Reconocimiento de Formas y Análisis de Imágenes

Edición en formato Acrobat y Postscript para CDROM

La *Asociación Española de Reconocimiento de Formas y Análisis de Imágenes (AERFAI)* presenta una nueva iniciativa de publicación electrónica: Los Tutoriales de Reconocimiento de Formas y Análisis de Imágenes . Esta publicación está formada por un ámplio conjunto de artículos de enfoque fundamentalmente didáctico que cubren tanto los aspectos teóricos como las aplicaciones.

Créditos

Contenidos

Codificación de imagen.

Luis Torres, Philippe Salembier
Departamento de Teoría de la Señal y Comunicaciones
Universitat Politècnica de Catalunya

Restauración.

M.J. Yzuel, J. Campos. Universidad Autónoma de Barcelona.
M.S. Millán, J. Pladellorens. Universidad Politécnica de Catalunya.

Reconstrucción.

Jordi Núñez
Departament d'Astronomia i Meteorologia
Universitat de Barcelona

Filtraje y Mejora.

Ramon Roman
Universidad de Granada

Procesamiento Óptico.

Santiago Vallmitjana, Ignacio Juvells, Arturo Carnicer
Laboratori d'Òptica
Universitat de Barcelona

Visión por Computador

Introducción a la Visión por Computador.

Juan José Villanueva
Centro de Visión por Computador
Universidad Autómoma de Barcelona.

Segmentación de Imágenes.

Falcón A., Mendez J., Hernández F., Cabrera J., Lorenzo J.
Grupo de Inteligencia Artificial y Sistemas
Departamento de Informática y Sistemas
Universidad de Las Palmas G.C.

Morfología Matemática.

Jordi Vitrià , Maria Vanrell
Centre de Visió per Computador
Universitat Autònoma de Barcelona

Clasificación de Imágenes.

N. Pérez de la Blanca, F. Cortijo y R. Molina
Departamento de Ciencias de la Computación e Inteligencia Artificial
ETS de Ingeniería Informatica
Universidad de Granada

Análisis e Interpretación de Imágenes

Jaime López Krahe
École Nationale Superieure des Telécommunications
Paris, Francia

Visión 3D.

René Alquezar
Universitat Politécnica de Catalunya

Reconocimiento del Habla

Introducción al reconocimiento del habla.

José B. Mariño Acebal
Departamento de Teoría de la Señal y Comunicaciones
Universitat Politècnica de Catalunya

Preproceso y Segmentación de Señales Vocales.

Antonio J. Rubio Ayuso, José C. Segura Luna, Antonio M. Peinado Herreros
Dept. de Electrónica y Tecnología de Computadores
Facultad de Ciencias, Universidad de Granada

Metodologías de Reconocimiento del Habla.

P.Aibar, J.M.Benedí, F.Casacuberta, M.J.Castro, A.Marzal, F.Prat
Universidad Politécnica de Valencia

Modelización Acústica-Fonética.

Climent Nadeu y Javier Hernando
Dept. Teoria del Senyal i Comunicacions
Universitat Politècnica de Catalunya

Tecnologías del Habla.

Eduardo Lleida Solano
Departamento de Ingeniería Eléctrica e Informática
Universidad de Zaragoza.

Aplicaciones

Análisis de Imágenes en Biomedicina.

Domènec Ros, Javier Pavía, Ignacio Juvells
Universitat de Barcelona

Análisis de Imágenes en Radiología.

J. Serrat
Centro de Visión por Computador
Universidad Autónoma de Barcelona

Análisis de Imágenes en Geología.

Ángel Martínez Nistal , Modesto Montoto (*,**)
Servicio de Proceso de Imágenes, Universidad de Oviedo. *
Departamento de Geología, Área de Petrología y Geoquímica, Universidad de Oviedo. **

Análisis de imágenes microscópicas en medicina.

Augusto Moragas, Magdalena García-Bonafé, Inés de Torres
Unidad de Anatomía Patológica, Departamento de Ciencias Morfológicas
Facultad de Medicina de la UAB y Ciudad Sanitaria Universitaria Valle de Hebrón, Barcelona.

Análisis de Imágenes Astronómicas.

R. Molina, J.A. García y N. Pérez de la Blanca
Departamento de Ciencias de la Computación e I.A.
E.T.S. de Ingeniería Informática
Universidad de Granada.

Análisis de documentos

Enric Martí
Centre de Visió per Computador
Universitat Autònoma de Barcelona

Software de Procesamiento y Análisis de Imágenes.

Francisco Javier Sánchez
Centre de Visió per Computador
Universitat Autònoma de Barcelona

Arquitecturas Especializadas.

Alícia Casals, Antonio B. Martínez
Departamento de Ingeniería de Sistemas, Automática e Informática Industrial
Universitat Politècnica de Catalunya.

Edición patrocinada por

Reconocimiento de Formas y Análisis de Imágenes

Modelización acústico-fonética

Climent Nadeu y Javier Hernando
Dept. Teoria del Senyal i Comunicacions
Universitat Politècnica de Catalunya

Resumen

Los sistemas actuales de reconocimiento automático del lenguaje oral se basan en dos etapas básicas de procesado: la parametrización, que extrae la evolución temporal de los parámetros que caracterizan la voz, y el reconocimiento propiamente dicho, que identifica la cadena de palabras de la elocución recibida con ayuda de los modelos que representan el conocimiento adquirido en la etapa de aprendizaje. Tomando como línea divisoria la palabra, dichos modelos son de tipo acústico-fonético o gramatical. Los primeros caracterizan las palabras incluidas en el vocabulario de la aplicación o tarea a la que está orientado el sistema de reconocimiento, usando a menudo para ello modelos de unidades de habla de extensión inferior a la palabra, es decir, de unidades subléxicas. Por otro lado, la gramática incluye el conocimiento acerca de las combinaciones permitidas de palabras para formar las frases o su probabilidad. Queda fuera del esquema la denominada comprensión del habla, que utiliza adicionalmente el conocimiento semántico y pragmático para captar el significado de la elocución de entrada al sistema a partir de la cadena (o cadenas alternativas) de palabras que suministra el reconocedor.

Capítulo

Siguiente .... Anterior
Índice

Tema 4, Capítulo 4

MODELIZACION ACUSTICO-FONETICA

Climent Nadeu y Javier Hernando

Dept. Teoria Senyal i Comunicacions
Universitat Politècnica de Catalunya

1. INTRODUCCION

Los sistemas actuales de reconocimiento automático del lenguaje oral se basan en el esquema de la Figura 1. En él se distinguen dos etapas básicas de procesado: la *parametrización*, que extrae la evolución temporal de los parámetros que caracterizan la voz, y el *reconocimiento* propiamente dicho, que identifica la cadena de palabras de la elocución recibida con ayuda de los modelos que representan el conocimiento adquirido en la etapa de aprendizaje [Rabiner 93]. Tomando como linea divisoria la palabra, dichos modelos son de tipo *acústico-fonético* o *gramatical*. Los primeros caracterizan las palabras incluidas en el vocabulario de la aplicación o tarea a la que está orientado el sistema de reconocimiento, usando a menudo para ello modelos de unidades de habla de extensión inferior a la palabra, es decir, de *unidades subléxicas*. Por otro lado, la gramática incluye el conocimiento acerca de las combinaciones permitidas de palabras para formar las frases o su probabilidad. Queda fuera del esquema la denominada *comprensión* del habla, que utiliza adicionalmente el conocimiento semántico y pragmático para captar el significado de la elocución de entrada al sistema a partir de la cadena (o cadenas alternativas) de palabras que suministra el reconocedor.

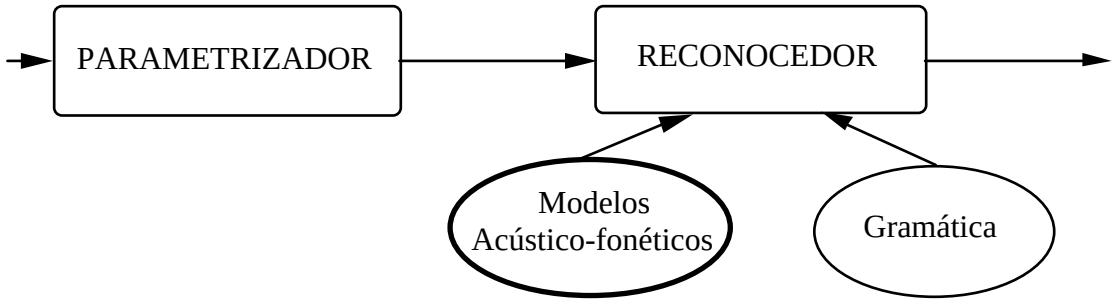

Figura 1. Esquema básico de un sistema de reconocimiento del habla.

El objeto de estudio del presente capítulo es la modelización de las unidades acústico-fonéticas. Puesto que dichos modelos acústico-fonéticos estan expresados en términos de los *parámetros*

o modelos de señal que representan la voz, deberemos empezar nuestra presentación con el estudio de las formas básicas de llevar a cabo la parametrización de la señal de voz, lo cual será tratado en la siguiente sección. Una vez establecidos dichos parámetros, podremos pasar en la sección tercera a la consideración de las unidades de habla más apropiadas y a los modelos matemáticos que las caracterizan.

2. MODELIZACION DE LA SEÑAL DE VOZ

Para modelar las unidades del habla se requiere una representación de la propia señal que describa sus características relevantes desde el punto de vista del reconocimiento [Picone 93]. Esta representación o modelo se expresa como evolución temporal de un conjunto de parámetros. El esquema de la Figura 2 muestra las distintas etapas en que hemos dividido el proceso de obtención de los parámetros comúnmente utilizado, donde el análisis se realiza por *tramos* , es decir, desplazando regularmente a saltos la ventana a través de la cual se observa la señal.

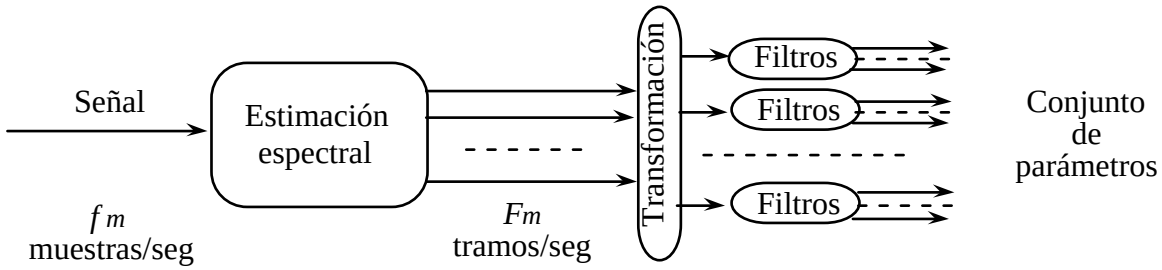

Figura 2. Esquema típico de determinación del vector de parámetros para cada tramo de señal.

