UNIVERSIDAD POLITECNICA DE CATALUÑA

Departamento de Teoria de la señal y comunicaciones

TECNICAS DE PROCESADO Y REPRESENTACION DE LA SEÑAL DE VOZ PARA EL RECONOCIMIENTO DEL HABLA EN AMBIENTES RUIDOSOS

Autor: Francisco Javier Hernando Pericas
Director: Climent Nadeu i Camprubi

Barcelona, mayo 1993

Capítulo 6

RESULTADOS EXPERIMENTALES

En este capítulo se describen las pruebas experimentales más importantes realizadas en este trabajo para evaluar y comparar las distintas técnicas de reconocimiento del habla en ambiente ruidosos descritas en los capítulos anteriores.

Como ya se ha mencionado, para evaluar el comportamiento de dichas técnicas se ha optado por la implementación de un sistema de reconocimiento de palabras aisladas mediante modelos ocultos de Markov. El hecho de que el sistema sea de palabras aisladas permitirá prescindir de las implicaciones de los niveles de conocimiento superiores: sintáctico, semántico, pragmático,... Por otro lado, los modelos ocultos de Markov son los que en estos momentos proporcionan unas mejores prestaciones en todos los sistemas en desarrollo.

Además, para centrar el trabajo en el problema del ruido, se ha optado por realizar pruebas multilocutor, usando un número reducido de locutores, con un vocabulario pequeño y de poca confusibilidad, como es el de los dígitos.

En cuanto al tipo de ruido interferente, las primeras pruebas se han realizado añadiendo a la señal limpia, que se había utilizado para entrenar las referencias, ruido blanco, idealización que puede tener sentido para una parte, al menos, del ruido real. A continuación se han llevado a cabo pruebas con ruido real y efecto Lombard, utilizando

señales grabadas en el interior de un coche en diferentes condiciones de ruido: diferentes velocidades del vehículo y potencias del ventilador de la calefacción.

La estructura del capítulo es la siguiente. En el apartado 6.1 se describen con detalle las bases de datos y los tipos de pruebas antes mencionados. El apartado 6.2 está dedicado a la descripción del sistema de reconocimiento básico. En el apartado 6.3 se recogen los resultados obtenidos con ruido blanco. En primer lugar, en el apartado 6.3.1 se optimiza el sistema básico; seguidamente, se evalúan los distintos tipos de técnicas: parametrizaciones (apartado 6.3.2) y distancias alternativas (6.3.3), incorporación de energía y parámetros dinámicos (6.3.4), suavizado (6.3.5) y utilización de modelos semicontinuos y de múltiple etiquetado (6.3.6); y en el apartado 6.3.7 se analizan los resultados obtenidos combinando varias de estas técnicas. Por último, el apartado 6.4 está dedicado a los resultados obtenidos con ruido real de coche.

6.1. BASES DE DATOS Y TIPOS DE PRUEBAS

En este apartado se describen las bases de datos que se utilizarán para las pruebas realizadas con ruido blanco y ruido de coche, respectivamente. Debido a las diferentes características de ambas bases de datos se han diseñado distintos tipos de pruebas de reconocimiento para cada caso.

6.1.1. RUIDO BLANCO

La base de datos está formada por 10 locutores, 3 mujeres y 7 hombres, cada uno de los cuales ha pronunciado 10 veces los 10 dígitos en catalán. Hay, pues, 100 pronunciaciones distintas de cada uno de los 10 dígitos, lo que constituye una base de 1000 palabras. Esta palabras fueron grabadas en el laboratorio en condiciones que se han considerado libres de ruido.

El entrenamiento del sistema se realizó siempre con estas señales limpias. En primer lugar, estas señales fueron filtradas paso-banda de 100 a 3400 Hz con un filtro antialiasing, muestreadas de 8 kHz y cuantificadas con 12 bits por muestra. Debido a que los errores en la detección de principio y fin de las palabras pueden provocar una importante disminución en la tasa de reconocimiento, la señal de voz limpia digitalizada fue marcada manualmente para determinar los límites de cada palabra y estos límites fueron usados en todos los reconocimientos (ver figura 6.1).

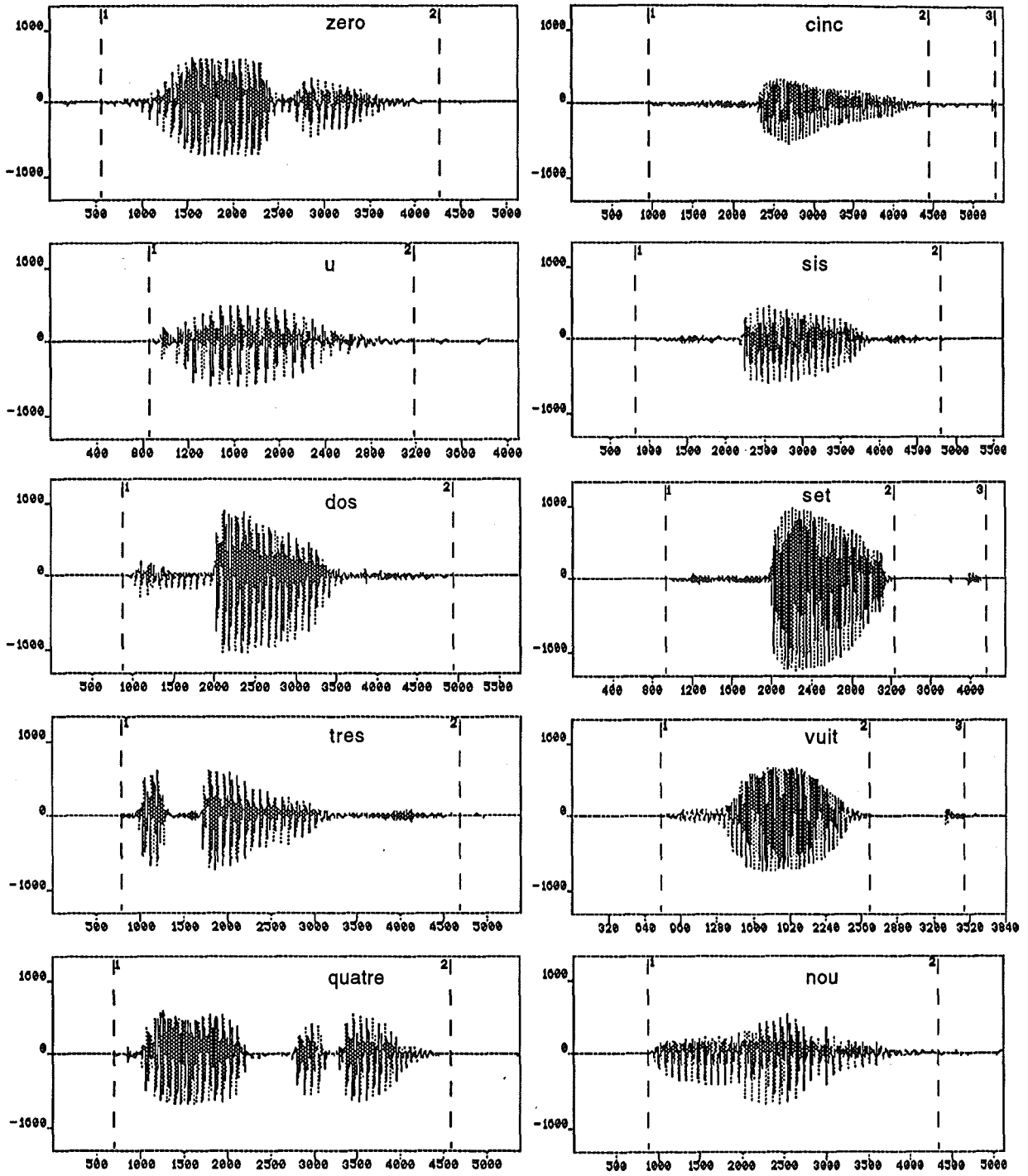

Fig. 6.1. Evolución temporal de las señales correspondientes a los diez dígitos en catalán.

En las palabras acabadas en consonante oclusiva -"cinc", "set", "vuit"- se consideró preciso poner otra marca, aparte de las de principio y fin de palabra, antes de la oclusiva final, ya que esta no se pronuncia en muchas ocasiones y cuando se pronuncia se hace de manera muy débil. Teniendo en cuenta que las tasas de reconocimiento considerando estas consonantes oclusivas fueron similares a las tasas obtenidas sin considerarlas, se optó finalmente por no incluir estas consonantes oclusivas en las pruebas experimentales aunque se hubieran pronunciado.

En la figura 6.1 se representa la evolución temporal de los 10 dígitos catalanes, correspondientes a distintas versiones de distintos locutores, escogidas para observar estas consonantes oclusivas comentadas.

En la fase de reconocimiento se utilizaron señales limpias, correspondientes a la base de datos original, y señales ruidosas, que se simularon sumando ruido blanco gaussiano de media nula a las señales limpias de manera que se obtuviesen las relaciones señal-ruido (SNR) deseadas. Hay que tener en cuenta, no obstante, que el ruido sumado a la señal es sólo una aproximación al ruido blanco gaussiano ideal, ya que consta de un número finito de muestras.

En concreto, se consideraron tres SNR distintas: 20, 10 y 0 dB. Estas dos últimas SNR pueden considerarse condiciones severas de ruido, ya que dan lugar a una fuerte degradación del comportamiento de un sistema de reconocimiento automático del habla que no haya sido diseñado teniendo en cuenta el problema de la robustez frente al ruido.

La relación señal-ruido (SNR) se definió de forma global, es decir, como cociente - expresado en dB- entre la potencia media de la señal correspondiente a la palabra entera sin ruido y la potencia del ruido, siendo esta última constante a lo largo de la señal. Evidentemente, esta SNR se cumple para la señal en su conjunto pero no para cada trama en particular. En concreto, las tramas de mayor potencia tendrán una SNR más elevada. La otra opción hubiera sido definir la SNR de forma segmental, es decir, sumar ruido de forma que todas las tramas tuvieran la misma SNR. Sin embargo, las SNR referidas serán globales, consideración más aproximada a la realidad.

Como ya se ha comentado, las marcas de inicio y fin de palabra correspondientes a las señales limpias se utilizaron en todos los reconocimientos, incluidos aquellos en que se contaminó la señal limpia con ruido blanco aditivo para obtener señal ruidosa. De este modo se elimina la influencia de los errores de detección de voz en la tasa de

reconocimiento, que es tanto más importante cuanto menor es la SNR de la señal de voz, y pueden estudiarse aisladamente los errores ocurridos en el propio proceso de reconocimiento.

Utilizando estas señales, se realizaron pruebas multilocutor. Para ello, se dividió la base de datos en dos bloques, cada uno de los cuales con la mitad de las versiones de cada dígito y locutor; es decir, 500 señales correspondientes a 5 versiones de los 10 dígitos de los 10 locutores. Con uno de los bloques se entrenó el sistema, el cuantificador vectorial y los diez modelos de cada palabra, y el otro bloque fue objeto de reconocimiento, en condiciones libres de ruido y en condiciones ruidosas. Por problemas de coste computacional, sólo se hizo una partición de la base y se realizaron dos pruebas alternando los papeles de de cada uno de los bloques. En este tipo de pruebas cada palabra es reconocida una sola vez en las mismas condiciones de ruido, por lo cual en cada tasa de reconocimiento presentada en el apartado de resultados con ruido blanco 6.3 han contribuido 1000 palabras reconocidas.

6.1.2. RUIDO DE COCHE

La base de datos está formada por 4 locutores, 2 mujeres y 2 hombres, cada uno de los cuales ha pronunciado 25 veces los 10 dígitos en italianos en el interior de un automóvil (Fiat Tipo) y en diferentes condiciones de ruido. Hay, pues 100 pronunciaciones distintas de cada uno de los diez dígitos, lo que constituye una base de 1000 palabras. Esta base procede del proyecto ESPRIT-ARS y pudo ser utilizada gracias a la intermediación de los Sres. G. Babini, de la compañía italiana CSELT, y J. Gómez Mena, de la Universidad Politécnica de Madrid.

La señal original había sido muestreada a 16 kHz y cuantificada con 16 bits por muestra y estaban marcados el inicio y el fin de cada palabra. Para poder comparar los resultados obtenidos con la base de datos anterior, las señales fueron filtradas paso-bajo con filtro antialiasing de frecuencia de corte discreta inferior a 1/4 y diezmadas por 2, se eliminaron los cuatro bits menos significativos de cada muestra y se copiaron las marcas de inicio y fin de cada palabra.

De estas 25 repeticiones de cada dígito y cada locutor: 5 repeticiones se realizaron con el coche parado y el motor y el ventilador parados, 10 con el coche parado, el motor en marcha y diferentes velocidades del ventilador, 5 con el coche

circulando a 70 km/h y diferentes velocidades del ventilador y 5 con el coche circulando a 130 km/h y diferentes velocidades del ventilador.

El entrenamiento del sistema, un cuantificador vectorial y un modelo para cada una de las diez palabras, se realizó con las 5 repeticiones de los 10 dígitos de cada locutor pronunciadas con el coche, el motor y el ventilador parados; es decir, un total de 200 señales. Para evitar los errores de detección de voz, se utilizaron las marcas de inicio y fin de palabra de la base de datos.

En la fase de reconocimiento intervinieron el resto de las señales de las bases de datos y también se utilizaron las marcas de inicio y fin de palabra. Se han presentado los resultados agrupados por la velocidad del vehículo y no por la velocidad del ventilador, pues se ha comprobado que el primer factor es el más importante a la hora de estudiar la degradación del comportamiento del sistema. Es conveniente hacer notar que en cada tasa de reconocimiento presentada en el apartado de resultados, han contribuido 400 señales en el caso de v = 0 (con el motor en marcha y diferentes velocidades del ventilador), 200 señales en el caso de v = 70 km/h (con diferentes velocidades del ventilador) y 200 señales en el caso de v =130 km/h (con diferentes velocidades del ventilador).

6.2. SISTEMA BASICO DE RECONOCIMIENTO

La etapa de parametrización del sistema básico de reconocimiento consta de los siguientes procesos (ver figura 6.2): preénfasis de factor a, división de la señal en tramas de 30 ms de duración con un desplazamiento de 15 ms, enventanado de la trama usando la ventana de Hamming, aplicación del método de autocorrelación de la técnica clásica de predicción lineal de orden p (cálculo de las p+1 primeras autocorrelaciones utilizando el estimador sesgado clásico de la autocorrelación y utilización del algoritmo de Levinson-Durbin para hallar los coeficientes del predictor a_k), obtención de L parámetros cepstrales LPC c(n) a partir de los p coeficientes de predicción mediante la recursión (4.8)-(4.10) y ponderación de dicho vector de características mediante una ventana cepstral w(n): rectangular, seno realzado, inversa de la desviación típica o rampa.

