por ejemplo, las vocales y algunos sonoros tienen mayor amplitud) , pero también del énfasis del fonema dentro del mensaje. En general parece existir cierta correlación entre el valor de la frecuencia fundamental y el incremento de la potencia del fonema respecto la media. Al igual que el *pitch* y la duración, se incrementa en las sílabas acentuadas. Otra consideración es que la potencia decae notablemente hacia el final de las frases.

Asignación de fonemas a unidades disponibles

Una vez que sabemos qué alófonos queremos concatenar y con qué características prosódicas podría parecer que el siguiente módulo debería tomar unas unida-

des básicas, modificarlas y concatenarlas. En realidad esto es lo que realiza el módulo de síntesis, pero previamente se deben relacionar los alófonos con las unidades básicas de síntesis que utiliza el sistema de conversión de texto en voz. Para obtener síntesis de alta calidad no es posible utilizar directamente los alófonos como unidades básicas de síntesis puesto que estas unidades están muy afectadas por la presencia de los sonidos adyacentes. El tracto vocal evoluciona de forma continua al articular los sonidos por lo que la transición entre los sonidos con los que se articulan los fonemas se realiza de forma gradual. De tal forma que, por ejemplo, las características espectrales del alófono /b/ en la palabra *Buran,* se acercan gradualmente hacia las del alófono /u/. Por tanto, si como unidad básicas se utiliza el alófono/b/extraído de la palabra*Burán*, será adecuado para ponerlo antes de /u/ pero no de otros alófonos.

Para solucionar este problema algunos sistemas modelan explícitamente la transición entre fonemas, pero la mayor parte de los sistemas actuales utilizan los difonemas como unidades básicas en la síntesis. Un difonema comprende parte de dos fonemas centrado precisamente en la transición. Así, la palabra *Burán*, se puede representar mediante los difonemas *$•b b•u u•r r•a a•n n•$*, dónde *$* indica silencio. Al concatenar el difonema *P•b* con *b•u,* la unión entre las unidades se realiza en la parte más estacionaria de */b/,* por lo que es más sencillo conseguir transiciones *suaves* entre segmentos.

Aún así, en algunos casos, la influencia de un alófono alcanza a varios de los fonemas adyacentes. Por ejemplo, el análisis espectral de los sonidos /tra/ revela que en realidad lo que se pronuncia es /tara/, con una primera *a* de corta duración. Por tanto, el difonema *t•r* debe depender de la vocal a la que precede. En general, cuanto mayor sea el tamaño de las unidades mejor es la calidad de la síntesis pero mayor es el número de ellas y por tanto mayor es la memoria que se requiere.

Una solución de compromiso consiste en utilizar mayoritariamente difonemas e introducir agrupaciones mayores en las situaciones más problemáticas. En castellano con inventarios de alrededor de 500 difonemas y 100 grupos mayores puede conseguirse buena calidad. Dicho inventario, a 16 kHz y codificando la señal únicamente mediante la ley A exige poco más de 1Mbyte de memoria.

Síntesis del habla

El último elemento del sistema de síntesis es el que toma unidades de síntesis elementales pregrabadas y las modifica para que tengan las características prosódicas deseadas. Existen distintos métodos, la mayoría derivados de sistemas de codificación. A continuación se describirán muy brevemente los dos más utilizados aunque en la literatura se encuentran otros métodos que también proporcionan buenas prestaciones. En todos los casos, el modificar la potencia de la señal consiste en añadir un factor de ganancia. Por tanto, sólo comentaremos lo que se refiere a frecuencia fundamental y duración.

MÉTODOS BASADOS EN PREDICCIÓN LINEAL

Estos métodos se basan en un modelo del aparato fonador como una excitación, que modela el aire que fluye de los pulmones y la vibración de las cuerdas vocales, y un filtro, que modela las cavidades bucales y nasales. El filtro se actualiza frecuentemente (cada varios milisegundos) de forma que represente correctamente los distintos estados que atraviesa el sistema tracto vocal. Esta descripción es válida entre otras para síntesis *LPC*, multipulso o *CELP*.

Para modificar la duración y la frecuencia fundamental se actúa sobre la excitación que es donde reside la información referente a la vibración de las cuerdas vocales. En el caso de la frecuencia fundamental para señales sonoras, la señal de excitación debe contener componentes periódicos. La periodicidad de la excitación (el *pitch*) en muchos métodos es un elemento del modelo que puede modificarse fácilmente en la fase de síntesis. Para modificar la duración puede tomarse la longitud deseada de respuesta del filtro a la excitación.

Métodos basados en solape temporal síncrono

Estos métodos han sido introducidos recientemente y proporcionan muy buena calidad con muy poca complejidad. En la fase de análisis, la señal se descompone en ventanas síncronas con la frecuencia de *pitch*. Estas ventanas, en la fase de síntesis, se suman acercándolas o alejándolas, para incrementar o disminuir el *pitch*. Las ventanas además pueden repetirse u omitirse para modificar la duración.

Síntesis por formantes

La síntesis de formantes suele prestar menor calidad que los otros métodos presentados, pero sus requerimientos de memoria son mínimos. No suele realizarse sobre difonemas sino que representa mediante ciertos parámetros cada uno de los alófonos y utiliza reglas de coarticulación e interpolación en las transiciones.

La voz sintética se genera sumando las respuestas de un banco de filtros a una señal de excitación. La excitación queda determinada mediante la potencia, el tipo de excitación (periódica o aleatoria) y, para el caso de periódicas, relacionada con sonidos sonoros, la frecuencia fundamental. En cuanto al banco de filtros, cada filtro paso-banda modela un formante quedando especificado por la frecuencia central y el ancho de banda.

BIBLIOGRAFÍA

El siguiente artículo se ha utilizado en la primera sección de esta comunicación y es una excelente referencia para valorar el impacto de las tecnologías del habla en lo referente a las telecomunicaciones:

L.R. RABINER: "Applications of Voice Processing to Telecommunications". Proceedings of the IEEE, pp. 199-228. February 1994, Vol. 82, No 2.