El primer apartado de la presente sección aborda el problema de la captación de las características relevantes mencionadas, las cuales están asociadas al *espectro de potencia* de la señal. En el segundo apartado se exponen las formas básicas de realizar la estimación de los parámetros espectrales de cada tramo. A continuación, en el apartado 2.3, se describen la transformación y las operaciones de filtrado de las secuencias temporales de dichos parámetros que permiten aumentar la capacidad de discriminación del reconocedor. Y, finalmente, el apartado 2.4 trata de las modificaciones o técnicas alternativas que se utilizan cuando la voz ha sido generada en condiciones ambientales adversas.

2.1 El problema de la representación espectral de la voz

La señal de voz se puede modelar simplificadamente como respuesta de un filtro variante en el tiempo cuando es excitado por una señal cuyo espectro es o bien plano (segmento sordo) o un tren de líneas espectrales situadas a frecuencias múltiplos de la frecuencia fundamental o frecuencia de vibración de las cuerdas vocales (segmento sonoro) [O'Shaugnessy 90]. Por consiguiente, la envolvente del espectro de la señal coincide con la respuesta espectral del filtro en ambos casos. Así pues, la evolución temporal de la envolvente espectral transporta casi toda la información relativa al habla. De acuerdo con el modelo anterior, únicamente queda al margen de la envolvente la información referente a la frecuencia fundamental (tono de la voz), la cual es raramente utilizada por ahora en reconocimiento, si exceptuamos el caso de lenguas tonales como el mandarín.

De este modo, el objetivo del análisis espectral de la señal de voz con vistas al reconocimiento del habla es la estimación de la envolvente espectral, es decir, $S(\omega,n)$, donde ω es la variable frecuencia (angular) y *n* indica el índice temporal. Para ello es necesario llevar a cabo la desconvolución de la señal, es decir, la separación en la respuesta del filtro de las contribuciones debidas a la excitación y al propio filtro, para quedarse sólo con esta última.

Para poder llevar a cabo un reconocimiento con buenas prestaciones es importante que el número de variables que intervienen en el proceso sea lo más bajo posible, sin perder información significativa. La razón estriba en que la base de datos de entrenamiento siempre es limitada, con lo que cuantos más grados de libertad tenga el modelo menos fiables serán los valores entrenados y, por otro lado, más costoso será el proceso de reconocimiento. Así pues, la envolvente espectral estimada ha de poder ser representada en forma paramétrica, o sea, con un número reducido de parámetros.

Cada estimación acarrea un cierto error aleatorio. Suponiendo gausianidad, el error debido a la estimación espectral puede representarse con los dos primeros momentos, media y varianza [Papoulis 91]. El sesgo asociado a la media es el causante de distintos efectos, el más relevante de los cuales es -probablemente- la pérdida de resolución [Marple 87]. Puesto que $S(\omega,n)$ es una función bidimensional, dicha pérdida repercute en ambas dimensiones, tiempo y frecuencia. Una adecuada resolución temporal resulta importante para seguir bien las transiciones, es decir, los cambios que se producen en las características del habla en torno de las fronteras entre sonidos (y también en el interior de algunos sonidos). Una adecuada resolución frecuencial posibilita captar los pequeños detalles de la envolvente espectral y, en consecuencia, aumentar la potencia discriminativa de la representación. En conclusión, podemos considerar que las medidas más significativas del comportamiento del estimador

espectral de un sistema de reconocimiento son la varianza y las resoluciones frecuencial y temporal.

Idealmente, el analizador espectral debería obtener una función $S(\omega,n)$ que capturara el dinamismo de la señal oral. Esto significa que en las transiciones debería afinarse más la descripción temporal y lo mismo puede decirse de la descripción frecuencial en las bandas con mayor significación dentro de un segmento dado de señal. Por el contrario, en las regiones tiempo-frecuenciales donde las características de la señal se estabilizan o son poco relevantes para la discriminación, una representación paramétrica parsimoniosa evitaría transferir al reconocedor valores de los parámetros redundantes o irrelevantes. Aunque ha habido intentos para conseguir este tipo de representación flexible mediante herramientas recientes del procesado de señal basadas en una representación conjunta tiempo-frecuencia, en reconocimiento del habla sigue imperando la representación completamente rígida expuesta en la Figura 2, basada en una partición de la señal en tramos de igual longitud, en cada uno de los cuales se representa el espectro de idéntica manera. Sin embargo, existen numerosos trabajos que, partiendo de dicha representación estática, realizan un postratamiento de la secuencia de parámetros a fin de detectar segmentos temporales acústicamente homogéneos que son susceptibles de ser caracterizados como una entidad diferenciada (segmentación de traza, descomposición temporal, etc; véase, p. ej., [Vidal 90],[Lleida 90]).

En principio, y para un determinado número de parámetros, un reconocedor de habla debería ser capaz de sacar provecho de unas (supuestamente) mejores prestaciones de la estimación espectral para aumentar correspondientemente la tasa de aciertos. Sin embargo, lo cierto es que ello no depende exclusivamente del estimador espectral sino también del propio reconocedor y, principalmente, del grado de adaptación que exista entre la representación espectral y el tipo de medida de distancia o de probabilidad que emplee el reconocedor.

En los sistemas actuales de reconocimiento, existen diversas maneras de realizar la estimación espectral [Picone 93]. El resto de apartados de la presente sección efectuará un breve recorrido por las técnicas más significativas.

2.2 Estimación de la envolvente espectral de un tramo de señal

Existen múltiples procedimientos para la estimación del espectro de un tramo de señal. De hecho, la estimación espectral es un apartado importante del tratamiento digital de señal [Marple 87]. Dado que nuestro problema es la estimación de la envolvente (es decir, no nos hace falta determinar los armónicos debidos a la vibración de las cuerdas vocales) la disciplina de estimación espectral nos ofrece -básicamente- como solución dos enfoques alternativos que parten de las dos herramientas fundamentales del tratamiento digital de señal, a saber, el filtro

digital y la transformada discreta de Fourier. Las técnicas empleadas en los sistemas de reconocimiento se basan en uno u otro enfoque y, a veces, en la combinación de ambos.

En el primer enfoque (denominado, a menudo, paramétrico) se trata de determinar los coeficientes del filtro -que sólo contiene polos- que aparece en el modelo de la señal de voz mencionado anteriormente, de forma que se minimize una medida cuadrático-media del error de la aproximación [Atal 67] [Itakura 75]. La técnica comúnmente utilizada, denominada *LPC* (codificación por predicción lineal) utiliza como punto de partida de la estimación los $p+1$ primeros valores de la autocorrelación de la señal, donde p es el número de polos del filtro. Puesto que p suele presentar un valor próximo a 10 (para frecuencia de muestreo de 8Khz), mientras que el periodo fundamental de la voz es bastante superior para voces normales, este tipo de estimador consigue desconvolucionar eficientemente la señal, reduciendo al mismo tiempo la varianza, por el hecho de prescindir de los demás valores de la autocorrelación, que son menos fiables.

El segundo enfoque (denominado no-paramétrico) es la base de las técnicas más utilizadas en la actualidad [Picone 93]. Por esta razón lo vamos a considerar con más detenimiento, desarrollando una vía explicativa que nos permitirá resaltar la íntima relación existente entre dos técnicas de estimación bien distintas a primera vista, como son la basada en el *banco de filtros* y la *mel-cepstrum*.

En este tipo de técnicas, el espectro se contempla dividido en bandas, al contrario de lo que ocurre en la estimación de tipo LPC, donde se estima globalmente la banda entera. El concepto fundamental es, pues, el de potencia en una banda espectral. Si $S(\omega)$ es el espectro de potencia, la potencia en la banda de ancho W, centrada entorno la frecuencia ω_0 se define como

$$P_W(\omega_0) = \frac{1}{2W}\int_{\omega_0 - W}^{\omega_0 + W} S(\omega)\, d\omega \qquad (1)$$

es decir, la parte de potencia de la señal que cae dentro de la banda en cuestión.

El periodograma [Marple 87] es el estimador más immediato de la potencia de la señal a una frecuencia ω y viene dado por la expresión

$$P(\omega) = \frac{1}{N}\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\omega n}\right|^2 \qquad (2)$$

en la que $x(n)$ es la señal, $w(n)$ la ventana a través de la cual se “ve” la señal y N su longitud. El ancho de banda viene determinado por la forma de la ventana. El periodograma no puede ser utilizado directamente en reconocimiento del habla por tres razones fundamentales: 1) presenta alta varianza; 2) incluye tanto la envolvente espectral como los armónicos de la frecuencia

fundamental; y 3) no puede representarse con un conjunto reducido de parámetros. Como veremos a continuación, los tres inconvenientes pueden ser superados acudiendo a un promediado de distintos periodogramas para estimar la potencia en una banda centrada en ω_m, es decir,

$$P_W(\omega_m) = \sum_k d(k) \left| \sum_n h_k(n) x(n) e^{-j\omega_m n} \right|^2 \qquad (3)$$

donde, como puede observarse, k indica cada uno de los periodogramas que se promedian y $d(k)$ es el factor de ponderación que reciben. Obsérvese que la diferencia entre periodogramas radica únicamente en la forma de la secuencia $h_k(n)$ que multiplica la señal $x(n)$.

Cuando

$$h_k(n) = w(n) e^{j\omega_k n}, \quad \omega_k = \frac{2\pi}{N} k \qquad (4)$$

k es un índice frecuencial, y los distintos valores que se promedian en (3) resultan de desplazar ω_k la variable frecuencia del periodograma de (2). Así pues, en este caso podemos interpretar (3) como promediado de valores del periodograma en torno ω_m. Se trata del estimador espectral utilizado por la parametrización denominada mel-cepstrum [Davis 80].

Cuando

$$h_k(n) = h(n - kM) \qquad (5)$$

k es un índice temporal, y $h(-n)$ es la respuesta impulsional de un filtro cuya banda de paso coincide con la banda centrada en ω_m. Así pues, en este caso podemos interpretar (3) como promediado de periodogramas obtenidos a partir de distintos segmentos de la señal de longitud N centrados en las muestras kM. Es el estimador utilizado por la parametrización basada en los valores de potencia a la salida de un banco de filtros paso-banda [Dautrich 83].