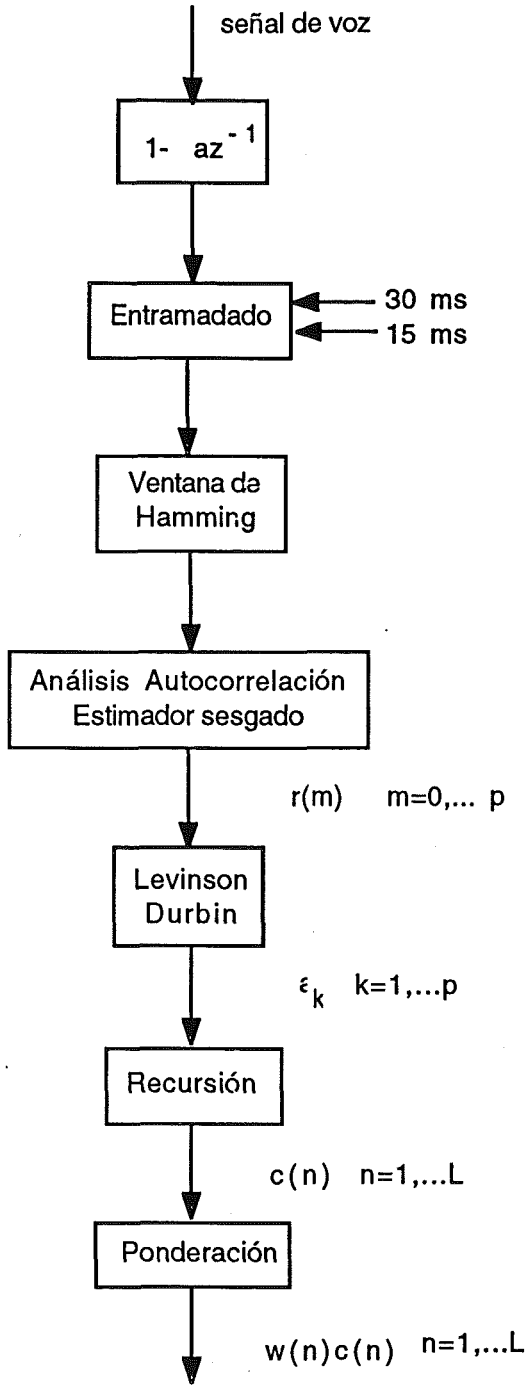

Fig. 6.2. Etapa de parametrización del sistema básico

Los vectores cepstrales ponderados son cuantificados vectorialmente utilizando distancia euclídea y un diccionario de M palabras-código que se ha construido previamente mediante el algoritmo jerárquico descrito en la figura 5.9. En la construcción del diccionario se utiliza también distancia euclídea y el criterio de convergencia se basa en la disminución relativa de la distorsión media, inferior a 1 %. Por último, se utiliza un valor de μ aleatorio de distribución uniforme entre -0.01 y 0.01 para la perturbación de los centroides.

Cada palabra se caracteriza por un modelo de Markov discreto de izquierda a derecha de N estados. Se utiliza una inicialización aleatoria de las matrices A y B y se fuerza que la secuencia comience en el estado 1, $\pi_1 = 1$ y $\pi_i = 0$ para $i \neq 1$, y termine en el estado N, $\beta_T(N) = 1$ y $\beta_T(i) = 0$ para $i \neq N$. El entrenamiento se realiza mediante el algoritmo de Baum-Welch para múltiples secuencias de observaciones, utilizando escalado dinámico de probabilidades, y la decisión de finalización del algoritmo se realiza en base a la variación relativa de la probabilidad de generación, de forma que se itera hasta que esta no rebasa el 1%. Posteriormente, se aplica la técnica de floor-smoothing con un valor de $\delta = 10^{-3}$. La evaluación de las probabilidades a posteriori se realiza mediante el algoritmo de Viterbi, utilizando compresión logarítmica de probabilidades.

6.3. RESULTADOS OBTENIDOS CON RUIDO BLANCO

En este apartado se exponen los resultados obtenidos con ruido blanco aplicando las técnicas descritas en esta memoria. En primer lugar, en el apartado 6.3.1 se optimiza el sistema el sistema básico. Seguidamente, se evalúan los distintos tipos de técnicas: parametrizaciones alternativas, distancias alternativas, incorporación de energía y parámetros dinámicos, suavizado y utilización de modelos semicontinuos y de múltiple etiquetado. Finalmente, en el apartado 6.3.7 se analizan los resultados obtenidos combinando varias de estas técnicas.

6.3.1. OPTIMIZACION DEL SISTEMA BASICO

El objetivo de este apartado es optimizar los parámetros del sistema básico de reconocimiento descrito en el apartado 6.2 desde el punto de vista de robustez al ruido blanco aditivo: grado de preénfasis a, estructura y número de estados N del modelo de

Markov, tamaño M del diccionario del cuantificador vectorial, orden de predicción lineal p y ventana de ponderación cepstral w(n) de longitud L.

Para poder llevar a cabo esta optimización realizando un número de pruebas razonables, estos parámetros fueron optimizados normalmente por separado y no de forma conjunta [Her91b]. Para ello, se utilizó la experiencia acumulada en reconocimiento automático del habla en el grupo de investigación en que se ha realizado esta tesis.

Por lo que se refiere al número de palabras-código del diccionario del cuantificador vectorial, en el grupo de investigación en que se ha realizado esta tesis, en la aplicación del reconocimiento de los números en castellano mediante semisílabas [Mar90], se suele trabajar con un valor de 64 para cuantificar los vectores cepstrales. En pruebas realizadas en esta tesis, con la base de datos y el tipo de pruebas elegidos para los experimentos con ruido blanco descritos en el apartado 6.2, se observó que esta era una buena elección.

En cuanto a la estructura del modelo de Markov, en [Mar90] se trabaja con modelos de izquierda a derecha, con un número de estados que depende de la duración media de cada semisílaba y permitiendo solamente transiciones al estado actual y a los dos siguientes. Se trata, por tanto, de modelos de Bakis con $\Delta = 2$ (ver apartado 5.2.3.1). En pruebas realizadas en esta tesis con los dígitos catalanes se observó que en presencia de ruido era aconsejable un número relativamente alto de estados e imponer $\Delta = 1$, es decir, sólo permitir transiciones al estado actual o al siguiente. En este trabajo este tipo de modelos se denotarán como modelos sin saltos, ya que obligan a generar observaciones en todos los estados. El compromiso coste computacional-tasa de reconocimiento llevó a considerar modelos de 10 estados.

Por otro lado, trabajos preliminares de reconocimiento de dígitos en presencia de ruido blanco realizados mediante la técnica de comparación de patrones muestran que la no utilización de preénfasis y el uso de ventanas cepstrales crecientes y órdenes relativamente elevados de predicción lineal robustecen el sistema de reconocimiento [Ras90]. Las pruebas realizadas en esta tesis confirmaron la no conveniencia de preénfasis y, entre los órdenes de predicción y los tipos de ventanas cepstrales considerados, resultó que la elección óptima era orden 12 y ventana rampa.

6.3.1.1. PREENFASIS

Como se ha comentado en apartado 3.2.2 de esta memoria, es común la aplicación de un filtro paso-alto de primer orden llamado de preénfasis sobre la señal de voz digitalizada, cuya función de transferencia es $H(z) = 1 - az^{-1}$.

Este filtro tiene un cero de transmisión que depende del valor de a, tal como se muestra en la figura 6.3. Para a = 0 el filtro es plano y para a = 1 existe un cero de transmisión en la frecuencia cero. En la práctica, en condiciones libres de ruido, se suele utilizar a = 0.95, con lo cual se enfatiza la zona de altas frecuencias y se desenfatizan las bajas frecuencias. Debido a que el espectro de la señal de voz suele tener una caída de 6 dB por octava, con ello se consigue reducir el margen dinámico del espectro y disminuir así problemas de precisión numérica en la implementación.

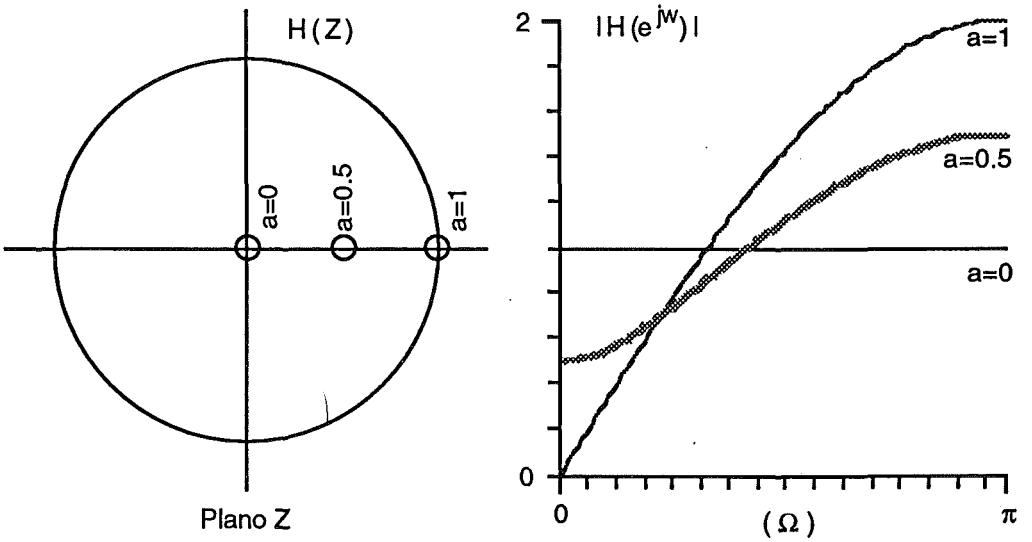

Fig. 6.3. Representación de H(Z) en el plano Z y de su módulo en el círculo unidad para distintos valores de a.

Sin embargo, en el caso de que la voz esté perturbada por la presencia de ruido blanco, el efecto del preénfasis ya no es deseable. El ruido blanco es espectralmente plano, mientras que la señal de voz tiene concentrada su energía en las bajas frecuencias. En consecuencia, si se aplica un filtro paso-alto (por ejemplo, a = 0.95)

resulta que se realza la zona del espectro en que el ruido presenta mayor potencia relativa respecto a la señal.

En la tabla 6.1 se muestran las tasas de reconocimiento expresadas en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral rampa de orden 12, con preénfasis (a=0.95) y sin preénfasis (a=0).

Preénf. /SNR(dB)	sin ruido	20	10	0
a = 0	99.8	98.9	89.5	54.2
a = 0.95	99.7	98.3	84.7	44.0

Tabla 6.1. Influencia del preénfasis

Puede observarse que se obtienen mejores resultados cuando no se realiza preénfasis y este efecto es más acusado a medida que disminuye la relación señal-ruido.

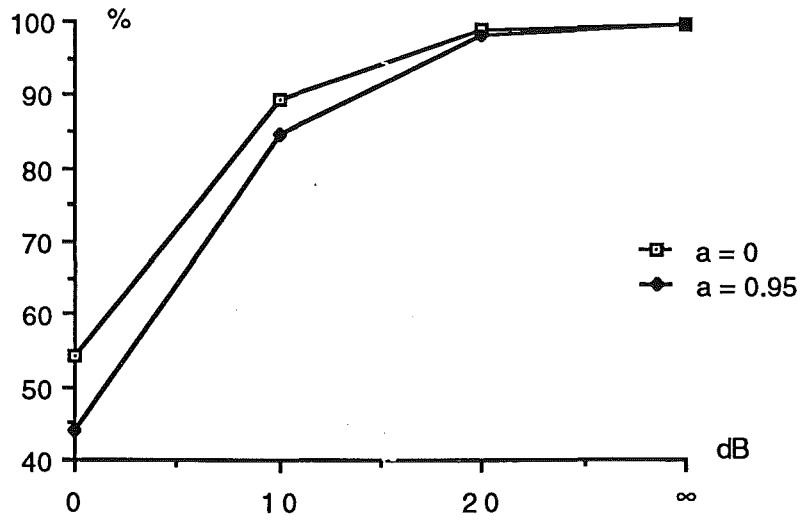

Fig. 6.4. Influencia del preénfasis

En la figura 6.4, se representan gráficamente los resultados de la tabla 6.1. A partir de estos resultados, las pruebas siguientes con ruido blanco se realizarán siempre sin preénfasis. En las figuras, la condición de ausencia de ruido se denotará como SNR = ∞.

En el caso de condiciones libres de ruido, tendría que ser mejor el resultado en el caso de utilizar preénfasis. Sin embargo, se ha de tener en cuenta que no se puede asignar una fiabilidad estadística importante a una diferencia de una décima en estas tasas, ya que sólo representa un error absoluto en un resultado de reconocimiento.

6.3.1.2. ESTRUCTURA DEL MODELO DE MARKOV

En la figura 6.5. se muestran las tasas de reconocimiento en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, sin préenfasis, utilizando un diccionario de 64 palabras-código, ventana cepstral rampa de orden 12 y modelos de izquierda a derecha en que sólo se permiten transiciones al estado actual y al siguiente, en función de N, el número de estados de cada modelo.

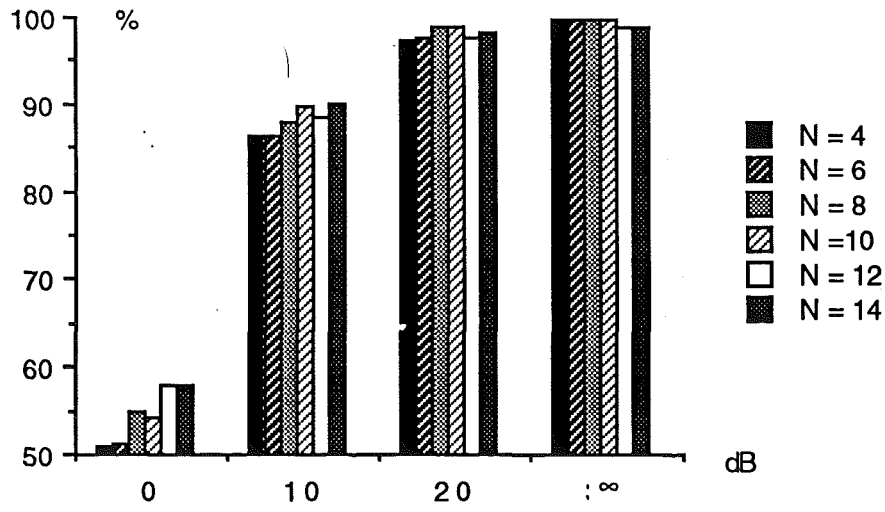

Fig. 6.5. Influencia del número de estados N del modelo de Markov

Se observa que en presencia de ruido las prestaciones aumentan con el número de estados, al menos si el número de estados no es mayor que 14. Seguramente, si se aumenta en exceso el número de estados aparecen problemas de entrenabilidad que hacen descender las prestaciones del sistema. El compromiso coste computacional-tasa de reconocimiento llevó a considerar en este trabajo modelos de 10 estados. Todas las pruebas que siguen a continuación están realizadas con modelos de 10 estados.

En cuanto a las transiciones permitidas en el modelo, se han considerado dos posibilidades: permitiendo sólo transiciones al estado actual y al siguiente (abreviadamente, modelos sin saltos), como en [Wil88], o permitiendo transiciones al estado actual y a los dos siguientes (modelos von saltos) [Mar90]. En la tabla 6.2 se muestran las tasas de reconocimiento en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, sin préenfasis, utilizando un diccionario de 64 palabras-código, ventana cepstral rampa de orden 12 y modelos de izquierda a derecha de 10 estados, sin saltos y con saltos.

Saltos /SNR(dB)	sin ruido	20	10	0
No	99.8	98.9	89.5	54.2
Sí	99.8	98.9	88.1	52.0

Tabla 6.2. Influencia de la estructura del modelo

Como puede observarse la diferencia es muy pequeña entre los dos casos, a favor de los modelos modelos sin saltos. Sin embargo, hay que destacar que este último tipo de modelos tiene menos parámetros para entrenar, concretamente menos probabilidades de transición no nulas.

En vista de los resultados de la figura 6.5 y la tabla 6.2, en las pruebas que siguen a continuación se utilizarán modelos de 10 estados sin saltos, es decir, permitiendo únicamente transiciones al estado actual y al siguiente.