En el primer caso tenemos un promediado frecuencial y en el segundo un promediado temporal. Ambos estimadores espectrales permiten superar los tres inconvenientes antes aludidos del periodograma. En primer lugar, es obvio que el hecho de promediar consigue reducir la varianza de la estimación; de hecho, el periodograma no es un estimador consistente, pero sí los obtenidos con sus promedios. Asimismo, el promediado consigue eliminar casi completamente los armónicos de la frecuencia fundamental, realizando por tanto una desconvolución de la señal. Por último, espaciando convenientemente las distintas frecuencias ω_m a lo largo del eje frecuencial, se obtiene con (3) un conjunto de parámetros espectrales $P_W(\omega_m)$ que representan la envolvente espectral (su número suele oscilar entre 12 y 24). Dicho espaciamiento puede ser

lineal o no serlo. En este segundo caso se utilizan espaciamientos de tipo logarítmico basados en observaciones del funcionamiento de la audición [O'Shaugnessy 90], como la escala mel usada en la parametrización mel-cepstrum, donde los anchos de banda W se toman dependientes del índice *m*.

Como hemos visto, cada estimador espectral presenta su propia forma de llevar a cabo la desconvolución de la señal y también de fijar el compromiso -mencionado en el apartado anterior- entre la varianza del error de la estimación y las resoluciones temporal y frecuencial. Sin embargo, este compromiso depende sobre todo de dos parámetros básicos presentes en todos los estimadores: 1) El ancho de banda efectivo W_e en los estimadores que determinan la potencia de las bandas espectrales, o el orden del modelo p (número de polos) en la estimación LPC. 2) La longitud efectiva de la ventana temporal L_e. Para un determinado valor de L_e ,W_e (o p) controla el compromiso entre varianza y resolución frecuencial. Para un determinado valor de W_e (o p), L_e controla el compromiso entre varianza y resolución temporal.

Desafortunadamente, no existen todavía comparaciones suficientemente amplias y fiables de los distintos tipos de parametrización de la señal de voz que permitan llegar a conclusiones definitivas sobre sus prestaciones en reconocimiento. Dichas prestaciones pueden depender de diversos factores: tipo de sistema (p. ej. los sistemas basados en redes neuronales suelen emplear como parámetros las energías a la salida de un banco de filtros), tarea de reconocimiento, frecuencia de muestreo, ambiente acústico donde se produce la voz, etc. El uso extendido de modelos de señal no-paramétricos es debido seguramente a la sencillez de su implementación a partir de la FFT (algoritmo eficiente de cálculo de la transformada discreta de Fourier) y a la flexibilidad que comporta la partición en subbandas, ya que trabajando en subbandas espectrales resulta sencillo incorporar la escala mel o ciertas técnicas de robustez al ruido.

2.3 Posprocesado de los parámetros espectrales

Dado un estimador espectral, existe la posibilidad de mejorar sus prestaciones dentro del sistema de reconocimiento posprocesando la representación bidimensional tiempo-frecuencia obtenida. Puesto que los reconocedores normalmente utilizan o bien distancia euclídea o bien funciones de densidad de probabilidad gausianas (cuya evaluación también involucra el cálculo de la distancia euclídea respecto de la media) un posprocesado eficaz implicará, aunque sea implícitamente, una adaptación de los parámetros espectrales a la forma de la medida euclídea.

En primer lugar, es corriente transformar en el posprocesado los parámetros de cada tramo al dominio del cepstrum (transformada inversa de Fourier del logaritmo del espectro [Oppenheim

89]) y ponderarlos de forma que se normalice la varianza de cada coeficiente cepstral [Tokhura 87]. A continuación se procesa cada una de las secuencias temporales de coeficientes cepstrales mediante una aproximación de la primera o primeras derivadas y los parámetros diferenciales resultantes se utilizan en lugar de los parámetros espectrales o como complemento de ellos. En nuestra explicación vamos a contemplar la ponderación cepstral como filtrado frecuencial y las derivaciones como filtrado temporal, de acuerdo con lo expresado en la Figura 2. Esta -poco corriente- interpretación nos permitirá observar lo que ocurre en el dominio transformado respectivo y de esta forma mostrar el cambio que estos tipos de posprocesado de los parámetros operan en el compromiso entre varianza y resolución establecido por el estimador espectral.

A) Filtrado en el dominio frecuencial

La ponderación de la secuencia de coeficientes cepstrales se realiza con una ventana de longitud finita M que iguala aproximadamente la varianza (coincidente con la potencia) de los coeficientes cepstrales. La ventana empleada suele tener forma de rampa (recta de pendiente unidad) [Hanson 87] o seno remontado [Juang 86]. Se puede demostrar que ambas implican una cierta derivación de $S(\omega,n)$ respecto de la variable ω, seguida de la convolución (de funciones periódicas) con la transformada de Fourier de un pulso discreto [Nadeu 94]. De esta manera, ambas ventanas de ponderación realizan una operación combinada de derivación y alisado. El alisado reduce la varianza del error de la estimación espectral, ya que cancela los coeficientes cepstrales de índice superior a M, que son los menos fiables. Sin embargo, esta cancelación produce también pérdida de resolución espectral al incrementar el ancho de banda efectivo *We* de la estimación. La componente derivativa de la operación consigue aumentar en cierta modo la resolución a base de emfatizar los coeficientes cepstrales de índices altos con respecto a los de índices bajos.

Gracias a la igualación o normalización de la varianza de los parámetros, la distancia euclídea usada en el reconocedor coincide con la distancia de Mahalanobis [Duda 73], siempre y cuando se pueda suponer incorrelación entre parámetros, lo cual es razonable con los coeficientes cepstrales [Picone 93] (y ésta es posiblemente la clave del éxito de los coeficientes cepstrales cuando se usa la distancia euclídea en reconocimiento). Por otro lado, cuando el reconocedor emplea modelos probabilísticos, las funciones de densidad gausianas ya incorporan dicho tipo de normalización.

A) Filtrado en el dominio temporal

Los sistemas de reconocimiento usan este tipo de filtrado lineal para dos objetivos distintos. En primer lugar, cuando las señales estan distorsionadas linealmente por el sistema de adquisición (micrófono, canal telefónico, etc) y dicha distorsión no es idéntica para todas ellas, la tasa de error aumenta sustancialmente. La distorsión únicamente contribuye con un término aditivo a

los coeficientes cepstrales, de forma que si es invariante en tiempo dentro del intervalo de observación, se puede cancelar su influencia eliminando la componente continua de las secuencias temporales de coeficientes cepstrales. Para conseguirlo pueden emplearse filtros con polos y ceros diseñados a medida (véase, como muestra, [Hirsch 91] [Hermansy 91]), los cuales presentan como características comunes un cero a frecuencia cero y un polo real cercano al valor unidad; este último controla la frecuencia de corte de la banda atenuada entorno al origen. Puesto que al cero a frecuencia cero le corresponde en el dominio temporal la diferencia finita de una muestra y el polo realiza un filtrado paso-bajo, este tipo de filtro lleva a cabo la misma operación combinada de derivación y alisado que observamos en la ponderación cepstral del apartado anterior, si bien esta última opera en el dominio frecuencial.

El segundo objetivo del filtrado lineal es más general. En efecto, en la actualidad se ha generalizado el uso de los parámetros diferenciales (también denominados delta-parámetros, o características dinámicas) como complemento de los parámetros suministrados por el estimador espectral [Furui 86]. Estos parámetros son del mismo tipo que los mencionados en el párrafo anterior, pues se obtienen realizando también una derivación alisada de las secuencias de parámetros espectrales en el entorno de cada tramo. Los denominaremos parámetros filtrados, ya que son el resultado de filtrar las secuencias temporales de parámetros espectrales. Los filtros correspondientes son de respuesta impulsional finita, relativamente corta (es corriente una longitud de 5 tramos). Hasta el momento presente se han utilizado básicamente dos tipos distintos. El primero responde a la igualdad [Lee K.F. 89]

$$d_m(n) = c_m(n - \frac{N}{2}) - c_m(n + \frac{N}{2}) \qquad (6)$$

donde $c_m(n)$ es la secuencia temporal del parámetro espectral m y $d_m(n)$ es la secuencia del parámetro filtrado correspondiente. Aplicando sucesivamente el filtro a la secuencia $c_m(n)$ se obtienen los parámetros filtrados de grados sucesivos.

El segundo filtro tiene como respuesta impulsional el polinomio de Legendre discreto de primer grado, basándose en la ecuación [Furui 86]

$$d_m(n) = \sum_{i=1}^{N/2} i \left[c_m(n-i) - c_m(n+i)\right] \qquad (7)$$

Cuando $c_m(n)$ son los coeficientes cepstrales, $d_m(n)$ se suele denominar delta-cepstrum. Los parámetros filtrados de grado sucesivo pueden obtenerse con los distintos polinomios de Legendre discretos o aplicando sucesivamente el de primer grado, aunque variando el orden N del filtro. Prácticamente todos los sistemas actuales utilizan al menos el parámetro filtrado de primer grado, muchos usan también el de segundo grado y algunos incluso el de tercer grado.

Por poner un ejemplo típico en reconocimiento del habla continua, en [Lee C.H.90] se usa $d_m(n)$ con $N=4$ para el primer parámetro filtrado y se obtiene el segundo aplicando la misma ecuación al primero con $N=2$. Los dos filtros resultantes se presentan en la Figura 3.

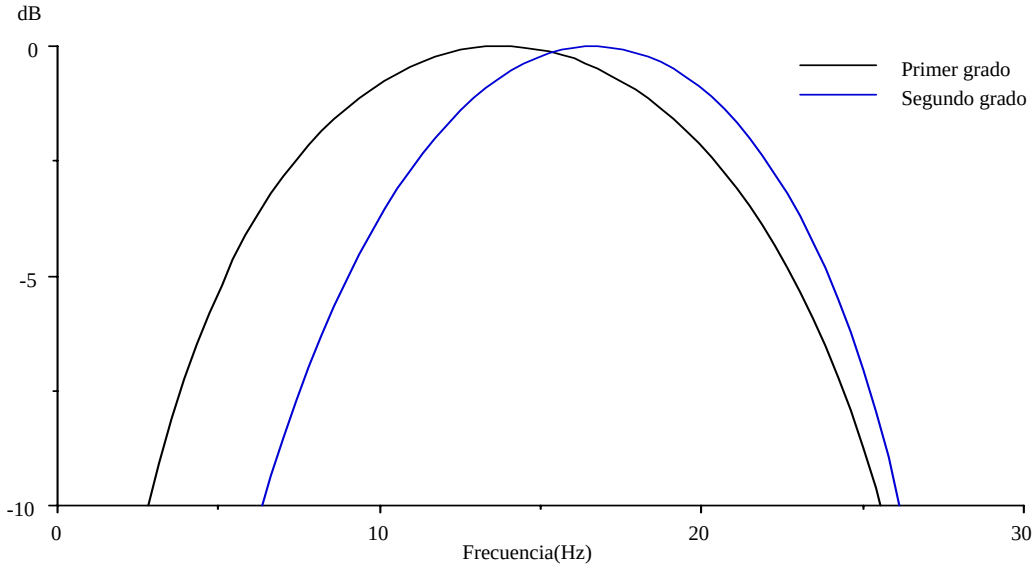

Figura 3. Respuesta frecuencial de los dos filtros utilizados en [Lee C.H. 92] para obtener las secuencias de parámetros filtrados.

Al igual que en los filtros canceladores de la distorsión lineal, tanto (6) como (7) son filtros paso-banda que combinan un cero a frecuencia cero (es decir, una derivación; obsérvese que si $c_m(n)$ es constante con n, la salida $d_m(n)$ es nula) con un alisado (la combinación lineal abarca un intervalo temporal superior al de dos muestras consecutivas). Vemos por tanto que, análogamente al filtrado frecuencial, en ambos tipos de parámetros filtrados también se produce un cambio en el compromiso varianza-resolución merced a la operación combinada derivación-alisado.

2.4 Técnicas específicas para reconocimiento en condiciones adversas

El comportamiento de los sistemas actuales de reconocimiento, que en general son diseñados suponiendo que las condiciones ambientales en que van a operar son tolerables y no van a afectar de forma sustancial la señal de voz, se degrada ostensiblemente en entornos reales, donde las condiciones pueden ser claramente desfavorables o adversas.

En general, tales condiciones adversas consisten en la presencia de ruido ambiente (de oficina, de coche,...), en la reverberación de la propia sala y en distorsiones y ruidos introducidos por

los transductores y el canal de transmisión (micrófonos, canal telefónico,...). Además, también han de tenerse en cuenta las variaciones en el modo de articular del hablante debidas a su reacción psicológica al entorno ruidoso (efecto Lombard).

Se ha observado que el principal problema lo constituye el desajuste entre las condiciones ambientales de entrenamiento y reconocimiento. La degradación de la tasa de reconocimiento debida a unas condiciones adversas durante el reconocimiento es mucho más drástica cuando se entrena el sistema con señales libres de ruido que cuando se entrena con señales inmersas en similares condiciones a las de reconocimiento. Aunque este resultado apunta a una posible manera de aumentar la robustez del sistema a las condiciones adversas, el problema estriba en el hecho de que la disponibilidad de datos de entrenamiento que reflejen las condiciones de reconocimiento es raras veces realista (por ser desconocidas estas condiciones, por ser difíciles de obtener, por ser variables en el tiempo,...).

Por estas razones, se requieren soluciones más elaboradas para resolver el problema. En los últimos años, se han propuesto en este sentido gran variedad de métodos y algoritmos en varias de las etapas del sistema de reconocimiento [Juang 91], entre los que destacan: 1) Métodos de mejora de la señal de voz, como la utilización de varios micrófonos [Viswanathan 86], la cancelación adaptativa de ruido [Widrow 75] o la substracción espectral [Boll 79]. 2) Nuevas parametrizaciones de la señal de voz, bien tratando de emular la capacidad auditiva humana [Ghitza 86], bien desde el punto de vista de procesado de la señal [Hernando 93]. 3) Medidas de distancia robustas [Mansour 89]. 4) Enmascaramiento de ruido, que consiste en eliminar en el proceso de comparación la contribución de las bandas frecuenciales cuya energía sea inferior a un determinado umbral [Klatt 79].

Sin embargo, a pesar de esta profusión de métodos, el reconocimiento del habla cuando el entorno es especialmente adverso no ha encontrado todavía una solución satisfactoria ni en el caso de reconocimiento de palabras aisladas, dependiente del locutor y con vocabularios pequeños.

A modo de ejemplo, en el resto del apartado se presentan los principios que inspiran tres de estos métodos: la mejora de la señal de voz mediante substracción espectral, la parametrización de la voz basada en modelos auditivos y la distancia de proyección cepstral.

1) *Substracción espectral*

Si no se consideran las características del canal y se supone que el ruido es aditivo e incorrelado con la señal de voz, el espectro de potencia de la señal ruidosa $S_y(\omega)$ resulta ser la suma de los espectros de la señal limpia $S_x(\omega)$ y del ruido $S_n(\omega)$. Estimando $S_y(\omega)$ a partir de la señal

ruidosa y $S_n(\omega)$ a partir de una señal que contiene solamente ruido (se supone que éste es estacionario), el espectro de la señal limpia se modela con la simple ecuación

$$S_x(\omega) = S_y(\omega) - S_n(\omega) \qquad (8)$$

(se impone un umbral mínimo positivo, ya que la aplicación directa de esta fórmula puede dar lugar a una estimación inconsistente -negativa- del espectro). A partir del resultado de esta substracción espectral [Boll 79], puede reconstruirse la señal temporal utilizando la fase de la señal ruidosa como fase de la señal limpia (ver Figura 4).

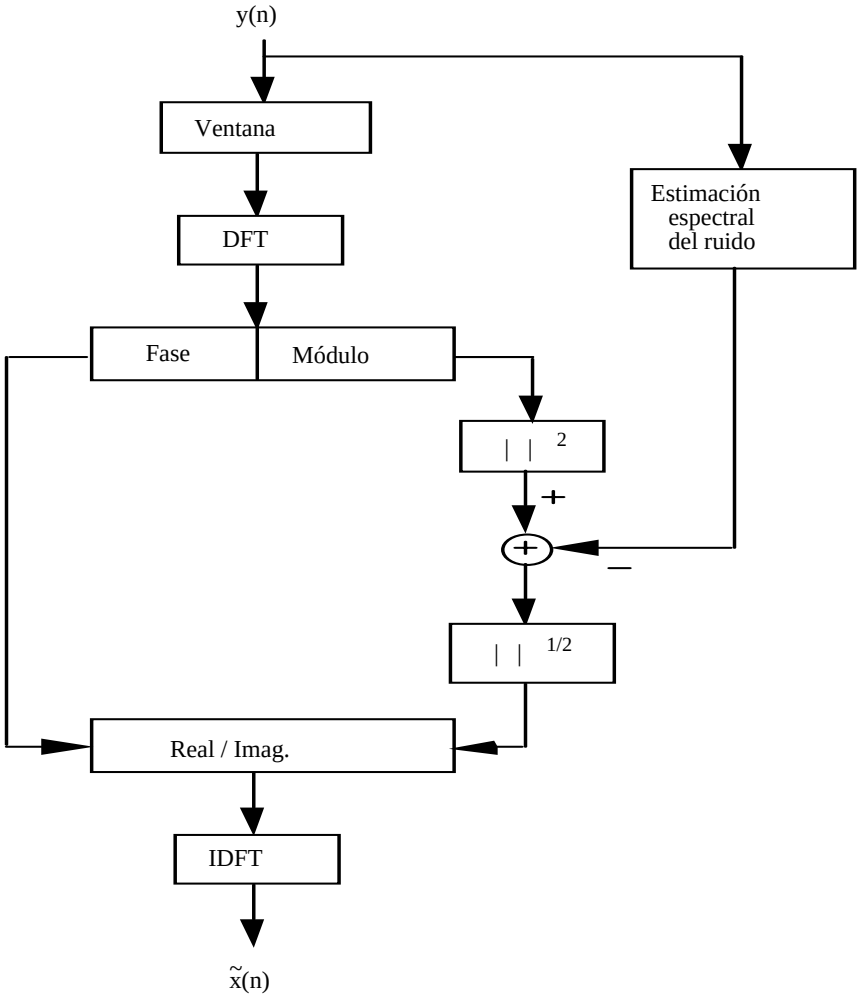

Figura 4. Algoritmo simplificado del método de substracción espectral. DFT e IDFT son los algoritmos utilizados para realizar las transformadas de Fourier directa e inversa, respectivamente.

La propiedad más sobresaliente de la substracción espectral es su simplicidad. La mayor dificultad es la obtención de una buena estimación del espectro de potencia del ruido, que suele hacerse promediando espectros de potencia localizados en tramos donde se supone que la señal contiene sólo ruido.

2) *Parametrizaciones basadas en modelos auditivos*

Es un hecho bien conocido que el sistema auditivo humano es más robusto que cualquier sistema automático no sólo frente al ruido aditivo y las distorsiones en general sino también frente a cualquier factor de variabilidad de la voz (como el modo de articulación del locutor debido a sus características personales, su estado emocional, la influencia del entorno, etc.). Por tanto, es de esperar que un sistema de reconocimiento del habla sea más robusto a todos estos factores si la etapa de representación de la señal de voz imita las características fisiológicas o psicoacústicas del oído humano.

Basándose en esta premisa, se han realizado varios intentos de modelar el sistema auditivo para aumentar la robustez del sistema de reconocimiento. Entre ellos destaca el modelo EIH (*Ensemble Interval Histogram*) [Ghitza 86], que pretende imitar el patrón temporal de descarga de las fibras del nervio auditivo. En primer lugar, el modelo EIH separa la señal de voz en la banda de 100-3200 Hz usando 85 filtros que simulan el poder discriminativo en frecuencia de la cóclea; después se detectan los intervalos en que la salida de cada filtro excede un cierto umbral; finalmente, se calculan los histogramas de estos intervalos. El histograma conjunto resultante es asimilable a un espectro, pero con las no-linealidades y la resolución no uniforme en frecuencia que son característicos del procesado auditivo humano.

3) *Distancia de proyección cepstral*

Evidencias tanto analíticas como experimentales indican que el ruido blanco aditivo provoca una reducción de la norma del vector cepstral, la cual resulta perjudicial cuando se utiliza en reconocimiento la distancia euclídea cepstral. Sin embargo, la orientación del vector queda más o menos intacta, resultado que sugiere el uso de una operación de proyección para calcular la medida de distancia -o la probabilidad- en el caso de que el sistema sea entrenado en condiciones libres de ruido pero las condiciones de reconocimiento sean desconocidas [Mansour 89]. En pruebas de reconocimiento en ambiente ruidoso, estas técnicas ha mostrado un buen comportamiento, incluso en presencia de efecto Lombard.

3 MODELIZACION DE LAS UNIDADES DE HABLA

Las características de los modelos acústico-fonéticos empleados por el sistema de reconocimiento dependen de la tarea en la que éste vaya a ser utilizado. Cuando dicha tarea de reconocimiento presenta cierta complejidad a nivel fonético, por ejemplo cuando el número de palabras que se utilizan es elevado, se recurre habitualmente al modelado de unidades de menor extensión, o sea, unidades de tipo subléxico, como partes constituyentes de los modelos de las palabras. La exposición comienza con la descripción de los distintos tipos de unidades utilizadas en reconocimiento del habla, ponderando sus ventajas e inconvenientes, y las alternativas existentes para la composición de las palabras a partir de unidades subléxicas. Posteriormente, se aborda el modelado probabilístico de las unidades, tanto en lo que respecta al tipo de modelos y su topología como al entrenamiento automático de dichos modelos a partir de los datos disponibles y su evaluación en los tests de reconocimiento.