6.3.1.3. TAMAÑO DEL DICCIONARIO DEL CUANTIFICADOR VECTORIAL

Por lo que se refiere al tamaño del diccionario del cuantificador, en la figura 6.6 se muestran las tasas de reconocimiento en tanto por ciento obtenidas en la

aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, sin préenfasis, utilizando modelos de izquierda a derecha de 10 estados sin saltos, ventana cepstral rampa de orden 12 y un diccionario de M palabras-código.

Puede observarse que en condiciones libres de ruido aumentan los errores de reconocimiento si se utilizan un número demasiado bajo de palabras-código. Ello es debido a que se pierde discriminación entre diferentes tipos de sonidos. Sin embargo, en condiciones severas de ruido es conveniente utilizar un diccionario de pocas palabras-código, lo cual puede justificarse si se tiene en cuenta que en este caso se producirán menores errores de cuantificación debidos al ruido cuanto mayor sea la región del espacio de características asignada a cada palabra-código.

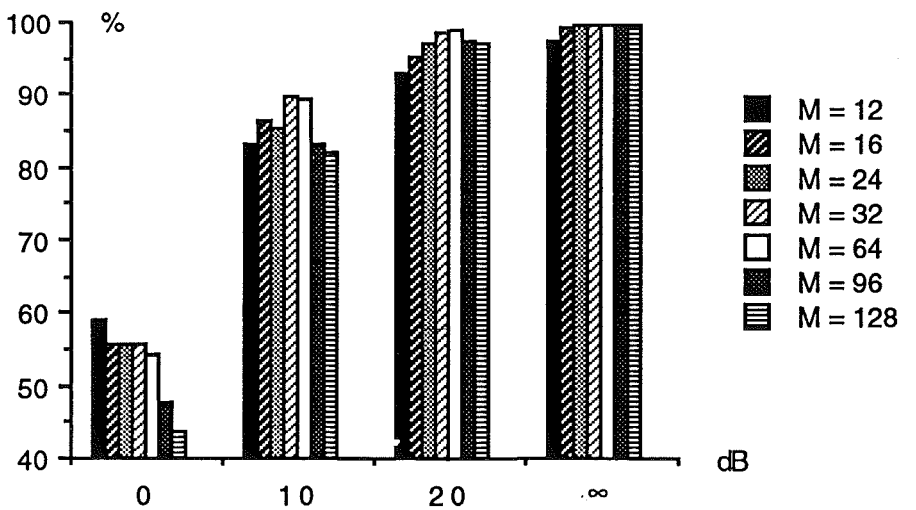

Fig. 6.6. Influencia del número M de palabras-código

Como puede verse, los valores de M para los que se obtienen mejores resultados globalmente, teniendo en cuenta las diversas condiciones de ruido, son 32 y 64. Para no perder discriminación entre diferentes sonidos en pruebas futuras y, siguiendo la experiencia en reconocimiento del habla del grupo de investigación en que se ha realizado esta tesis [Mar90], se decidió tomar 64 como tamaño óptimo del diccionario del cuantificador vectorial, es decir, como número óptimo de palabras-código.

6.3.1.4. ORDEN DE PREDICCION LINEAL Y PONDERACION CEPSTRAL

Debido a que el orden de predicción lineal y la forma y longitud de la ventana de ponderación cepstral están estrechamente relacionados, se optimizaron ambos factores conjuntamente. Para ello, se consideró la experiencia aportada por pruebas preliminares con sistemas de comparación de patrones [Ras90], que indicaba que en presencia de ruido blanco eran preferibles órdenes de predicción relativamente altos y ventanas de ponderación crecientes.

Teniendo en cuenta que el orden de predicción p normalmente usado en reconocimiento del habla es la frecuencia de muestreo expresada en kHz (en el sistema usado es 8) se realizaron pruebas de reconocimiento con órdenes 8, 12 y 16 y las ventanas cepstrales más usadas en reconocimiento del habla: rectangular de p coeficientes, seno realzado, inversa de la desviación típica y rampa.

En la tabla 6.3 se muestran las tasas de reconocimiento en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, sin préenfasis, utilizando modelos de izquierda a derecha de 10 estados sin saltos, diccionario de 64 palabras-código y los órdenes de predicción lineal y ventanas cepstrales indicados.

Orden	Vent. cep. /SNR(dB)	sin ruido	20	10	0
8	Rectangular	99.8	74.2	36.6	22.2
	Seno	99.8	92.8	56.8	27.0
	Inv. desv. típica	99.9	97.7	80.0	37.7
	Rampa	99.7	95.7	72.3	34.1
12	Rectangular	99.8	66.1	34.0	22.8
	Seno	99.7	96.2	73.7	29.0
	Inv. desv. típica	99.7	97.8	84.0	41.8
	Rampa	99.8	98.9	89.5	54.2
16	Rectangular	99.9	73.0	35.5	22.2
	Seno	100	94.0	60.2	19.6
	Inv. desv. típica	99.9	97.7	73.5	32.3
	Rampa	99.8	93.2	70.7	41.2

Tabla 6.3. Influencia del orden de predicción y la ponderación cepstral

En esta tabla puede observarse que en ausencia de ruido ni el orden del modelo ni el tipo de ventana cepstral son importantes en la aplicación que se está considerando. Sin embargo, en presencia de ruido blanco, los resultados resultan ser muy sensibles a ambos factores.

Para los tres órdenes considerados las prestaciones de las ventanas rectangular y seno realzado son muy pobres, mientras que las ventanas crecientes, como la inversa de la desviación típica y la rampa superan ampliamente estos resultados. Concretamente, para orden 8 la ventana rampa es superada ligeramente por la ventana inversa de la desviación típica, para orden 12 la ventana rampa obtiene los mejores resultados y para orden 16 la bondad de las ventanas depende de la SNR considerada.

En cuanto a la variación de las prestaciones de las distintas ventanas cepstrales en función del orden de predicción, se observa que en todas las ventanas los resultados óptimos se obtienen para orden de predicción igual a 12, que es un orden relativamente alto si se tiene en cuenta que el orden usual sería 8. Obsérvese la mejora importante de prestaciones que se observa en la ventana rampa al pasar de orden 8 a orden 12.

La conclusión más importante es que se constatan los resultados obtenidos en los estudios preliminares con sistemas de comparación de patrones: son convenientes órdenes de predicción relativamente altos y ponderaciones cepstrales crecientes.

La conveniencia de este orden relativamente alto es debido al hecho de que los coeficientes de autocorrelación de orden bajo están más contaminados por este tipo de ruido, aproximación al ruido blanco ideal, que los coeficientes de orden alto. Ordenes del modelo demasiado altos, sin embargo, dan lugar de nuevo a resultados pobres debido a la aparición de picos espurios en la estimación espectral. La causa de estos picos espurios es el hecho de que la varianza de los coeficientes de autocorrelación estimados mediante el estimador sesgado clásico aumenta con el índice debido a que al aumentar este disminuye el número de términos que intervienen en la estimación. Hay, pues, un compromiso robustez al ruido-varianza en los coeficientes de la autocorrelación, que en las pruebas consideradas da lugar a que lo mejores resultados se produzcan en torno al orden 12, para una frecuencia de muestreo de 8 kHz.

En cuanto a la conveniencia de ventanas cepstrales crecientes, las razones son análogas. En presencia de ruido blanco los coeficientes cepstrales de orden bajo están

más contaminados que los coeficientes cepstrales de orden alto y, por tanto, es conveniente desenfatizar los coeficientes cepstrales de orden bajo.

Sin tener en cuenta los resultados en ausencia de ruido, la ventana rampa de orden 12 supera ampliamente al resto de las combinaciones orden del modelo-ventana cepstral de la tabla 6.3 para todas las relaciones señal-ruido. Por ello, usando la ventana rampa se hicieron pruebas para un rango más amplio de órdenes de predicción alrededor de 12 (ver figura 6.7). Los mejores resultados se obtienen para p =12. Por tanto, la ventana rampa de orden 12 será la usada en las siguientes pruebas experimentales, a menos que se indique lo contrario.

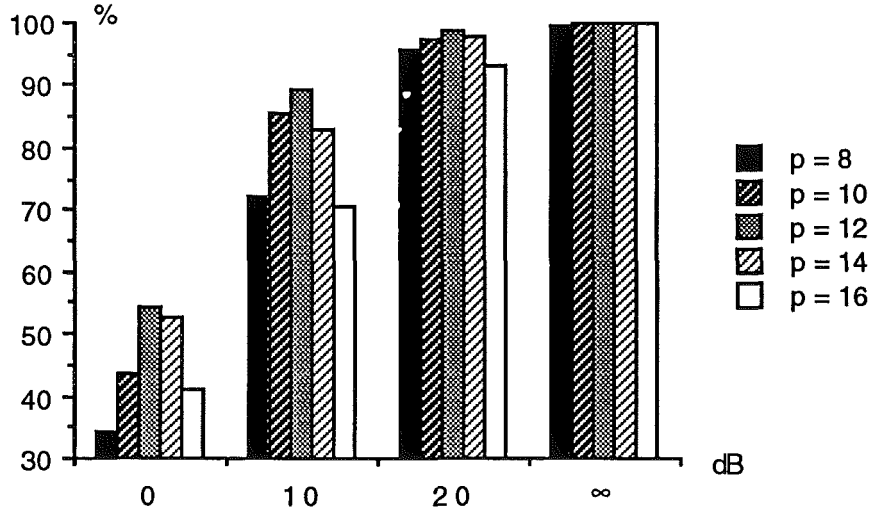

Fig. 6.7. Influencia del orden de predicción con ventana cepstral rampa

En la figura 6.8 se presentan gráficamente las tasas de reconocimiento expresadas en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos, orden de predicción 12 y las cuatro ventanas cepstrales consideradas.

En esta gráfica se observa claramente la necesidad de una ponderación de los coeficientes cepstrales para mejorar los resultados en presencia de ruido y que la ponderación óptima de las cuatro consideradas es la ventana rampa.

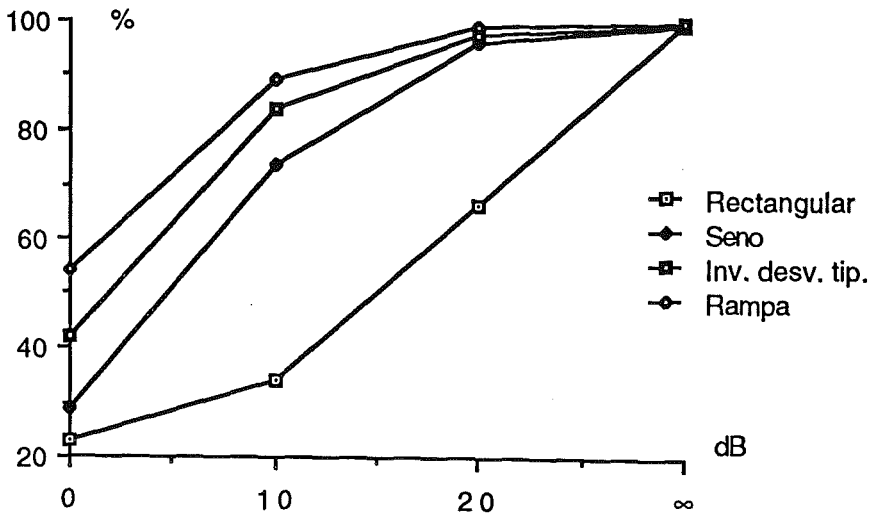

Fig. 6.8. Influencia de la ventana cepstral con orden de predicción 12

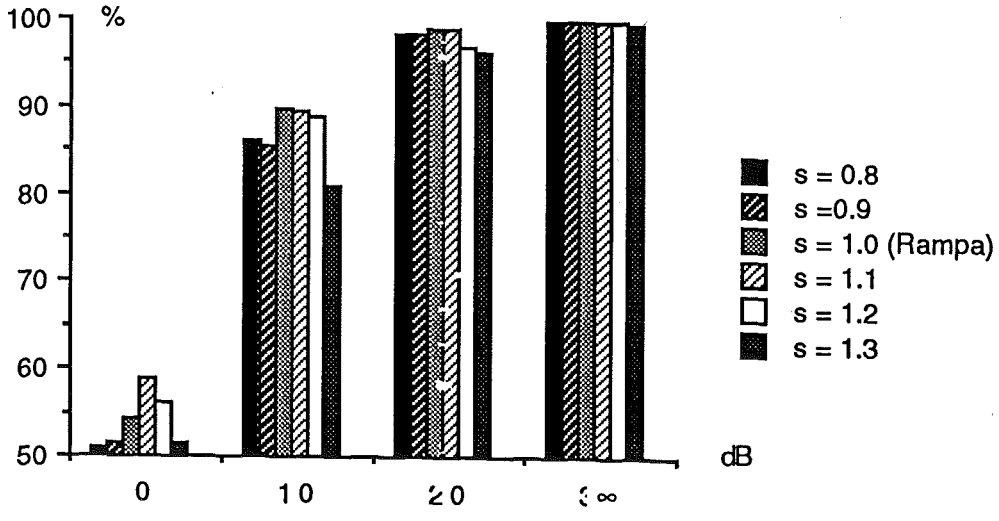

Fig. 6.9. Generalización de la ventana cepstral rampa

Una generalización de la ventana rampa que puede mejorar los resultados de la misma es la siguiente ponderación cepstral

$$w(n) = n^s, \qquad (6.1)$$

donde s es un parámetro de ajuste para optimizar las prestaciones del sistema. Para n=1, se obtiene la ventana rampa, con la cual se han obtenido muy buenas prestaciones.

En la figura 6.9 se muestran las tasas de reconocimiento en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, sin preénfasis, utilizando modelos de izquierda a derecha de 10 estados sin saltos, diccionario de 64 palabras-código, orden de predicción lineal 12 y la ponderación cepstral indicada en (6.1) para varios valores de s en torno a la unidad.

Puede observarse en dicha figura, que en las condiciones más severas de ruido consideradas, 0 dB, se obtienen resultados mejores utilizando (6.1) con s =1.1 que usando rampa. Ello es debido a que cuanto mayor es el nivel de ruido, más contaminados están los primeros coeficientes cepstrales respecto a los siguientes y, por tanto, necesitan un mayor desénfasis, lo cual se obtiene aumentando s por encima de 1.

6.3.2. PARAMETRIZACIONES ALTERNATIVAS

En este apartado se analiza el comportamiento de las parametrizaciones alternativas a la predicción lineal clásica descritas y/o presentadas en el capítulo 3 de esta memoria en reconocimiento de habla ruidosa. En el apartado 6.3.2.1 se muestran las prestaciones de los sistemas HOYWE, OHOYWE y OYWE, descritos en el apartado 3.3 de la memoria. El apartado 6.3.2.2 está dedicado a la parametrización MIAC, propuesta en el apartado 3.5.2. En el apartado 6.3.2.3 se recogen los resultados obtenidos con la parametrización OSALPC, presentada en el apartado 3.5.3, y se relacionan con las prestaciones de la representación SMC, revisada en el apartado 3.5.4. Por último, el apartado 6.3.2.4 está dedicado a la transformación bilineal de frecuencias.

6.3.2.1. HOYWE, OHOYWE Y OYWE

Este apartado está dedicado a comparar las prestaciones de las técnicas de estimación espectral descritas en el apartado 3.3 de la memoria, HOYWE (Ecuaciones de Yule-Walker de Orden Superior), OHOYWE (Ecuaciones de Yule-Walker de Orden Superior Sobredeterminadas) y OYWE (Ecuaciones de Yule-Walker Sobredeterminadas), con la técnica de predicción lineal clásica utilizada en el sistema básico de reconocimiento con que se ha trabajado hasta ahora.

En la figura 6.10 se ha esquematizado la etapa de parametrización del sistema de reconocimiento que se corresponde con estas técnicas de análisis espectral. Hay bastantes diferencias con respecto a la figura 6.2, correspondiente con la técnica de predicción clásica.

Se ha omitido el bloque de preénfasis, pues ya se ha visto anteriormente que este es perjudicial en reconocimiento de habla en presencia de ruido blanco. Para la estimación del modelo autorregresivo de orden p, no se han de calcular las p+1 primeras autocorrelaciones como en el caso de la predicción lineal clásica, sino las necesarias para construir los sistemas de ecuaciones correspondientes a cada parametrización: de m=1 a 2p, en el sistema HOYWE, de m=1 a M, en el sistema OHOYWE, y de m=0 a M, en el sistema OYWE.

En las pruebas experimentales se ha utilizado un valor de M = N/2, siendo N la longitud de la trama de la señal de voz. Se ha escogido este valor porque este tipo de estimadores espectrales no son muy sensibles al valor de M, si se encuentra en un rango bastante amplio alrededor de este valor N/2, y este valor permitirá realizar una comparación más fiable con las técnicas SMC y OSALPC. Como N es igual a 240, 30 ms a 8 kHz de frecuencia de muestreo, M es igual a 120.

Se ha utilizado el mismo estimador sesgado de la autocorrelación que se usa en la predicción lineal clásica. En la técnica clásica, se suele utilizar este estimador debido principalmente a que asegura la estabilidad del filtro del modelo. En el caso de las parametrizaciones consideradas el uso de este estimador no garantiza estabilidad, pero el uso del estimador no sesgado no ha mejorado las prestaciones del sistema.

Con las estimaciones de la autocorrelación obtenidas se construyen los sistemas de ecuaciones respectivos: (3.50) en el caso de HOYWE, (3.54), en el caso de OHOYWE y (3.56), en el caso de OYWE. Los sistemas sobredeterminados han sido resueltos por el método de mínimos cuadrados. La aplicación de técnicas de descomposición en valores

singulares de la matriz de autocorrelación y de mínimos cuadrados totales [Sto92] no han mejorado las prestaciones del sistema.

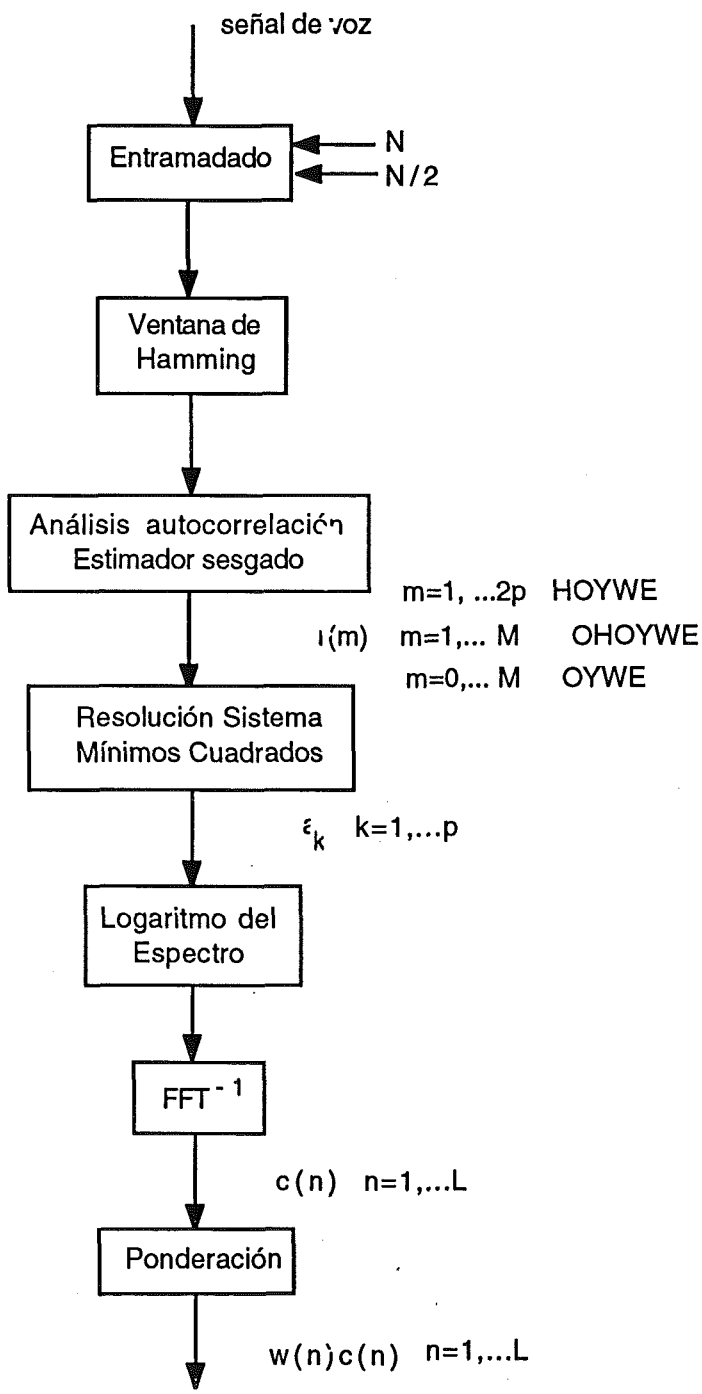

Fig. 6.10. Etapa de parametrización de las técnicas HOYWE, OHOYWE y OYWE

Al no garantizarse estabilidad en estos métodos de estimación espectral, la recursión (4.8)-(4.10), que se utiliza en el sistema básico de reconocimiento para calcular el cepstrum a partir de los coeficientes de predicción, no puede ser utilizada en general. En su lugar, se ha hallado el cepstrum aplicando su definición en su dominio frecuencial mediante FFT's de 512 puntos.

En la tabla 6.4 se comparan las tasas de reconocimiento expresadas en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral de orden 12, con las obtenidas sustituyendo en el mismo sistema de reconocimiento la predicción lineal clásica por las técnicas de análisis espectral HOYWE, OHOYWE y OYWE.

Param. /SNR(dB)	sin ruido	20	10	0
YWE (LPC)	99.8	98.9	89.5	54.2
HOYWE	97.0	88.6	66.9	46.6
OHOYWE	99.5	97.7	81.3	43.1
OYWE	99.9	95.9	66.9	31.7

Tabla 6.4. Comparación de la predicción lineal clásica, YWE (LPC) con las técnicas HOYWE, OHOYWE y OYWE.

Puede que el uso de una ventana rampa de orden 12 no sea la condición óptima para cada técnica de parametrización, pero puede ayudar a comparar sus prestaciones. Además, se ha observado que las técnicas HOYWE, OHOYWE y OYWE son menos sensibles a cambios en el orden del modelo y en la ventana cepstral que la técnica clásica.

Los resultados de la tabla 6.4 pueden observarse de forma gráfica en la figura 6.11. La conclusión fundamental que se extrae de estos resultados es que estas técnicas se comportan peor que la predicción lineal clásica. Es de destacar, no obstante, que la técnica OYWE se comporta de forma excelente en ausencia de ruido [Her92c].

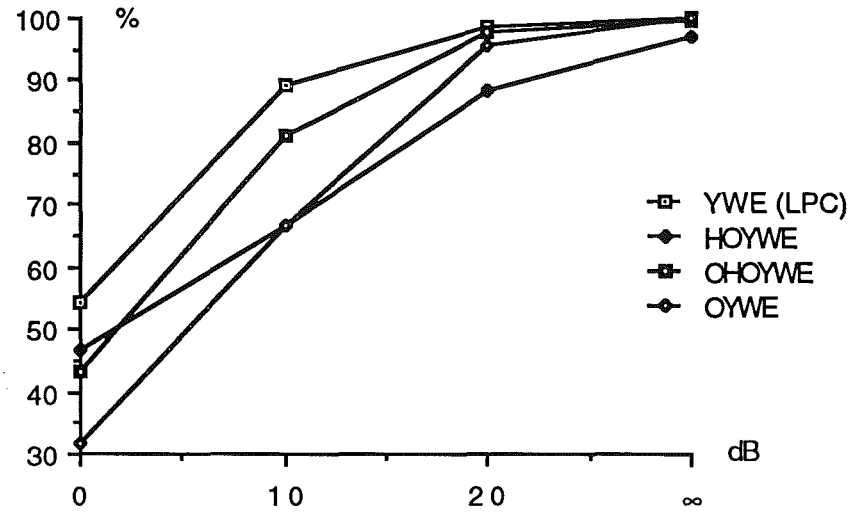

Fig. 6.11. Comparación de la predicción lineal clásica, YWE (LPC) con las técnicas HOYWE, OHOYWE y OYWE.

6.3.2.2. MIAC

Las prestaciones obtenidas en reconocimiento del habla por la técnica de parametrización MIAC (Modelado Inverso de la Autocorrelación Causal), propuesta en el apartado 3.5.2 de esta memoria, son bastante pobres.

Aparecen fuertes oscilaciones en la secuencia cepstral debidas a la existencia de un cero del transformador de Hilbert en la frecuencia π y al bajo contenido a altas frecuencias del espectro de la señal de voz.

Estas prestaciones han conseguido mejorarse multiplicando el valor en el origen de la parte causal de la autocorrelación por un factor K, lo cual equivale a sumar un pedestal de ruido blanco de valor K-1 al espectro. Con ello se aplana el espectro y se suavizan las oscilaciones del cepstrum. Además, el hecho de añadir ruido al espectro mejora el reconocimiento en ambiente ruidoso.

En la figura 6.12 se ha esquematizado la etapa de parametrización del sistema de reconocimiento basado en esta técnica. Hay algunas diferencias con respecto a la figura 6.2, correspondiente con la técnica de predicción clásica.

Se ha omitido el bloque de preénfasis, pues ya se ha visto anteriormente que este es perjudicial en reconocimiento de habla en presencia de ruido blanco. En lugar de utilizar el algoritmo de Levinson-Durbin, se resuelve el sistema triangular (3.86), sustituyendo r(0)/2 por Kr(0)/2, mucho más simple de resolver que dicho algoritmo. Se ha utilizado el mismo estimador sesgado de la autocorrelación que se usa en la predicción lineal clásica y se calculado el cepstrum a partir de su definición en el dominio frecuencial mediante FFT's de 512 puntos por las mismas razones expuestas en el caso de las técnicas del apartado anterior. La imposibilidad de utilizar la recursión (4.8)-(4.10) para calcular los coeficientes cepstrales a partir de los de predicción hace que esta técnica de parametrización no resulte tan eficiente como podía suponerse en un principio.

En la tabla 6.5 se comparan las tasas de reconocimiento expresadas en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral de orden 12, con las obtenidas sustituyendo en el mismo sistema de reconocimiento la predicción lineal clásica por la técnica de parametrización MIAC, para varios valores de K.

Param. /SNR(dB)	sin ruido	20	10	0
LPC	99.8	98.9	89.5	54.2
MIAC (K=1)	95.8	63.3	29.9	14.3
MIAC (K=2)	96.6	95.0	87.6	66.2
MIAC (K=5)	96.3	95.0	88.7	69.7
MIAC (K=10)	96.6	94.1	85.5	58.7
MIAC (K=20)	96.6	94.3	82.9	54.1

Tabla 6.5. Comparación de la predicción lineal clásica LPC con la técnica MIAC

En la tabla de resultados puede observarse que la técnica MIAC sin la ponderación K en el valor en el origen de la parte causal de la autocorrelación da lugar a unas tasas de reconocimiento muy pobres respecto a la predicción lineal clásica. Utilizando el factor de ponderación K las prestaciones de la técnica MIAC mejoran de

forma muy significativa, pero sólo consiguen superar a las de la predicción lineal clásica en las condiciones más severas de ruido consideradas para algunos valores de K.

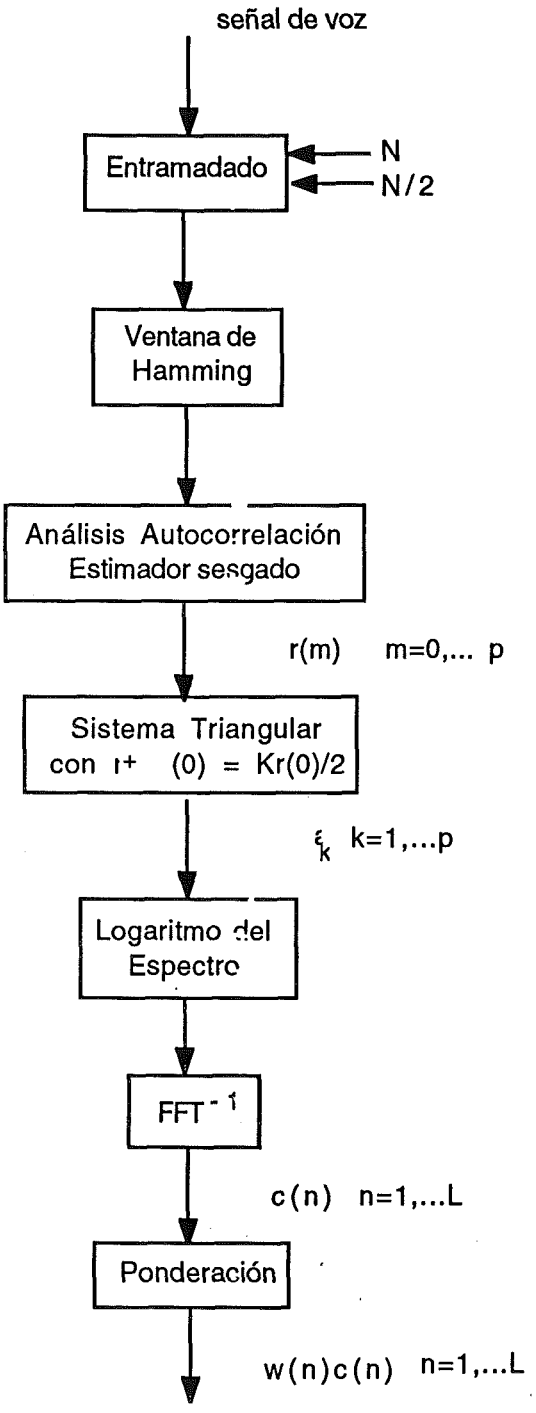

Fig. 6.12. Etapa de parametrización de la técnica MIAC

6.3.2.3. OSALPC Y SMC

Antes de analizar las prestaciones de las técnicas OSALPC y SMC, se va a estudiar el comportamiento en reconocimiento de habla ruidosa del sistema de ecuaciones sobredeterminado (3.87) presentado en el apartado 3.5.3 de esta memoria como enlace entre las técnicas MIAC y OSALPC.

Al igual que en las técnicas HOYWE, OHOYWE y OYWE, y por las mismas razones, no se aplicó preénfasis a la señal de voz, se utilizó el estimador segado de la autocorrelación, se tomo M igual a 120, se resolvió el sistema por el método de mínimos cuadrados y se calculó el cepstrum a partir de su definición en el dominio frecuencial con FFT´s de 512 puntos.

En la figura 6.13 se representa la etapa de parametrización de un sistema de reconocimiento basado en la resolución del sistema (3.87). Teniendo en cuenta que la resolución de este sistema consiste en un modelado autorregresivo en el dominio de la autocorrelación, se consideró que podía ser conveniente un enventanado de Hamming de la parte causal de la secuencia de la autocorrelación antes de resolver el sistema.

Para hacer un estudio completo, se consideraron cuatro posibles combinaciones de enventanado: la combinación (1,1) consiste en enventanar con Hamming la señal de voz y la parte causal de la autocorrelación, la (1,2) consiste en enventanar con Hamming la señal de voz y con ventana rectangular la parte causal de la autocorrelación y así sucesivamente.