3.1 Unidades

Cuando el tamaño del vocabulario es reducido, como es el caso de muchas aplicaciones de palabras aisladas o conectadas, los sistemas de reconocimiento del habla suelen utilizar la palabra como unidad de habla. Resultados experimentales muestran que esta elección proporciona las mejores prestaciones [Rosenberg 83] .

Sin embargo, la utilización de la palabra como unidad de reconocimiento plantea varios problemas graves cuando se trabaja con vocabularios extensos (más de 1000 palabras). En primer lugar, se necesita una enorme cantidad de datos de entrenamiento para asegurarse que cada palabra aparece el número de veces suficiente para entrenar explícitamente cada modelo de palabra. También pueden desbordarse los requerimientos de memoria y la complejidad de la búsqueda. Finalmente, en muchas tareas es conveniente proporcionar al usuario la posibilidad de añadir nuevas palabras al vocabulario, lo cual obliga a pronunciar muchas realizaciones de las palabras nuevas si se utiliza la palabra como unidad de reconocimiento.

Por ello, cuando el número de palabras que se utilizan en una tarea es elevado la palabra deja de ser una unidad de reconocimiento práctica y se recurre habitualmente al modelado de unidades de menor extensión, unidades de tipo subléxico. Esta opción permite superar en parte los problemas mencionados, pero añade un nuevo nivel al proceso de reconocimiento, la modelización léxica. En este nivel las unidades subléxicas se han de combinar para formar palabras de acuerdo con un diccionario que describe todas las palabras del léxico en términos de las unidades utilizadas.

3.1.1 Tipos de unidades subléxicas

En la definición de las unidades subléxicas pueden distinguirse dos alternativas básicas, según la distinta aplicación de los conocimientos acústico y lingüístico.

Una primera alternativa consiste en definir las unidades subléxicas confiando enteramente en la similitud acústica medible en la señal de voz [Svendsen 87] [Vidal 90]. Se trata de encontrar un conjunto de modelos de segmentos acústicos que cubran el espacio de señal y utilizar estos modelos como unidades de reconocimiento. Las señales de entrenamiento son segmentadas y los segmentos resultantes se agrupan con criterios de homogeneidad acústica (por ejemplo, mediante cuantización de segmentos, es decir, de matrices) para crear los modelos de las unidades. El proceso puede ser iterativo, de forma que se vayan determinando alternadamente las unidades y la segmentación [Shiraki 88].

Este método tiene la ventaja de que la definición de las unidades es consistente con la segmentación y con los modelos resultantes. Sin embargo, presenta el grave inconveniente de que se ha de recurrir a métodos automáticos para construir a posteriori un diccionario de palabras basado en las unidades acústicas obtenidas para la aplicación, ya que dichas unidades acústicas son difícilmente interpretables en términos lingüísticos. Se ha demostrado que un conjunto de 256-512 unidades acústicas es apropiado para un rango amplio de aplicaciones [Lee C.H. 89].

La alternativa más usada y a la que nos referiremos en adelante consiste en definir a priori un conjunto básico de unidades basándose fundamentalmente en criterios lingüísticos (fonéticos), pero no se hace ninguna hipótesis a priori acerca del mapeado entre observaciones acústicas y unidades lingüísticas. Tal mapeado es aprendido enteramente en la etapa de entrenamiento. Por un lado, la definición lingüística de estas unidades permite realizar la modelización léxica de un modo simple, ya que que puede utilizarse un diccionario diseñado a priori utilizando conocimientos lingüísticos. Por otro lado, las unidades son modeladas acústicamente. Los modelos resultantes son esencialmente descripciones acústicas de las unidades lingüísticas tal como están representadas en las palabras que ocurren en un conjunto de entrenamiento dado, teniendo en cuenta las restricciones de ligadura impuestas por el diccionario. Se podría decir que se trata de unidades pseudolingüísticas. Utilizaremos el término consistencia para designar el grado de homogeneidad del conjunto de segmentos acústicos que son representados por una determinada unidad lingüística. A continuación, describiremos las principales unidades propuestas hasta el momento, ponderando sus ventajas e inconvenientes.

1) Alófono

La elección del alófono (*allophone* o *phone*, en la literatura en lengua inglesa) como unidad subléxica [Bahl 80] presenta varias ventajas, siendo la más importante el tamaño del conjunto de este tipo de unidades. Su número es sustancialmente menor al resto de unidades, de 20 a 60. De esta forma se obtiene un alto número de ocurrencias de cada unidad en el corpus de entrenamiento. Por tanto, son fácilmente entrenables. Por otro lado, permiten modelizar nuevas palabras no presentes en el corpus de entrenamiento del sistema.

Sin embargo, la realización acústica de un alófono puede depender en gran medida del contexto acústico en que ocurre debido a que nuestros órganos articulatorios no pueden moverse instantáneamente de una posición a otra (coarticulación). La utilización de los alófonos como unidades de reconocimiento supone que un alófono en cualquier contexto es equivalente al mismo alófono en cualquier otro contexto, lo cual es totalmente erróneo. En realidad, los modelos de alófonos están muy afectados por el contexto y los parámetros del modelo reflejan muchas señales acústicas diferentes para el mismo alófono, incluyendo demasiada variabilidad acústica. Los modelos de palabra, sin embargo, son más consistentes acústicamente, ya que absorben los efectos contextuales internos a la palabras. Se puede decir que mientras los modelos de palabra carecen de generalidad, los modelos de alófonos generalizan en exceso.

2) Sílabas y semisílabas

Una manera de modelar los efectos coarticulatorios es usar unidades de habla más extensas que los alófonos, como las sílabas [Hunt 80] y las semisílabas [Rosenberg 83] [Mariño 90]. Estas unidades abarcan los grupos de alófonos que contienen los efectos contextuales intrasilábicos, que son los más severos, reduciendo así la variabilidad de los modelos de alófonos y aumentando la consistencia. Sin embargo, en ambos extremos de las sílabas y en un extremo de las semisílabas sigue produciéndose el fenómeno de la coarticulación. Otro problema es el gran número de estas unidades (en castellano hay miles de 500 semisílabas y el número de sílabas es un orden de magniud superior), lo cual crea problemas de entrenabilidad, ya que existen menos pronunciaciones de cada unidad disponibles en la base de datos o corpus necesariamente finito con la que se entrenan los modelos.

Para modelar convenientemente los efectos coarticulatorios entre sonidos adyacentes (la coarticulación entrre sonidos separados por uno o más sonidos suele ser muy débil en la mayor parte de las combinaciones posibles), se han considerado dos definiciones de unidades: modelado explícito de transiciones, o dialófonos, y alófonos dependientes del contexto.

3) Dialófonos

Los dialófonos (*diphones*, en la literatura en lengua inglesa) modelan pares de alófonos sin el uso de sus partes estacionarias, es decir, van del centro de un alófono al centro del siguiente.

Estas unidades incorporan explícitamente todos los fenómenos contextuales y de transición, reduciendo la variabilidad de los modelos de alófonos. En principio, presentan problemas de entrenabilidad debido su gran número (en lugar de N alófonos, hay N^2 dialófonos). Sin embargo, utilizando conjuntamente los alófonos con tan solo los dialófonos que contribuyen a la discriminación se reduce el inventario de unidades a un tamaño adecuado para el entrenamiento y se mejoran los resultados obtenidos usando únicamente alófonos [Cravero 86]. La estrategia consiste en crear modelos de alófonos simples para sonidos relativamente estacionarios, tales como consonantes fricativas y vocales tónicas, y dialófonos en los casos en que los efectos contextuales son significativos, como oclusivas seguidas de vocales o nasales seguidas de vocales.

4) Alófonos dependientes del contexto

Un método para aumentar la consistencia de los alófonos y modelar los efectos contextuales entre sonidos adyacentes consiste en usar un modelo diferente de alófono para cada contexto. Este contexto suele referirse al alófono inmediatamente anterior (alófono dependiente del contexto izquierdo), al posterior (del contexto derecho) o a ambos (de los contextos derecho e izquierdo) (*triphone*, en la literatura en lengua inglesa) [Bahl 80].

Por supuesto, el número de alófonos dependientes del contexto es muy alto (si se consideran 30 alófonos habrá del orden de 27.000 alófonos dependientes de los contextos derecho e izquierdo) y, por tanto, el entrenamiento individualizado de cada uno de ellos da lugar a modelos de baja calidad. Sin embargo, como estos modelos se refieren a alófonos en contextos específicos pueden ser interpolados con modelos de alófonos más generales, mejor entrenados pero menos consistentes. Los pesos de esta interpolación pueden ajustarse manualmente [Chow 86] o de forma automática [Lee K.F. 89].

En este tipo de modelado la consistencia se logra a costa de un nivel de detalle demasiado fino, que no tiene en cuenta la similitud de los efectos contextuales entre sonidos. Una mayor generalización puede conducir a un importante ahorro de memoria y a unos modelos con más alto nivel de entrenabilidad. En la bibliografía se encuentran algunas propuestas basadas en el conocimiento fonético [Derouault 87]. Sin embargo, esta tarea es complicada y tediosa. Otras propuestas agrupan de forma automática modelos correspondientes a contextos similares. Algunos autores utilizan una regla de reducción de unidades basada en el número de ocurrencias en el corpus de entrenamiento, resultando una distribución diferente de alófonos independientes y dependientes de contexto según el umbral de ocurrencias y el corpus de entrenamiento utilizados [Lee C.H. 90]. Lee K.F. realiza un agrupamiento (*clustering*) de modelos utilizando una medida de similitud basada en la cantidad de información perdida cuando dos modelos se mezclan [Lee K.F. 89] (en general, la combinación de algoritmos de agrupamiento y partición (*splitting*) de modelos permite un control preciso del número de ellos

para conseguir un compromiso óptimo entre entrenabilidad y consistencia para un conjunto de entrenamiento dado [Pieraccini 89]).

En la Figura 5 se muestra la transcripción de un texto ortográfico en términos de las unidades subléxicas descritas. En el ejemplo aparecen dos alófonos correspondientes al mismo fonema: la D aproximante y la d oclusiva. También se observan dos alófonos dependientes del contexto correspondientes al mismo alófono: o-s-p y e-s-$.