En la tabla 6.6 se comparan las tasas de reconocimiento expresadas en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral de orden 12, con las obtenidas sustituyendo en el mismo sistema de reconocimiento la predicción lineal clásica por la parametrización correspondiente a la figura 6.13. Se ha seguido el código de enventanado descrito en el párrafo anterior.

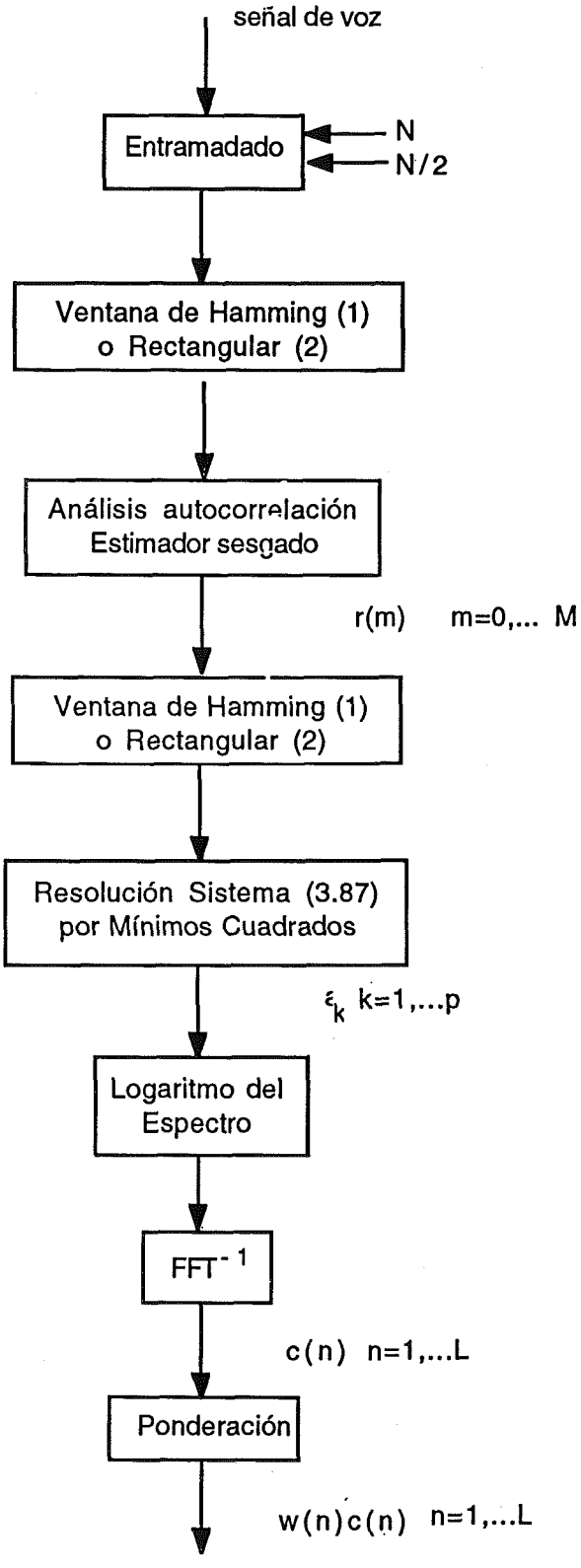

Fig. 6.13. Etapa de parametrización usando el sistema (3.87)

Param. /SNR(dB)	sin ruido	20	10	0
LPC	99.8	98.9	89.5	54.2
Sist. (3.87) (1,2)	97.6	95.2	90.4	74.7
Sist. (3.87) (2,2)	96.9	95.5	90.3	76.4
Sist. (3.87) (1,1)	98.7	98.3	93.3	75.2
Sist. (3.87) (2,1)	99.0	98.1	93.6	75.5

Tabla 6.6. Comparación de la predicción lineal clásica con la parametrización de la Fig. 6.13

Como puede observarse, se obtienen resultados excelentes para todas las combinaciones de enventanado en condiciones severas de ruido, 10 y 0 dB, comparando con la predicción lineal clásica LPC.

El comportamiento a SNR altas es, sin embargo, algo peor que en el caso de la predicción lineal clásica, lo cual es explicable dado que el sistema (3.87) supone un modelo todo-polos de la parte causal de la autocorrelación de la señal de voz y, por tanto, no es consistente con el modelo lineal de producción de voz, según el cual a la señal de voz le corresponde un modelo todo-polos y a la parte causal de su autocorrelación un modelo de ceros y polos, con igual número de ceros y polos.

Considerando los resultados a SNR altas se observa que es conveniente enventanar la parte causal de la autocorrelación con una ventana de Hamming. Los resultados de las dos últimas combinaciones de la tabla son mucho mejores que los de las dos anteriores a ∞ y 20 dB de SNR. En lo sucesivo se considerará enventanado (2,1), ventana rectangular para la señal y ventana de Hamming para la parte causal de la autocorrelación, pues da resultados ligeramente superiores a la combinación (1,1).

La ponderación del valor en el origen de la parte causal de la autocorrelación multiplicando por un factor K no ha dado lugar a una mejora de resultados, por lo cual no se considerará esta técnica en las siguientes pruebas.

Sustituyendo en la parametrización de la figura 6.13 el sistema de ecuaciones (3.87) por el sistema (3.89), ver apartado 3.5.3 de la memoria, se obtiene una implementación de la técnica OSALPC (*One-Sided Autocorrelation Linear Predictive Coding*, Predicción Lineal de la Parte Causal de la Autocorrelación). En este caso, se

obtienen las siguientes tasas de reconocimiento: 98.9, 98.0, 93.0 y 75.6 para ∞, 20, 10 y 0 dB de SNR, respectivamente.

Como ya se ha descrito en el apartado 3.5.3 es posible implementar esta parametrización haciendo uso del algoritmo de Levinson-Durbin, con lo cual se garantiza la estabilidad del filtro del modelo y se reduce considerablemente el coste computacional.

Esta implementación de la técnica de parametrización OSALPC se observa en la figura 6.14. Se ha omitido el primer bloque de enventanado, al tratarse de una ventana rectangular. Por otro lado, como el algoritmo de Levinson-Durbin garantiza estabilidad, se pueden calcular los coeficientes cepstrales a partir de los coeficientes de predicción haciendo uso de la recursión (4.8)-(4.10).

En la tabla 6.7 se comparan las tasas de reconocimiento expresadas en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral de orden 12, con las obtenidas sustituyendo en el mismo sistema de reconocimiento la predicción lineal clásica por la parametrización OSALPC correspondiente a la figura 6.14. Puede observarse que las tasas de reconocimiento proporcionadas por la técnica OSALPC difieren algo de las mencionadas tres párrafos antes. Estas diferencias son debidas a la distinta implementación de la técnica.

Param. /SNR(dB)	sin ruido	20	10	0
LPC	99.8	98.9	89.5	54.2
OSALPC	98.6	97.7	94.9	79.0

Tabla 6.7. Comparación de la predicción lineal clásica LPC con la técnica OSALPC

En la tabla 6.7 se observa claramente el excelente comportamiento de la técnica OSALPC en condiciones severas de ruido, 10 y 0 dB de SNR. Por las razones ya comentadas, se produce un empeoramiento de la tasa de reconocimiento cuando hay poco ruido.

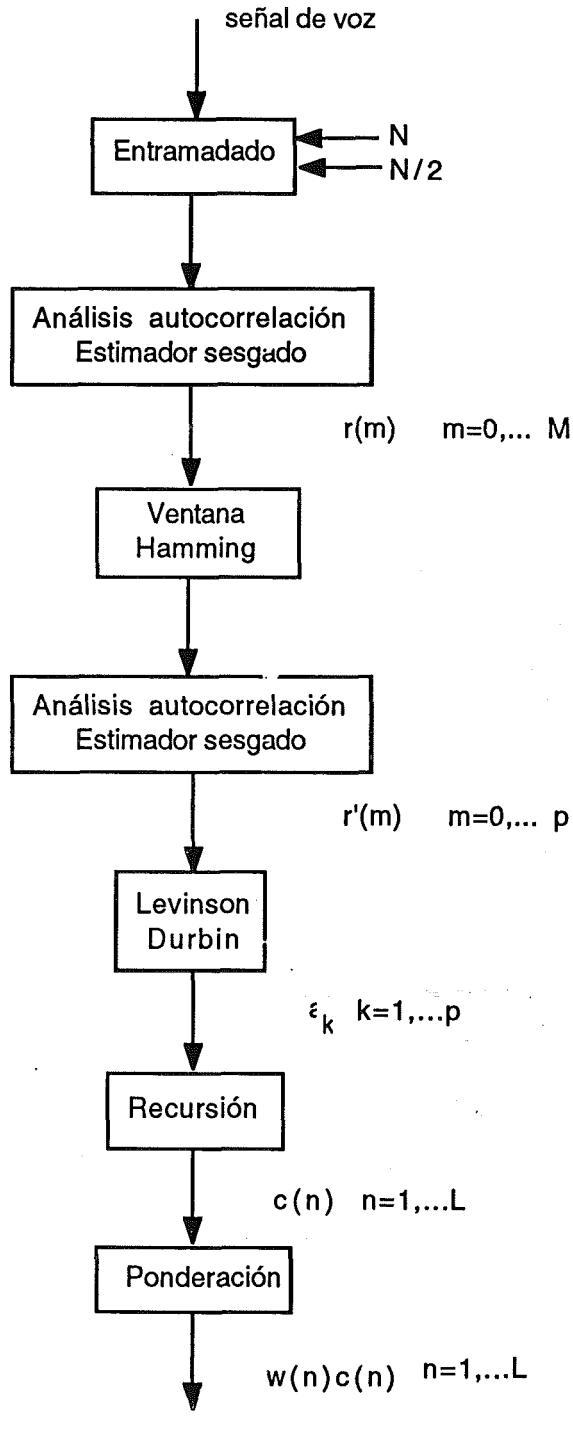

Fig. 6.14. Etapa de parametrización correspondiente a la técnica OSALPC

Por otro lado, se vio en el apartado 3.5.4 de la memoria la estrecha relación existente entre la técnica propuesta en esta memoria, la representación OSALPC, y la técnica propuesta recientemente por Mansour y Juang, la representación SMC. Además, se comprobaron las buenas propiedades del estimador de la autocorrelación usado por la técnica SMC, la coherencia.

Seguidamente, se van a comparar las prestaciones de ambas técnicas y también se estudiará la posibilidad de utilizar el estimador coherencia en la técnica OSALPC, que en este caso se denotará como OSALPC2. En esta nueva variante OSALPC2 también se anulará el valor de la autocorrelación en el origen, como se hace en la técnica SMC.

En la tabla 6.8 se comparan las tasas de reconocimiento obtenidas con el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral de orden 12, con las obtenidas sustituyendo en el mismo sistema la predicción lineal clásica por las técnicas OSALPC, SMC y OSALPC2, implementadas tal como se muestra en la Fig. 6.15.

Param. /SNR(dB)	sin ruido	20	10	0
LPC(YWE)	99.8	98.9	89.5	54.2
OSALPC	98.6	97.7	94.9	79.0
SMC	99.0	97.0	89.2	67.5
0SALPC2	99.4	98.4	94.7	72.2

Tabla 6.8. Comparación de la predicción lineal clásica con las técnicas OSALPC y SMC

En la figura 6.16 están representados gráficamente los resultados de la tabla 6.8, en la que se comparan las prestaciones de las técnicas LPC, SMC, OSALPC y OSALPC2.

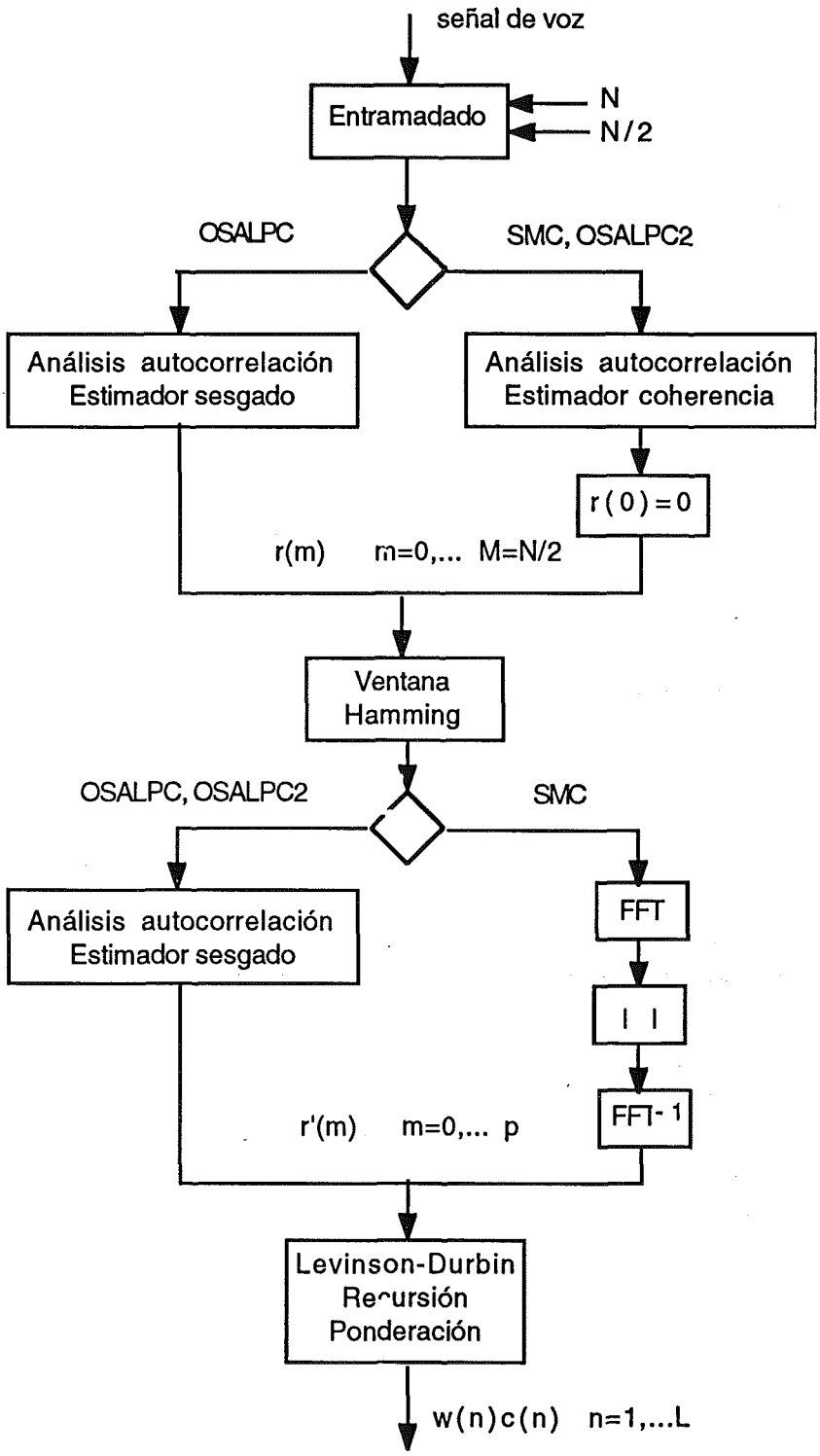

Fig. 6.15. Etapa de parametrización de las técnicas OSALPC, SMC y OSALPC2

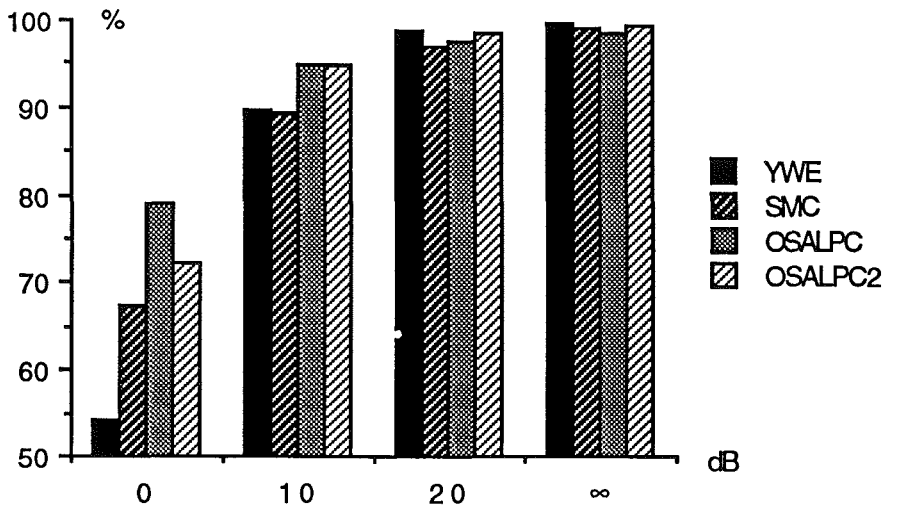

Fig. 6.16. Comparación de la predicción lineal clásica con las técnicas OSALPC y SMC

En la figura se observa claramente que las técnicas OSALPC, OSALPC2 y SMC mejoran ostensiblemente las tasas de reconocimiento obtenidas por la predicción lineal clásica (YWE o LPC) en las condiciones de ruido más severas consideradas, 0 dB de SNR. De estas tres, destaca por sus prestaciones en estas condiciones de ruido la técnica OSALPC. A 10 dB de SNR las técnicas LPC y SMC son comparables y son superadas ampliamente por las técnicas OSALPC y OSALPC2. A 20 dB de SNR la mejor técnica es la clásica LPC, pero la técnica OSALPC2 obtiene resultados bastante similares. Por último, en condiciones libres de ruido también la técnica clásica es la que obtiene mejores resultados y la técnica OSALPC2 es la que más se acerca a sus prestaciones.