Texto ortográfico:	dos padres
Alófonos:	$ d o s p a D r e s $
Sílabas:	$ dos pa Dres $
Semisílabas:	$ do os pa a Dre es $
Dialófonos:	$d do os sp pa aD Dr re es s$
Alófonos dep. contexto:	$-d-o d-o-s o-s-p s-p-a p-a-D a-D-r D-r-e r-e-s e-s-$

Figura 4. Transcripción de una frase en términos de unidades subléxicas ($ denota silencio).

5) Alófonos dependientes de la palabra

El contexto también puede referirse a la palabra a la que corresponde el alófono. Se habla entonces de alófonos dependientes de la palabra [Chow 86]. Esta opción suele aplicarse exclusivamente a las llamadas *palabras funcionales.* Se trata de palabras gramaticales (artículos, conjunciones, preposiciones, verbos cortos, etc.) que en el habla continua suelen ser pronunciadas de manera poco clara, con fuertes efectos contextuales que no pueden predecirse del contexto formado por los alófonos vecinos. Son, por tanto, difíciles de reconocer. Estas palabras son muy frecuentes y, por ello, sus errores afectan de forma apreciable a la tasa global de reconocimiento. Sin embargo, esta alta frecuencia y el reducido número de dichas palabras facilitan el modelado de alófonos dependientes de la palabra en estos casos para eliminar estos errores [Lee K.F. 89].

Para finalizar cabe decir que la elección de un conjunto de unidades adecuado para el reconocimiento no es en absoluto un problema resuelto y sistematizado. En una aplicación

determinada, consideraciones prácticas, entre las que hay que destacar la lengua utilizada, determinan la elección para la que se obtiene un mejor compromiso consistencia-entrenabilidad. La combinación de diferentes tipos de unidades puede ser, en muchos casos, la mejor opción.

3.1.2. Composición de palabras

Cuando se utilizan unidades subléxicas, el modelado léxico requiere la construcción de un diccionario que proporcione la transcripción de cada palabra en términos de las unidades seleccionadas. El diccionario puede construirse a priori suponiendo que la pronunciación de cada palabra consiste en una una única secuencia de unidades [Baker 75], como en el ejemplo de la Figura 5. Si se entrenan los modelos con esta hipótesis de una pronunciación por palabra, las pronunciaciones alternativas serán absorbidas en los modelos de unidades. Con este diccionario tan sencillo se pueden obtener resultados razonables, pero los modelos habrán sido contaminados por dichas pronunciaciones.

La alternativa es representar cada palabra del diccionario mediante una red de pronunciaciones. Un método popular de generación de la red es comenzar con la pronunciación clásica de la palabra y a continuación aplicar reglas fonológicas para construir incrementalmente la red [Cohen 74]. Estas reglas pueden insertar, borrar o sustituir unidades en una palabra. Los modelos entrenados a partir de estas redes son más fieles a su definición lingüística que los entrenados con las formas básicas, pero se necesitan muchos más datos aprendizaje.

En el habla continua no suelen existir pausas entre palabras, de forma que el efecto de coarticulación afecta también a la parte final de una palabra y al inicio de la siguiente. Se suelen distinguir dos tipos diferentes de cambios de pronunciación en la unión de dos palabras que pueden ser causantes de errores en el reconocimiento: los cambios fuertes o drásticos, en los que un alófono puede ser omitido o substituido por otro, y los débiles, en los que el alófono alterado se aprecia como una variación del original. Los errores provocados por los cambios fuertes son los menos frecuentes y pueden ser eliminados o minimizados con reglas fonológicas apropiadas. En cambio, para los cambios débiles se ha demostrado que es más apropiado el uso de alófonos dependientes del contexto. Este tipo de unidades se denominan unidades interpalabra, en contraposición con las que consideran los efectos dentro de una palabra, llamadas unidades intrapalabra. En el ejemplo de la Figura 5 se han utilizado ambos tipos de unidades (otra opción hubiera sido considerar silencios entre palabras y utilizar sólo unidades intrapalabra).

3.2 Modelado probabilístico

Para llevar a cabo la modelización de las unidades de reconocimiento descritas existen distintas alternativas [Rabiner 93]: las plantillas empleadas en el alineamiento temporal dinámico (o no-lineal) [Sakoe 78], el modelado probabilístico [Rabiner 89], las redes neuronales [Lippmann 89], etc. La más usada actualmente es el modelado probabilístico con los modelos ocultos de Markov. En el apartado anterior, al hablar de unidades subléxicas, ya estábamos suponiendo implícitamente que eran modeladas de esta forma. En los siguientes apartados describiremos las estructuras de los modelos probabilísticos y los procesos de entrenamiento y reconocimiento utilizados.

3.2.1 Tipo de modelos

Los modelos ocultos de Markov (HMM, *Hidden Markov Models*, en la literatura inglesa) [Rabiner 89] se han convertido en la técnica predominante en reconocimiento del habla debido a la simplicidad de su estructura algorítmica y sus buenas prestaciones. Son una representación de un proceso estocástico que consta de dos mecanismos interrelacionados: una cadena de Markov de primer orden, que permanece oculta y describe la evolución temporal de la señal de voz, y un conjunto de funciones de densidad de probabilidad, asociadas cada una de ellas a un estado (o transición, en algunas variantes) de la cadena, que describen las observaciones o vectores de parámetros de la señal, modelando la naturaleza aleatoria del proceso de producción de voz.

La naturaleza de las probabilidades de generación de observaciones es la diferencia fundamental entre los distintos tipos de modelos. En los modelos discretos, se trata de funciones de densidad de probabilidad discretas, ya que las observaciones toman valores dentro de un conjunto -discreto- de símbolos, obtenidos a partir de los vectores de parámetros de la voz mediante cuantización vectorial. En los modelos continuos, los más usados en el caso de disponer de suficientes datos de entrenamiento, se trata de funciones de densidad de probabilidad paramétricas multivariadas, ya que las observaciones toman valores dentro de un espacio continuo multidimensional. También hay aproximaciones intermedias como los modelos semicontinuos y los de múltiple etiquetado [Hernando 93].

Debido a las fuertes restricciones temporales en el habla, suelen utilizarse modelos de izquierda a derecha, en los que sólo están permitidas las transiciones hacia delante. Una topología muy usada es la de la Figura 6, en que ningún estado puede ser "saltado" por el modelo en su evolución temporal. En otras ocasiones se utiliza una topología más general que permita saltar un estado [Bakis 76].

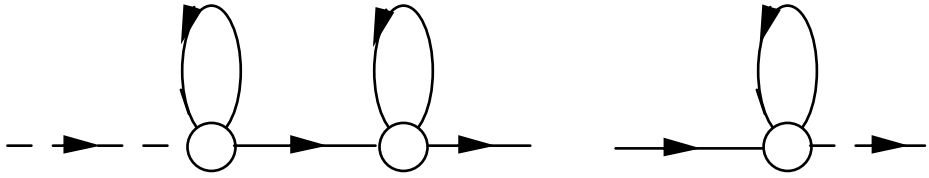

Figura 6. Modelo oculto de Markov de izquierda a derecha

Se han usado modelos como los descritos, con diverso número de estados, para representar palabras [Rabiner 85]. En teoría, un estado puede representar cualquier evento del habla. En la práctica, es difícil escoger el número de estados, y el número óptimo no está de acuerdo normalmente con la intuición. Por tanto, se suelen escoger más estados de los necesarios para la palabra más larga y se usa el mismo modelo para todas las palabras.

Esta misma topología se ha utilizado también para representar cualquiera de las unidades del habla descritas en el apartado anterior. Es muy común utilizar para el alófono la topología de la Figura 6 con tres estados, que representan la transición al alófono, la región estacionaria y la transición hacia fuera del alófono, respectivamente. Recientemente, se ha propuesto [Lee K.F. 89] otro modelo en el que hay además estados y transiciones alternativos para modelar explícitamente duraciones cortas. Si se usa exactamente el mismo modelo para entrenamiento y test, se ha comprobado experimentalmente que la topología de los modelos no es una cuestión crítica.

3.2.2 Entrenamiento y reconocimiento

Una vez definidos el tipo de modelos -discretos, continuos, semicontinuos, etc.- y su topología, se estiman los parámetros de dichos modelos de forma automática a partir de las señales disponibles y su transcripción en términos de las unidades de reconocimiento. A continuación describiremos un proceso de aprrendizaje de modelos continuos típico (ver Figura 7) [Young 92] realizado en varias etapas. Aunque no todas las etapas son necesarias, este esquema tiene la ventaja de englobar distintos tipos de tareas.

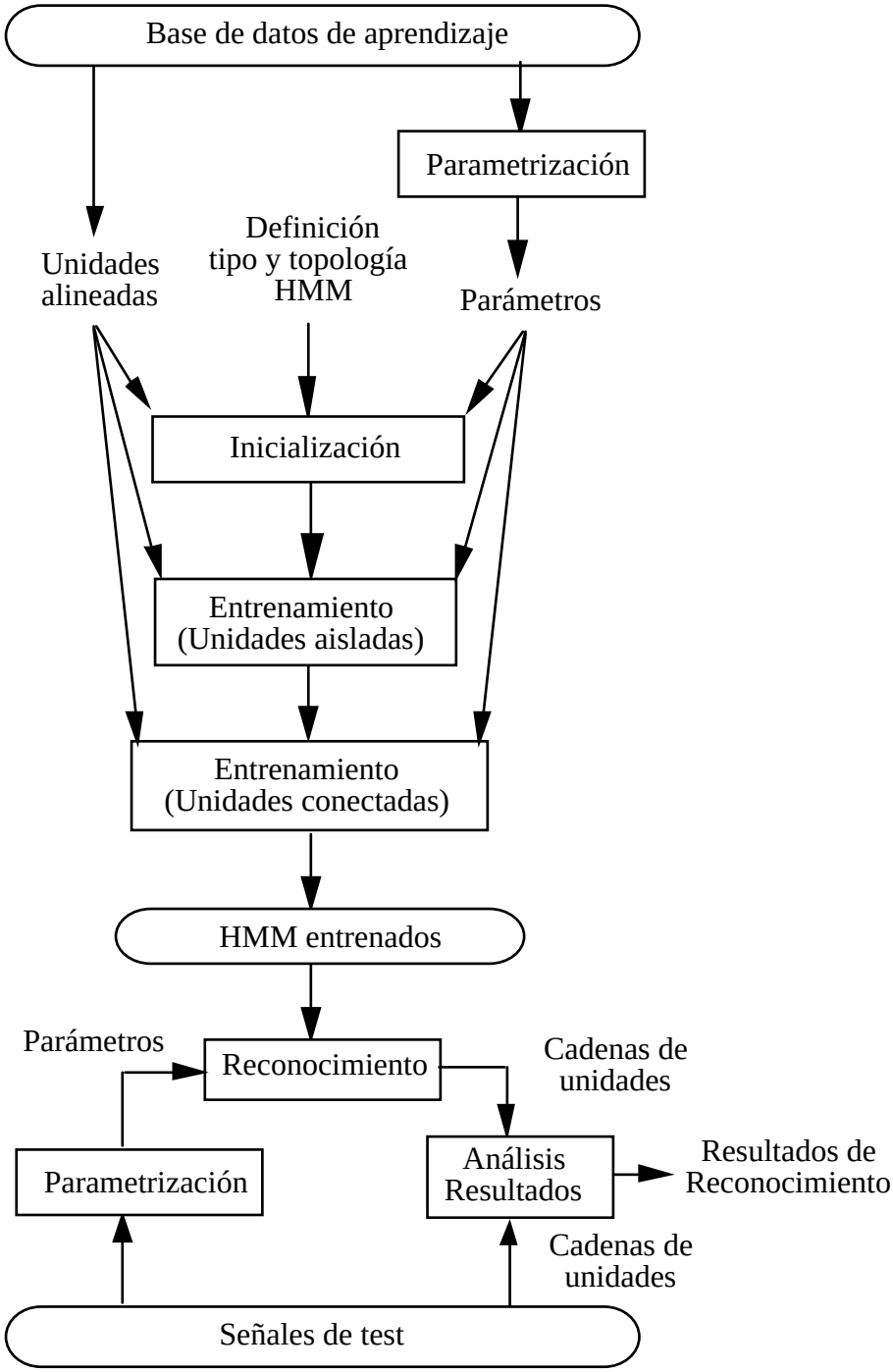

Figura 7. Diagrama de bloques de los procesos de entrenamiento y reconocimiento

1) Inicialización. Se calculan unos valores iniciales para los parámetros de los modelos de las unidades a partir de un corpus (suele ser una parte del corpus general de aprendizaje) del que se dispone su segmentación en términos de dichas unidades. Una forma eficiente de cálculo de obtener una estimación inicial de los parámetros consiste en segmentar uniformemente en estados todas las apariciones en dicho corpus de la unidad correspondiente a cada modelo y

estimar los parámetros del modelo a partir de la distribución obtenida de los vectores de parámetros de señal. Posteriormente, pueden utilizarse los modelos obtenidos para segmentar las señales mediante el algoritmo de Viterbi [Jelinek 76], que proporciona la secuencia de estados más probable dadas las observaciones, volver a calcular los parámetros del modelo a partir de la nueva distribución y así sucesivamente. Este algoritmo iterativo se conoce con el nombre de *segmental k-means* [Rabiner 85].

2) Entrenamiento aislado. A partir de los valores iniciales obtenidos en la etapa anterior y de las apariciones de las unidades en la parte segmentada del corpus, se reestiman los valores de los parámetros de cada modelo mediante el algoritmo iterativo de Baum-Welch [Jelinek 76], que proporciona la estimación de máxima verosimilitud de los parámetros de los modelos dadas las observaciones.

3) Entrenamiento conectado. Finalmente, se utilizan todas los datos del corpus de aprendizaje para reestimar de forma conectada los parámetros de los modelos de las unidades. Si se utilizan unidades subléxicas, en primer lugar se forman los modelos de las palabras concatenando los modelos de las unidades de acuerdo con el diccionario. Seguidamente, en el caso de habla continua, a partir de los modelos de las palabras se construyen los modelos de las frases que corresponden a las señales del corpus de aprendizaje concatenando dichos modelos de acuerdo con la transcripción en palabras. Se suele considerar un modelo opcional de silencio al principio y al final de la señal y otro entre palabras (ver Figura 8). Una vez construidos estos modelos, se aplica el algoritmo de Baum-Welch para reestimar los parámetros de los modelos a partir de todos los datos del corpus. Este entrenamiento conectado permite capturar los efectos contextuales y de transición entre unidades. Además, tiene la ventaja de que no es necesaria la segmentación manual de los datos de entrenamiento en términos de las unidades de reconocimiento, tarea que es tediosa y subóptima. Si se dispone de un conjunto de unidades y el diccionario correspondiente, lo único que se necesita son las señales parametrizadas correspondientes -en general- a frases y su transcripción en palabras.

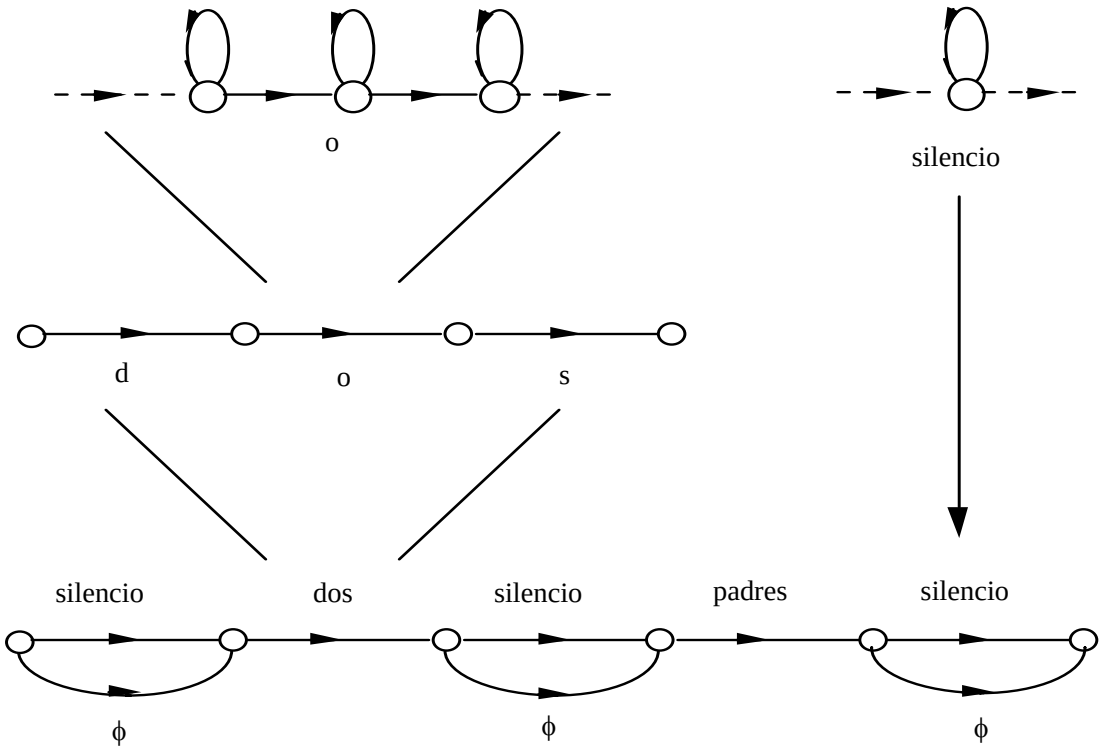

Figura 8. Creación del modelo de una frase a partir de los modelos de las unidades subléxicas y del modelo de silencio. La transición f indica que el silencio es opcional.

En reconocimiento de habla continua a partir de unidades subléxicas, el algoritmo de entrenamiento descrito proporciona unas buenas estimaciones de los modelos de unidades con muy pocas iteraciones en cada etapa. Las dos primeras etapas pueden simplificarse hasta una simple estimación a partir de una segmentación uniforme, aumentando en este caso las iteraciones necesarias en el algoritmo de Baum-Welch. El algoritmo *segmental k-means* presentado en 1) es también aplicable en la etapa 3) de entrenamiento conectado [Rabiner 85]. Se ha descrito el caso más sencillo de un modelo para cada unidad y una única transcripción en unidades para cada palabra, pero el algoritmo se puede generalizar fácilmente. En reconocimiento de palabras aisladas, utilizando la palabra como unidad de reconocimiento, el algoritmo queda reducido a las dos primeras etapas, de las cuales la primera puede reducirse a la segmentación uniforme.

Por un lado, es fundamental para un buen aprendizaje disponer de una base de datos o corpus de habla suficientemente amplio. Por otro lado, aunque idealmente convendría construir modelos acústico-fonéticos que permitieran utilizar el sistema en cualquier aplicación, resultan más efectivos modelos especializados en la tarea concreta de reconocimiento, formados a partir de un corpus adaptado a dicha tarea. De todas formas, ello no invalida la conveniencia de

disponer de bases de datos extensas que cubran fonéticamente las lenguas en las que vaya a operar un sistema, a fin de conseguir unos primeros modelos de tipo general que luego pueden ser adaptados a la aplicación específica.

Este es el objetivo -por ejemplo- de la base TIMIT [Garofolo 89] para el inglés. En castellano existirá próximamente la base denominada ALBAYZIN [Casacuberta 92], el corpus fonético de la cual ha sido diseñado para cumplir rigurosamente unos criterios de ocurrencia de los alófonos y sus contextos [Moreno 93]. Dichos criterios obedecen a un compromiso entre dos objetivos contradictorios: 1) la obtención de un número mínimo de ocurrencias, para posibilitar un entrenamiento adecuado; y 2) la similitud del número de ocurrencias en el corpus con el del habla natural -es decir, la empleada comúnmente.

La tarea de reconocimiento (indicada en la parte inferior de la Figura 7) se lleva a cabo con el algoritmo de Viterbi, teniendo en cuenta los modelos HMM de las unidades de habla, el diccionario de palabras y la gramática. Cuando la tarea presenta cierta complejidad, la gramática suele ser de tipo probabilístico, como la *N-grama*, que asigna una probabilidad a cada combinación de *N* palabras.

Durante el desarrollo de un sistema es importante evaluar sus prestaciones a fin de optimizarlas o, simplemente, para comparar los distintos sistemas. La evaluación puede realizarse a nivel de unidades subléxicas, de palabras o de frases. Cuando se trata de evaluar el sistema fijándose en las elocuciones enteras (palabras o frases) la medida resultante proviene simplemente de un conteo. Sin embargo, en el caso de querer evaluar el sistema a nivel de las unidades que se encadenan, se consideran tres tipos distintos de errores: sustitución de la unidad correcta por otra, borrado o supresión de la unidad correcta e inserción (incorrecta) de una unidad en la cadena. Habitualmente se emplea la siguiente tasa percentil de error de reconocimiento basada en la anterior clasificación de errores [Young 92]:

$$precision = \frac{C - I}{T} 100\% \qquad (9)$$

donde *C*, *I* y *T* son, respectivamente, el número de aciertos, el de inserciones y el número total de unidades presentes en el test.