Puede concluirse a partir de estos resultados que en condiciones libres de ruido la técnica clásica no es mejorada por las nuevas técnicas por la razones ya comentadas, pero la técnica OSALPC2 obtiene resultados muy aceptables. Considerando globalmente las cuatro condiciones de ruido, la técnica OSALPC2 parece ser un buen compromiso ente robustez y buenas prestaciones en ausencia de ruido y mejora las prestaciones de la representación SMC.

Hasta ahora sólo se han mostrado resultados de las técnicas basadas en la predicción lineal en el dominio de la autocorrelación utilizando ventana cepstral rampa de orden 12. En las tablas 6.9 y 6.10 se muestran las tasas de reconocimiento obtenidas con el sistema básico de reconocimiento, sin preénfasis, utilizando un

diccionario de 64 palabras-código, modelos de 10 estados sin saltos y varios ordenes de predicción y ventanas cepstrales, sustituyendo en el mismo sistema la predicción lineal clásica por las técnicas OSALPC2 y SMC, respectivamente.

Orden	Vent. cep. /SNR(dB)	sin ruido	20	10	0
8	Seno	97.3	95.5	82.6	44.2
	Inv. desv. típica	97.0	96.4	86.4	52.5
	Rampa	97.6	97.0	92.5	76.0
12	Seno	98.8	97.2	94.1	71.1
	Inv. desv. típica	98.8	98.3	93.3	68.4
	Rampa	99.4	98.4	94.7	72.2
16	Seno	99.3	98.7	94.4	76.8
	Inv. desv. típica	99.1	98.1	92.4	72.7
	Rampa	99.1	98.1	90.7	68.3

Tabla 6.9. Resultados de la técnica OSALPC2 para varios órdenes y ventanas cepstrales

Orden	Vent. cep. /SNR(dB)	sin ruido	20	10	0
8	Seno	98.3	97.1	90.0	63.6
	Inv. desv. típica	98.5	96.1	84.2	56.6
	Rampa	98.6	96.6	87.3	60.1
12	Seno	98.3	97.3	90.7	70.1
	Inv. desv. típica	98.6	96.3	84.8	58.2
	Rampa	99.0	97.0	89.2	67.5
16	Seno	98.2	97.1	89.6	65.4
	Inv. desv. típica	98.2	95.8	84.5	57.4
	Rampa	99.1	97.2	85.9	67.2

Tabla 6.10. Resultados de la técnica SMC para varios órdenes y ventanas cepstrales

A la vista de ambas tablas se concluye que los mejores resultados se obtienen usando la técnica OSALPC2 y ventana cepstral rampa de orden 12 o ventana seno de orden 16. Los resultados de ambas ventanas son similares, superando ligeramente la ventana seno de orden 16 a la ventana rampa de orden 12 en algunas condiciones. Sin embargo, teniendo en cuenta que la ventana rampa de orden 12 tiene longitud 12 y la

ventana seno de orden 16 tiene longitud 24, por razones obvias de cálculo y memoria se puede considerar que la ventana rampa de orden 12 es la óptima para esta técnica.

Además, se observa que la técnica OSALPC2 supera ampliamente la técnica SMC [Her92e], salvo en algunos resultados correspondientes a orden 8 que se alejan claramente de las mejores prestaciones obtenidas por la técnica OSALPC2.

Por último, también destaca el hecho de que la técnica OSALPC2 es menos sensible a cambios en el orden de predicción y en el tipo de ventana cepstral que la técnica clásica de predicción lineal, al menos para órdenes altos (ver tabla 6.3).

Otra variable que se había fijado hasta ahora era el valor de M, que se había tomado igual a N/2 para poder comparar en igualdad de condiciones las técnicas SMC y OSALPC. En la figura 6.17 se muestran los resultados que se obtienen al variar M utilizando la técnica OSALPC2, ventana cepstral rampa de orden 12 y el resto de variables idénticas a las pruebas de la tabla 6.9.

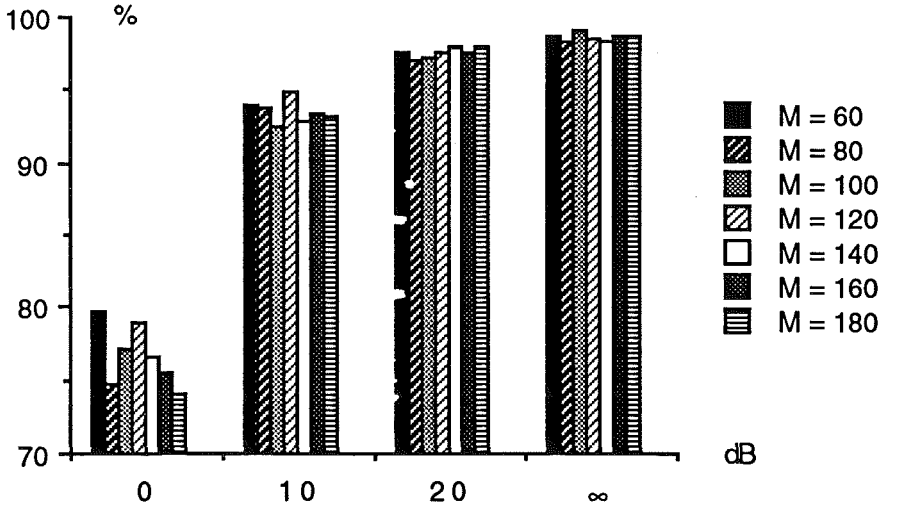

Fig. 6.17. Influencia de la longitud M de la secuencia de autocorrelación en la técnica OSALPC2

En la figura se observa que esta técnica no es muy sensible a las variaciones de M en torno al valor 120 y que es bastante razonable este valor para pruebas futuras considerando globalmente todas las condiciones estudiadas. También se observa que en

las condiciones más severas de ruido consideradas aparece un resultado bastante bueno para un valor muy bajo de M.

Se ha visto que el uso de técnicas basadas en la predicción lineal en el dominio de la autocorrelación, en lugar de sobre la señal misma, proporciona buenos resultados en condiciones ruidosas. Se puede aplicar esta idea iterativamente proponiendo la técnica OSA^{2}LPC2, que consistiría en aplicar la técnica OSALPC2 sobre la parte causal de la secuencia de autocorrelación en lugar de sobre la señal, la técnica OSA^{3}LPC2 y así sucesivamente. En cada autocorrelación se reduciría la longitud de la secuencia a la mitad y se aplicaría la ventana de Hamming sobre la secuencia objeto de predicción.

En la tabla 6.11 se comparan las tasas de reconocimiento obtenidas con el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral rampa de orden 12, sustituyendo en el mismo sistema la predicción lineal clásica por las técnicas OSALPC2, OSA^{2}LPC2 y OSA^{3}LPC2.

Param. /SNR(dB)	sin ruido	20	10	0
LPC (1)	99.8	98.9	89.5	54.2
OSALPC2 (2)	99.4	98.4	94.7	72.2
OSA^{2}LPC2 (3)	95.3	94.3	89.7	75.5
OSA^{3}LPC2 (4)	93.4	93.3	87.2	71.6

Tabla 6.11.Comparación de la técnica LPC con las técnicas OSALPC2, OSA^{2}LPC2 y OSA^{3}LPC2

Estos resultados pueden verse gráficamente en la figura 6.18. En esta figura la técnica LPC corresponde a la etiqueta (1), la OSALPC2 a la (2), la OSA^{2}LPC2 a la (3) y la OSA^{3}LPC2 a la (4). Puede observarse que en condiciones libres de ruido el comportamiento del sistema se degrada al realizar autocorrelaciones sucesivas. A 20 dB de SNR las prestaciones se mantienen bastante si sólo se aplica una vez el concepto de predicción lineal de la parte causal de la autocorrelación, que es el caso de la técnica OSALPC2, ampliamente comentada. A 10 dB las mejores prestaciones corresponden a la técnica OSALPC2. Finalmente, a 0 dB las mejores resultado son los de la técnica OSA^{2}LPC2. Por tanto, estas nuevas técnicas sólo serían aconsejables en condiciones severísimas de ruido.

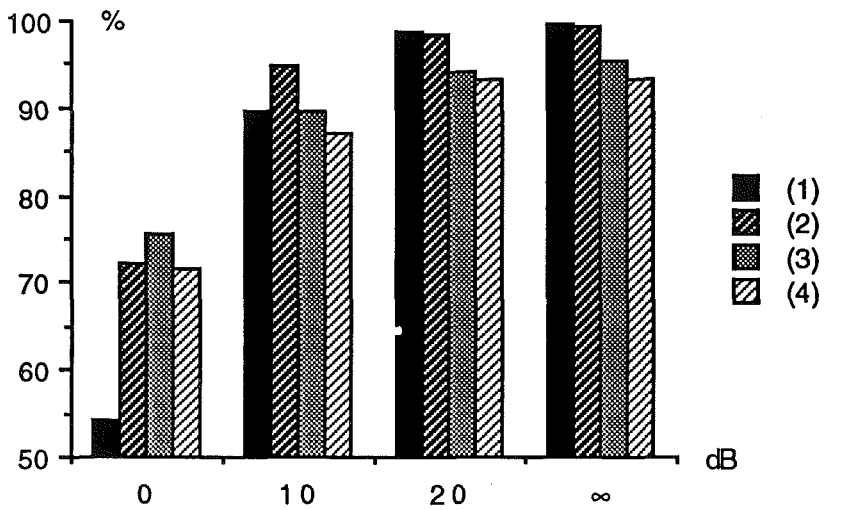

Fig. 6.18. Comparación de la técnica LPC con las técnicas OSALPC2, OSA^{2}LPC2 y OSA^{3}LPC2

Una manera de aprovechar los resultados anteriores sería reconocer con la suma de los espectros de estas técnicas. En la tabla 6.12 se comparan las tasas de reconocimiento obtenidas con el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral rampa de orden 12, sustituyendo en el mismo sistema la predicción lineal clásica (1) por combinaciones de las técnicas anteriores. Puede observarse en esta tabla que se pueden conseguir buenos compromisos entre las prestaciones en diversas condiciones de ruido aplicando esta idea.

Param. /SNR(dB)	sin ruido	20	10	0
(1)	99.8	98.9	89.5	54.2
(2)	99.4	98.4	94.7	72.2
(1)+(2)	99.4	98.9	92.5	73.1
(1)+(2)+(3)	99.5	98.1	93.6	76.0
(1)+(2)+(3)+(4)	99.7	97.9	93.4	77.0

Tabla 6.12. Comparación de la predicción lineal clásica con suma de espectros de otras técnicas basadas en el modelado AR de autocorrelaciones sucesivas

6.3.2.4. TRANSFORMACION BILINEAL

En este apartado se va analizar el comportamiento de la transformación bilineal de los coeficientes cepstrum LPC en reconocimiento de habla ruidosa. El objetivo de la transformación bilineal es emular la sensibilidad logarítmica en frecuencia del oído humano, para lo cual expande la zona de bajas frecuencias y comprime la de altas frecuencias. En presencia de ruido de espectro plano, es de esperar que esta transformación robustezca el vector de parámetros, ya que en la señal de voz la energía se concentra en las bajas frecuencias y, por tanto, estas son más robustas a este tipo de ruido.

En la tabla 6.13 se muestran las tasas de reconocimiento obtenidas con el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos, orden de predicción 12 y diferentes ventanas cepstrales, añadiendo al final de la etapa de parametrización del sistema básico la transformación bilineal con el parámetro α = 0.4. La matriz de transformación se truncó a una longitud adecuada para conseguir la precisión necesaria.

Vent. ceps. /SNR(dB)	sin ruido	20	10	0
Rectangular	99.9	96.8	78.6	38.4
Seno	100	97.5	74.8	24.6
Inv. desv. típ.	99.7	98.3	84.8	42.5
Rampa	99.9	69.1	34.3	20.4

Tabla 6.13. Comportamiento de la transformación bilineal con orden de predicción 12

Vent. ceps. /SNR(dB)	sin ruido	20	10	0
Rectangular	99.8	66.1	34.0	22.8
Seno	99.7	96.2	73.7	29.0
Inv. desv. típ.	99.7	97.8	84.0	41.8
Rampa	99.8	98.9	89.5	54.2

Tabla 6.14. Resultados sin transformación bilineal y orden de predicción 12

En la tabla 6.14 se muestran los tasas de reconocimiento correspondientes a las mismas pruebas de la tabla 6.13, pero sin realizar transformación bilineal (es decir, $\alpha = 0$), extraídas de la tabla 6.3.

A la vista de estos resultados, puede observarse que en ausencia de ponderación cepstral (ventana rectangular), la transformación bilineal robustece al cepstrum frente al ruido blanco aditivo. Ello es debido a que la transformación bilineal expande la zona de bajas frecuencias que es donde la señal de voz tiene más energía y, por lo tanto, es más robusta a este tipo de ruido. Sin embargo, cuando se utilizan las ponderaciones cepstrales usuales -seno e inversa de la desviación típica- la transformación bilineal no parece ayudar al reconocimiento de habla ruidosa. Por último, en el caso de la ventana rampa los resultados empeoran con la transformación bilineal. Ello puede ser debido a que la ventana rampa ha sido diseñada para el reconocimiento de habla ruidosa teniendo en cuenta la sensibilidad al ruido relativa de los diferentes coeficientes cepstrales del modelo LPC en ausencia de transformación bilineal [Her92b].