4. CONCLUSION

Hemos descrito en este capítulo las principales técnicas empleadas en la construcción de modelos para el reconocimiento acústico-fonético del habla, es decir, la conversión de una señal de voz en una cadena de palabras. Para ello hemos estudiado tanto el procesado de señal

que permite la extracción de los parámetros de la voz como el propio modelado probabilístico de las unidades. Aunque la línea explicativa apunta al reconocimiento del habla continua con vocabularios de gran tamaño, lo cual requiere el uso de unidades subléxicas, se han considerado también las tareas más simples, que permiten usar la palabra como unidad elemental (vocabulario reducido) o requieren un sistema menos complejo (palabras aisladas). Como es fácil imaginar, existe un gran número de técnicas publicadas en cada uno de los temas tratados. No obstante, el desarrollo de la explicación ha priorizado más la exposición -relativamente- pausada de las técnicas consideradas más relevantes que la enumeración o clasificación de todas las publicadas.

REFERENCIAS

[Atal 67] B.S. Atal y M.R. Schroeder, "Predictive Coding of Speech Signals", Proc. Conf. Commun. and Process., pp. 360-361, 1967.

[Bahl 80] L.R. Bahl, R. Bakis, P.S. Cohen, A.G. Cole, F. Jelinek, B.L. Lewis, R.L. Mercer, "Further Results on the Recognition of a Continuously Read Natural Corpus", Proc. ICASSP, Abril 1980.

[Baker 75] J.K. Baker, *Stochastic Modeling as a Means of Automatic Speech Recognition*, PhD Thesis, Computer Science Department, Carnegie Mellon University, Abril 1975.

[Bakis 76] R. Bakis, "Continuous Speech Recognition via Centisecond Acoustic States", 91st Meeting of the Acoustical Society of America, Abril 1976.

[Boll 79] S.F. Boll, "Supression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. ASSP, vol. 27, nº 2, pp. 113-120, 1979.

[Casacuberta 92] F. Casacuberta, R. García, J. Llisterri, C. Nadeu, J.M. Pardo, A. Rubio, “Desarrollo de corpus para investigación en tecnologías del habla”, Boletín SEPLN, No. 12, pp. 35-42, Julio 1992.

[Chow 86] Y.L. Chow, R.M. Schwartz, S. Roucos, O.A. Kimball, P. Price, G.F. Kubala, M.O. Dunham, M.A. Krasner, J. Makhoul, "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System", Proc. ICASSP, Tokio, Abril 1986.

[Cohen 74] P.S. Cohen, L.R. Mercer, "The Phonological Component of a Automatic Speech Recognition System", Proc. IEEE Synposium on Speech Recognition, Pittsburgh, PA, pp. 177-187, 1974.

[Cravero 86] M. Cravero, R. Pieraccini, F. Raineri, "Definition and Evaluation of Phonetic Units for Speech Recognition by Hidden Markov Models", Proc. ICASSP, Tokio, pp. 2235-8, 1986.

[Dautrich 83] B.A. Dautrich, L.R. Rabiner, T.B. Martin, "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition", IEEE Trans. ASSP, vol. 31, pp. 793-806, 1983.

[Davis 80] S.B. Davis y P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. ASSP, vol. 28, pp. 357-366, 1980.

[Derouault 87] A.M. Derouault, "Context-Dependent Phonetic Markov Models for Large Vocabulary Speech Recognition", Proc. ICASSP, Dallas, Abril 1987, pp. 360-363.

[Duda 73] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, Vol. I, John Wiley & Sons, 1973.

[Furui 86] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Trans. ASSP, vol. 34, nº 1, pp. 52-59, 1986.

[Garofolo 89] J.S. Garofolo, D.S. Pallet, "Use of CD-ROM for Speech Database Storage and Exchange", EUROSPEECH'89, pp. 309-12, Sept. 1989.

[Ghitza 86] O. Ghitza, "Auditory Nerve Representation as a Front-End for Speech Recognition in a Noisy Environment", Computer Speech and Language, nº 1, pp. 109-130, 1986.

[Hanson 87] B.A. Hanson, H. Wakita, "Spectral Slope Based Distortion Measures for All-Pole Models of Speech", IEEE Trans. ASSP, vol. 35, nº 7, pp. 968-973, 1987.

[Hermansky 91] H. Hermansky, N. Morgan, A. Bayya, P. Kohn, "Compensation for the Effect of the Communication Channel in Perceptual Linear Predictive (PLP) Analysis of Speech", Proc. EUROSPEECH-91, pp. 1367-1370, Génova, Septiembre 1991.

[Hernando 93] J. Hernando, *Técnicas de Procesado y Representación de la Señal de Voz para el Reconocimiento del Habla en Ambientes Ruidosos*, Tesis Doctoral, Universitat Politècnica de Catalunya, Barcelona, Mayo 1993.

[Hirsch 91] H.G. Hirsch, P. Meyer, H.W. Ruehl, "Improved Speech Recognition Using High-Pass Filtering of Subband Envelopes", Proc. EUROSPEECH-91, pp. 413-6, Génova, Septiembre 1991.

[Hunt 80] M.J. Hunt, M. Lennig, P. Mermelstein, "Experiments in Syllable-Based Recognition of Continuous Speech", Proc. ICASSP, Abril, 1980, pp. 880-883.

[Itakura 75] F. Itakura, "Minimum prediction residual principle applied to speech recognition", IEEE Trans. ASSP, vol. 35, 1975, pp. 67-72.

[Jelinek 76] F. Jelinek, "Continuous Speech Recognition by Statistical Methods", Proc. IEEE, vol. 64, pp. 532-536, Abril 1976.

[Juang 87] B. H. Juang, L.R. Rabiner, J. G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition", IEEE Trans. ASSP, vol. 35, nº 7, pp. 947-953, 1987.

[Juang 91] B.H. Juang, "Speech Recognition in Adverse Environments", Computer Speech and Language, vol. 5, pp. 275-294, 1991.

[Klatt 79] D.H. Klatt, "Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access", Journal of Phonetics, vol. 7, pp. 279-312, 1979.

[Lee C.H. 89] C.H. Lee, B.H. Juang, F.K. Soong, L.R. Rabiner "Word Recognition using Whole Word and Subword Models", Proc. ICASSP, Abril 1989, pp. 683-686.

[Lee C.H. 90] C-H. Lee, L.R. Rabiner, R. Pieraccini, "Speaker Independent Continuous Speech Recognition Using Continuous Density Hidden Markov Models", Speech Recognition and Understanding. NATO ASI Series,vol.75, Italia, 1990.

[Lee K.F. 89] K.F. Lee, "Automatic Speech Recognition. The Development of the SPHINX System", Kluwer Academic Publishers, 1989.

[Lippmann 89] R.P. Lippmann, "Review of Neural Networks for Speech Recognition", Neural Computation, 1989, pp. 1-38.

[Lleida 90] E. Lleida, *Compresión de Información en Reconocimiento del Habla*, Tesis Doctoral, Universitat Politècnica de Catalunya, Barcelona, 1990.

[Mansour 89] D. Mansour y B. H. Juang, "A Family of Distortion Measures Based upon Projection Operation for Robust Speech Recogniton", IEEE Trans. ASSP, vol. 37, nº 11, pp. 1659-1671, 1989.

[Mariño 90] J.B. Mariño, A. Bonafonte, A. Moreno, E. Lleida, C. Nadeu, E. Monte, "Recognition of Numbers by Using Demisyllables and Hidden Markov Models", Proc. EUSIPCO, Barcelona, Sept. 1990, pp. 1363-1366.

[Marple 87] S.L. Marple, Jr., *Digital Spectral Analysis with Applications*, Prentice-Hall, Inc., 1987.

[Moreno 93] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Mariño, C. Nadeu, “Albayzin speech data base: design of the phonetic corpus”, EUROSPEECH’93, Berlín, pp. 175-8, Sept. 1993.

[Nadeu 94] C. Nadeu, B.H. Juang, “Filtering spectral parameters for speech recognition” (próxima publicación).

[Oppenheim 89] A.V. Oppenheim, R.W. Schafer, *Discrete-TimeSignal Processing*, 2a. Ed., Prentice-Hall, 1989.

[O'Shaugnessy 90] D. O'Shaughnessy, *Speech Communcation, Human and Machine*, Addison-Wesley Series in Electrical Engineering: Digital Signal Processing, 1990.

[Papoulis 91] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3a. Ed., McGraw-Hill, 1991.

[Pieraccini 89] R. Pieraccini, A.E. Rosenberg, "Automatic Generation of Phonetic Units for Continuous Speech Recognition", Proc. ICASSP, Glasgow, Mayo 1989, pp. 623-626.

[Rabiner 85] L.R. Rabiner, B.H. Juang, S.E. Levinson, M.M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities", AT&T Technical Journal 64 (6), Julio-Agosto 1985, pp. 1211-33.

[Rabiner 89] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. IEEE, vol. 77, 1989, pp. 257-286.

[Rabiner 93] L.R. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[Rosenberg 83] A.E. Rosenberg, L.R. Rabiner, J. Wilpon, D. Kahn, "Demysillable-Based Isolated Word Recognition System", IEEE Trans. ASSP, vol. 31, Junio 1983, pp. 713-726.

[Sakoe 78] H. Sakoe, S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Words Recognition", IEEE Trans. ASSP, vol. 26, 1978, pp. 43-49.

[Shiraki 88] Y. Shiraki, M. Honda, “LPC Speech Coding based on Variable-Length Segment Quantization”, IEEE Trans. ASSP, Vol. 36, No. 9, Sept. 1988, pp. 1437-44.

[Svendsen 87] T. Svendsen, F.K. Soong, " On the Automatic Segmentation of Speech Signals", Proc. ICASSP, Dallas, Abril 1987, pp. 77-80.

[Tohkura 87] Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition", IEEE Trans, ASSP, vol. 35, nº 10, Octubre 1987.

[Vidal 90] E. Vidal, A. Marzal, “A Review and New Approaches for Automatic Segmentation of Speech Signals”, Proc. EUSIPCO, Barcelona, Sept. 1990, pp.

[Viswanathan 88] V.Viswanathan, C. Henry, "Evualuation of Multisensor Speech Input for Speech Recognition in High Ambient Noise", Proc. ICASSP, Tokio, Abril 1986, pp. 85-88.

[Widrow 75] B. Widrow, J.R. Glover, Jr., J.M. McCool, J. Kaunitz, C.S. Williams, R.H. Hearn, J.R. Zeidler, E. Dong, Jr., R.C. Goodlin, "Adaptive Noise Cancelling: Principles and Applications", Proc. IEEE, vol. 63, nº 12, 1975, pp.1692-1716.

[Young 93] S. Young, "HTK: Hidden Markov Model Toolkit V1.5", Cambridge University Engineering Department, Speech Group, Mayo 1993.