A este respecto, se ha publicado [O'Sh87] que la transformación mel, similar a la transformación bilineal, mejora el comportamiento de los sistemas de reconocimiento basados en el el cepstrum FFT, pero no el de los basados en el cepstrum LPC.

6.3.3. DISTANCIAS ALTERNATIVAS

En este apartado se analizarán las prestaciones de la distancia de Mahalanobis (apartado 6.3.3.1), las distancias de proyección (6.3.3.2) y el uso de energía y parámetros dinámicos (6.3.3.3), aspectos que se han revisado en el capítulo 4 de esta memoria.

6.3.3.1. DISTANCIA DE MAHALANOBIS

En la tabla 6.15 se comparan las tasas de reconocimiento obtenidas con el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos, orden de predicción 12 y ventanas cepstrales rampa e inversa de la desviación típica, con las obtenidas sustituyendo en el mismo sistema de reconocimiento las distancia euclídea por la distancia de Mahalanobis sobre vectores cepstrales sin ponderar (ventana rectangular). La distancia de

Mahalanobis se ha implementado de forma eficiente, tal como se explicó en el apartado 4.2 de la memoria.

Distancias /SNR(dB)	sin ruido	20	10	0
Mahalanobis	99.7	66.5	26.2	20.1
Inv. desv. típ.	99.7	97.8	84.0	41.8
Rampa	99.8	98.9	89.5	54.2

Tabla 6.15. Comparación de la distancia de Mahalanobis con la distancia euclídea ponderada

Como puede observarse en la tabla, las prestaciones de la distancia de Mahalanobis en condiciones ruidosas están muy por debajo de las distancias euclídeas ponderadas con la ventana inversa de la desviación típica y la ventana rampa. Entre otros factores, ello puede ser debido a una mala estimación de los elementos de fuera de la diagonal de la matriz de covarianza.

6.3.3.2. DISTANCIAS DE PROYECCION

En este apartado se mostrarán las prestaciones de la familia de distancias de proyección cepstral propuestas por D. Mansour y B.H. Juang [Man89b] y revisadas en el apartado 4.4 de esta memoria.

Estas distancias fueron propuestas por los citados autores para su utilización en un sistema de reconocimiento basado en comparación de patrones y, por tanto, fueron usadas como distancia entre vectores para el cálculo de la distancia acumulada en el algoritmo de programación dinámica. Las distancias d_{P3}, d_{P4} y d_{P5} se diferenciaban en diferentes ponderaciones de la norma del vector de test ruidoso, lo cual tiene sentido en el algoritmo de programación dinámica porque permite ponderar aquellas tramas de test de normas altas y, por tanto, más robustas al ruido.

Sin embargo, el sistema de reconocimiento usado en las pruebas experimentales de este apartado está basado en los modelos ocultos de Markov discretos. En este caso, la distancia entre vectores cepstrales se utiliza en la construcción del diccionario del cuantificador vectorial y en la cuantificación de los vectores cepstrales. En tales casos, la ponderación anterior no tiene ningún efecto en las prestaciones del sistema. Por ello,

en las pruebas experimentales de este trabajo, de estas tres distancias sólo se ha utilizado la distancia d_{P5}, por ser la más eficiente desde el punto de vista computacional.

Por otro lado, la distancia d_{P1} tampoco se ha considerado en estas pruebas experimentales, debido a que es necesaria una estimación previa del nivel de ruido para encontrar el valor óptimo de λ.

Por tanto, solamente se compararán las prestaciones de las distancias d_{P2} y d_{P5} con respecto a la distancia euclídea.

Para la utilización de estas distancias en la construcción del diccionario del cuantificador vectorial se ha tenido que afrontar el problema del cálculo del centroide, el vector que minimiza la distorsión medida de una distribución de vectores, que en el caso de la distancia euclídea es trivial, ya que coincide con la media aritmética de la distribución. Para el caso de la distancia d_{P1}, en [Jua92] se demuestra que el centroide coincide con el vector propio asociado al valor propio mayor de la matriz de covarianza de dicha distribución. Sin embargo, este procedimiento resulta prohibitivo desde el punto de vista computacional. En cuanto a la distancia d_{P5}, no se ha encontrado una fórmula cerrada para el cálculo del centroide.

Al no poder usar una fórmula cerrada para el cálculo del centroide, se hicieron pruebas utilizando como centroide el vector de la distribución que minimizaba la distancia a los demás. Sin embargo, se han conseguido resultados similares, e incluso ligeramente mejores, utilizado la media aritmética de la distribución. Debido a que esta última opción es mucho más eficiente desde el punto de vista computacional que la anterior se ha tomado como centroide la media aritmética de la distribución, como en el caso de distancia euclídea.

Antes de presentar los resultados obtenidos, es importante hacer notar que la simplificación (4.44) de la distancia d_{P5} sólo es válida en el proceso de cuantificación. Cuando se construye el diccionario del cuantificador vectorial, se ha de usar la expresión (4.43) si se quiere tomar como criterio de convergencia la disminución relativa de la distorsión global.

En la tabla 6.16 se comparan las tasas de reconocimiento obtenidas con el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos, ventana cepstral rampa de orden

12, con las obtenidas sustituyendo en el mismo sistema de reconocimiento las distancia euclídea por la distancias de proyección d_{P2} y d_{P5}.

Distancias /SNR(dB)	sin ruidc	20	10	0
euclídea	99.8	98.9	89.5	54.2
d_{P2}	99.8	97.9	87.4	62.0
d_{P5}	99.8	98.6	89.2	65.6

Tabla 6.16. Comparación de la distancia euclídea y las de proyección

Puede observarse en esta tabla que la distancia de proyección d_{P5} obtiene prestaciones mejores que la distancia d_{P1}. Ello es debido a que la distancia d_{P1} tiene en cuenta la disminución de la norma cepstral debida al ruido, mientras que la distancia d_{P5} tiene en cuenta la robustez del ángulo frente al ruido, efecto que predomina sobre el anterior. Por otro lado, se observa que la distancia de proyección d_{P5} supera ampliamente a la distancia euclídea en las condiciones más severas de ruido, mientras que en el resto de condiciones el comportamiento de las dos distancias es muy similar.

En la tabla 6.17 se muestran las tasas de reconocimiento obtenidas con el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y varios órdenes de predicción y ventanas cepstrales sustituyendo en el mismo sistema de reconocimiento las distancia euclídea por la distancia de proyección d_{P5}.

Orden	Vent. cep. /SNR(dB)	sin ruido	20	10	0
8	Seno	99.9	97.1	83.5	61.6
	Inv. desv. típica	99.9	97.0	82.7	56.6
	Rampa	99.8	98.1	89.5	66.1
12	Seno	99.8	96.8	83.2	61.0
	Inv. desv. típica	99.9	98.4	88.1	57.3
	Rampa	99.8	98.6	89.2	65.6
16	Seno	99.8	96.6	83.0	60.7
	Inv. desv. típica	99.8	98.5	87.3	57.6
	Rampa	99.8	98.7	89.0	65.3

Tabla 6.17. Resultados de la distancia d_{P5} para varios órdenes y ventanas cepstrales

Puede observarse que la utilización de la distancia de proyección d_{P5} robustece el sistema de reconocimiento frente a cambios en el orden de predicción y la ventana cepstral. Los mejores resultados se obtienen para las ventanas cepstrales rampa de orden 12 y 16, siendo los resultados muy similares en ambos casos.

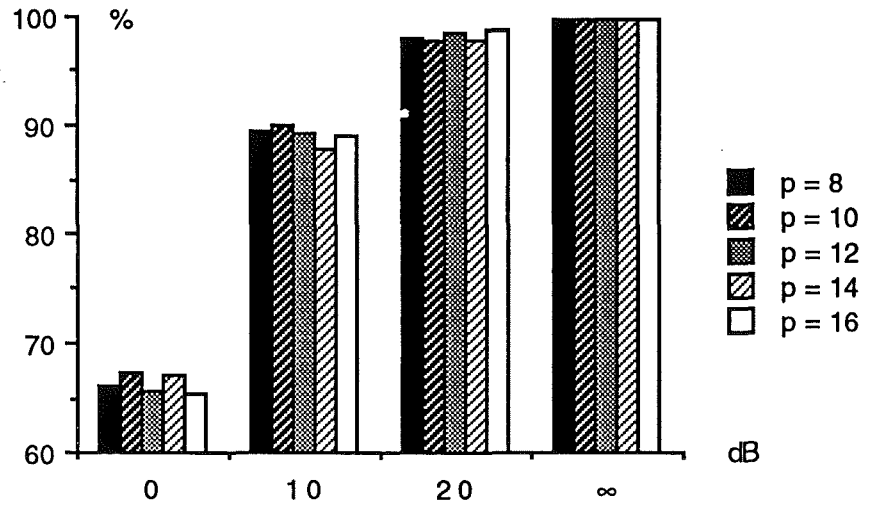

Fig. 6.19. Resultados con distancia d_{P5}, ventana rampa y varios órdenes

En la figura 6.19 están representados gráficamente los resultados obtenidos con distancia de proyección d_{P5}, ventana cepstral rampa y un amplio rango de valores de orden de predicción. El orden de predicción 12 puede considerarse un valor de compromiso entre las prestaciones a altas y bajas SNR.

En la figura 6.20 se comparan gráficamente las prestaciones de las distancias euclídea y de proyección d_{P5} para ventana cepstral rampa de orden 12. Puede observarse que el comportamiento de ambas distancias es muy similar a SNR altas y medias y la distancia d_{P5} supera a la de proyección en condiciones severas de ruido.

Por último, es interesante destacar que mediante el uso conjunto de la distancia de proyección con la ventana cepstral de orden 12 se han mejorado las resultados obtenidos por los autores que propusieron las distancias de proyección, que usaban una ventana cepstral seno desplazado de orden 8 (4.35)-(4.36).

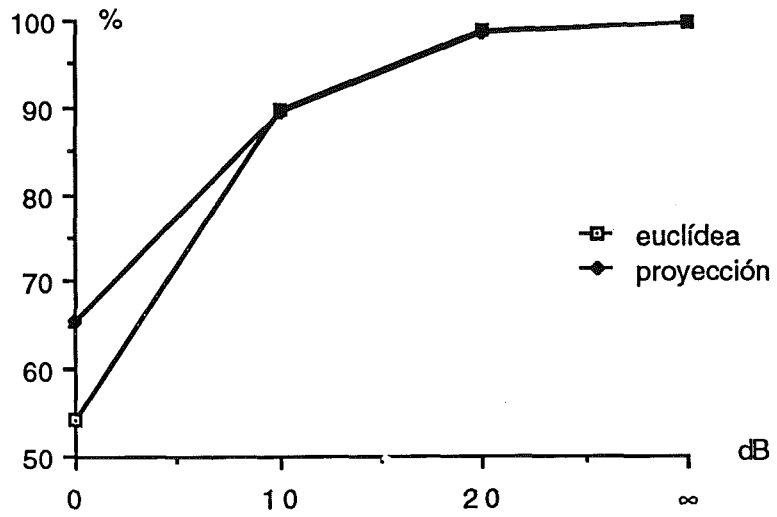

Fig. 6.20. Comparación de las distancias euclídea y de proyección d_{P5}

6.3.4. INCORPORACION DE ENERGIA Y PARAMETROS DINAMICOS

En este apartado se mostrarán los resultados obtenidos incorporando información de energía y parámetros dinámicos del cepstrum y de la energía al sistema de reconocimiento, tal como se describió en el apartado 4.5 de la memoria. En primer lugar, se incorporarán los parámetros dinámicos cepstrales (apartado 6.3.4.1) y posteriormente la energía y sus parámetros dinámicos (apartado 6.3.4.2)

6.3.4.1. PARAMETROS DINAMICOS CEPSTRALES

Como ya se comentó en el apartado 4.5 de la memoria existen fundamentalmente dos estrategias para incorporar varias informaciones a un sistema de reconocimiento basado en modelos ocultos de Markov discretos:

- Distancia compuesta, que consiste en construir un supervector concatenando con una ponderación adecuada los vectores y/o las componentes escalares que se desean utilizar y obtener un único símbolo correspondiente a este supervector usando distancia euclídea ponderada en el proceso de cuantificación vectorial.

- Diccionarios múltiples, en la que se cuantifican por separado cada una de las informaciones y se considera independencia estadística de las mismas en el entorno de los modelos ocultos de Markov.

En primer lugar se ha probado esta última estrategia, propuesta por Gupta [Gup87] y utilizada por K.F.Lee [Lee88a]. Las modificaciones en las fórmulas de evaluación, codificación y entrenamiento de los modelos ocultos de Markov en el caso de múltiples diccionarios se han descrito en el apartado 5.2.3.3 de la memoria.

En la figura 6.21 se muestran gráficamente las tasas de reconocimiento obtenidas con el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral de orden 12, añadiendo a la información del cepstrum que consideraba el sistema básico la información del delta-cepstrum, ver expresión (4.47), mediante la estrategia de diccionarios múltiples. Se muestra la variación de estos resultados en función del valor de K en la expresión (4.47), que define el intervalo de derivación desde t-K a t+K. Como distancia del cuantificador vectorial correspondiente al delta-cepstrum se ha elegido la misma que en el cuantificador del cepstrum (suponiendo siempre que los coeficientes delta-cepstrales se han calculado a partir de los coeficientes cepstrales no ponderados)

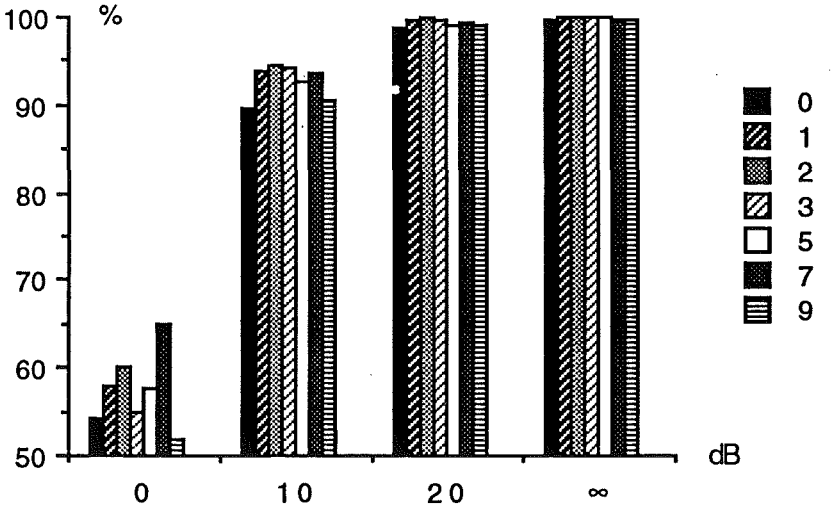

Fig. 6.21. Incorporación de la informaciór. delta-cepstrum mediante diccionarios múltiples

La columna en negro, correspondiente a K = 0, equivale a utilizar sólo la información de cepstrum. Por tanto, puede observarse en la figura que la incorporación de la información de delta-cepstrum aumenta las tasas de reconocimiento en todas las condiciones, ruidosas o libres de ruido. Se ha tomado para las pruebas siguientes el valor de K igual a 2, como valor de compromiso entre todas las condiciones consideradas. Observar el buen resultado obtenido con K igual a 7 y 0 dB de SNR.

En segundo lugar, se ha probado la estrategia de distancia compuesta. En este caso, en el que sólo hay información de cepstrum y delta-cepstrum, se construye un supervector concatenando los coeficientes cepstrales y delta-cepstrales y se utiliza como distancia del cuantificador la siguiente expresión

$$d_{WE}(T_1,T_2) = w_c \sum_{n=1}^{L} \left(w_c(n) \left(c_1(n)-c_2(n)\right)\right)^2 + w_{\Delta c} \sum_{n=1}^{L} \left(w_{\Delta c}(n) \left(\Delta c_1(n)-\Delta c_2(n)\right)\right)^2, \qquad (6.2)$$

donde T_1 y T_2 representan las dos tramas cuyas informaciones se comparan, $c_1(n)$ y $c_2(n)$ son los coeficientes cepstrales no ponderados de índice n correspondientes a las tramas T_1 y T_2, $\Delta c_1(n)$ y $\Delta c_2(n)$ son los coeficientes delta-cepstrales no ponderados de índice n correspondientes a las mismas tramas y w_c, $w_c(n)$, $w_{\Delta c}$ y $w_{\Delta c}(n)$ son ponderaciones, que pueden elegirse según diferentes criterios.

Dos posibles ponderaciones en la expresión (6.2) son

$$\begin{array}{lll} (a) & w_c = 1 & w_c(n) = \dfrac{1}{\sigma_{c(n)}} \\ & w_{\Delta c} = 1 & w_{\Delta c}(n) = \dfrac{1}{\sigma_{\Delta c(n)}}, \qquad (6.3) \\ (b) & w_c = \left(\prod_{n=1}^{L} \dfrac{1}{\sigma_{c(n)}} \right)^{\frac{1}{L}} & w_c(n) = n \\ & w_{\Delta c} = \left(\prod_{n=1}^{L} \dfrac{1}{\sigma_{\Delta c(n)}} \right)^{\frac{1}{L}} & w_{\Delta c}(n) = n, \qquad (6.4) \end{array}$$

donde $\sigma_{c(n)}$ y $\sigma_{\Delta c(n)}$ son las desviaciones típicas de los coeficientes c(n) y $\Delta c(n)$, respectivamente.

En la figura 6.22 se comparan las tasas de reconocimiento mostradas en la figura 6.21 para K = 2 (dicc. múlt.) con las obtenidas utilizando el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y vector cepstral de longitud 12 sin ponderar, añadiendo a la información del cepstrum que consideraba el sistema básico la información del delta-cepstrum mediante la estrategia de distancia compuesta (6.2), utilizando las ponderaciones (a) (6.3) y (b) (6.4). Por otro lado, teniendo en cuenta que el nuevo vector tiene longitud doble que los anteriores también se ha realizado una prueba utilizando la ponderación (6.4) y 128 palabras-código, en lugar de 64. A esta prueba se le ha asignado la etiqueta (c).

Se puede comprobar en esta figura que las mejores prestaciones se obtienen para la estrategia de múltiples diccionarios. Por tanto, esta será la técnica que se aplicará para incorporar parámetros dinámicos de orden superior.

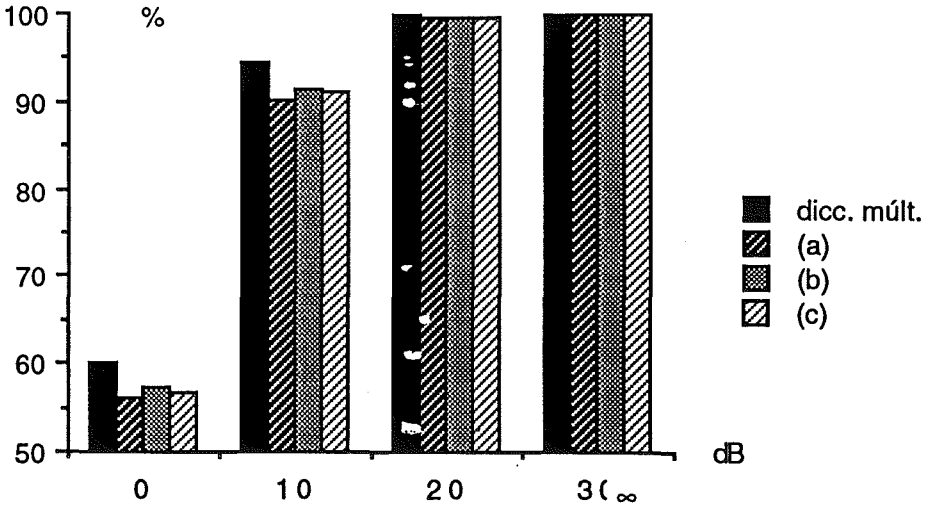

Fig 6.22.Comparación diccionarios múltiples-distancia compuesta para la incorporación de Δe

En la tabla 6.18 se comparan los resultados obtenidos sólo con información de cepstrum (c) y con las dos informaciones, cepstrum y delta-cepstrum (c,Δc). En esta misma tabla también se muestran los resultados obtenidos utilizando, en lugar del delta-cepstrum, el resultado de filtrar el cepstrum mediante el filtro paso-alto IIR (4.48) propuesto por Hirsch [Hir91]. En dicha tabla puede observarse que las prestaciones del delta-cepstrum y el cepstrum filtrado son muy similares y ambas mejoran ostensiblemente los resultados de utilizar sólo el cepstrum.

Parám. /SNR(dB)	sin ruido	20	10	0
c	99.8	98.9	89.5	54.2
c, Δc	99.9	99.9	94.4	60.2
c, c filtrado	99.9	99.8	94.1	59.7

Tabla 6.18. Prestaciones del delta-cepstrum y el cepstrum filtrado

A la hora de implementar los parámetros dinámicos de segundo orden, se han planteado dos alternativas: el uso de coeficientes de regresión de segundo orden y el uso del coeficiente de regresión de primer orden del delta-cepstrum, el delta-delta-cepstrum.

En las figuras 6.23 y 6.24 se muestran gráficamente las tasas de reconocimiento en tanto por ciento obtenidas utilizando el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral de orden 12, añadiendo a la información del cepstrum que consideraba el sistema básico la información del delta-cepstrum y del parámetro dinámico de segundo orden mediante la estrategia de diccionarios múltiples y la misma ventana de ponderación cepstral.

En la figura 6.23 el segundo parámetro dinámico se ha incorporado mediante el cálculo del delta-delta-cepstrum, el coeficiente de regresión de primer orden del delta-cepstrum. La variable de la gráfica indica el valor K en la expresión del cálculo del coeficiente de regresión, que define el intervalo de derivación desde t-K a t+K. Como se trata de una segunda derivada, el intervalo real de derivación es mayor que este, ya que hay que tener en cuenta el valor de K en el cálculo del delta-cepstrum.

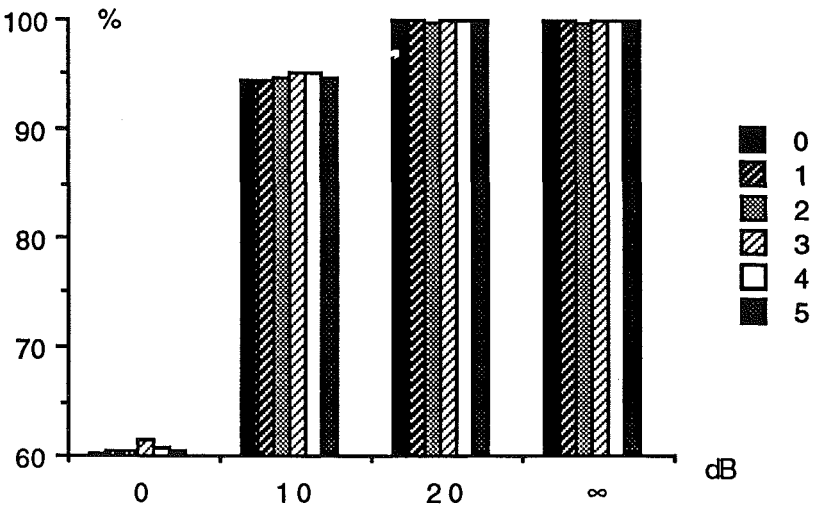

Fig. 6.23. Incorporación del delta-delta-cepstrum

En la figura 6.24 el segundo parámetro dinámico del cepstrum se ha implementado mediante el cálculo del coeficiente de regresión de segundo orden del cepstrum. La variable de la gráfica indica la longitud total del intervalo de derivación desde t-K a t+K de esta derivada de segundo orden.

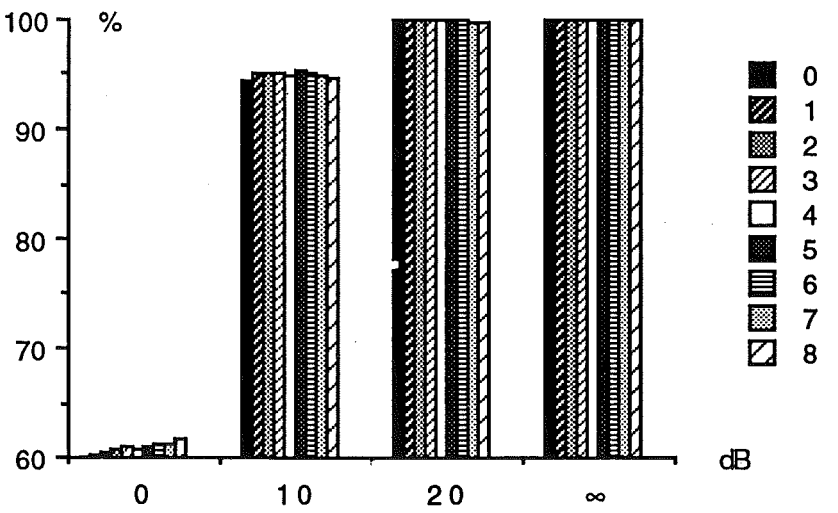

Fig. 6.24. Incorporación del coeficiente de regresión de segundo orden del cepstrum

Puede observarse que no hay mucha variación entre las prestaciones de las dos implementaciones del parámetro dinámico de segundo orden. Los mejores resultados se han obtenido utilizando el delta-delta-cepstrum con K=3.

En la tabla 6.19 se comparan los resultados obtenidos sólo con información de cepstrum (c), con dos informaciones, cepstrum y delta-cepstrum (c, Δc), y con tres informaciones, cepstrum, delta-cepstrum y delta-delta-cepstrum (c, Δc, $\Delta_2 c$), mediante la técnica de diccionarios múltiples.

Puede observarse que las prestaciones del sistema mejoran con la adición de un nuevo parámetro dinámico. Sin embargo, la mejora introducida por el delta-delta-cepstrum es menor que la introducida por el delta-cepstrum.

Parám. /SNR(dB)	sin ruido	20	10	0
c	99.8	98.9	89.5	54.2
c, Δc	99.9	99.9	94.4	60.2
c, Δc, $\Delta_2 c$	99.9	99.9	95.2	61.4

Tabla 6.19. Prestaciones del delta-cepstrum y del delta-delta cepstrum

Por último, se ha incorporado la información del parámetro dinámico de tercer orden del cepstrum en forma de delta-delta-delta-cepstrum, es decir, calculando el coeficiente de regresión de primer orden del delta-delta-cepstrum.

Los resultados obtenidos se muestran en la figura 6.25. La variable de la gráfica indica el valor K en la expresión del cálculo del coeficiente de regresión, que define el intervalo de derivación desde t-K a t+K. Como se trata de una tercera derivada, el intervalo real de derivación es mayor que este, ya que hay que tener en cuenta los valores de K en el cálculo del delta-cepstrum y del delta-delta-cepstrum.

El valor de K=0 indica la no incorporación del nuevo parámetro. Por tanto, puede observarse en la gráfica que la mejora introducida por el delta-delta-delta-cepstrum es muy pequeña. En caso de utilizar esta información, se podría elegir el parámetro K igual a 2.

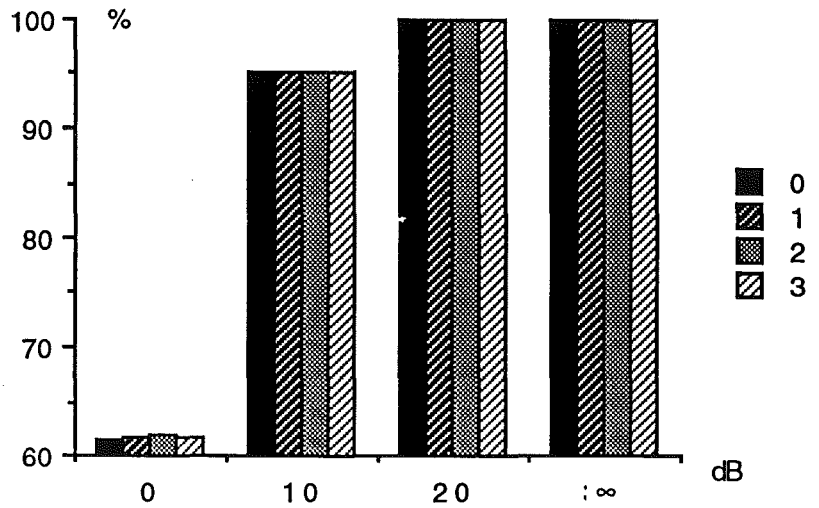

Fig. 6.25. Incorporación del delta-delta-delta-cepstrum

En la tabla 6.20 y en la figura 6.26 se comparan los resultados obtenidos sólo con información de cepstrum (c), con dos informaciones, cepstrum y delta-cepstrum (c, Δc), con tres informaciones, cepstrum, delta-cepstrum y delta-delta-cepstrum (c, Δc, $\Delta_2 c$), y con cuatro informaciones, cepstrum, delta-cepstrum, delta-delta-cepstrum y delta-delta-delta-cepstrum (c, Δc, $\Delta_2 c$, $\Delta_3 c$), mediante la técnica de diccionarios múltiples.

Parám. /SNR(dB)	sin ruido	20	10	0
c	99.8	98.9	89.5	54.2
c, Δc	99.9	99.9	94.4	60.2
c, Δc, $\Delta_2 c$	99.9	99.9	95.2	61.4
c, Δc, $\Delta_2 c$, $\Delta_3 c$	100	99.9	95.2	61.9

Tabla 6.20. Prestaciones del delta-cepstrum, del delta-delta-cepstrum y del delta-delta-delta-cepstrum

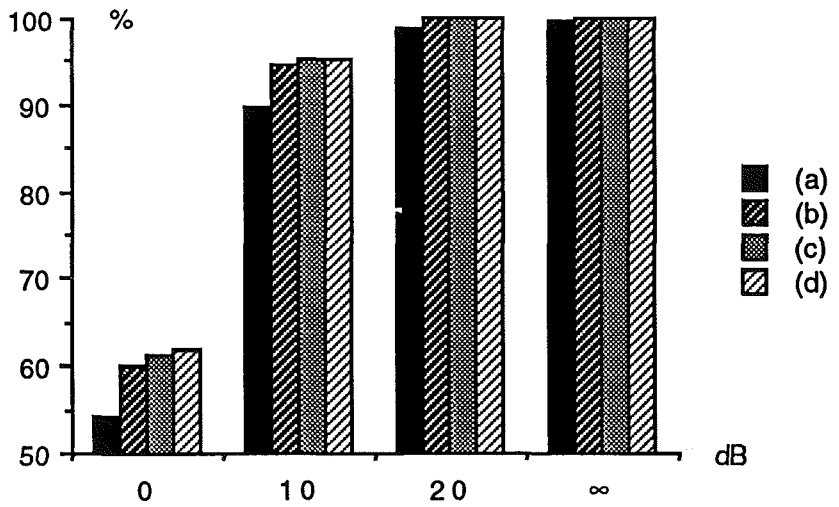

Fig. 6.26. Prestaciones del delta-cepstrum, del delta-delta-cepstrum y del delta-delta-delta-cepstrum

Puede observarse que las prestaciones del sistema mejoran con la adición de un nuevo parámetro dinámico. Sin embargo, la mejora introducida por el delta-delta-delta-cepstrum es ya muy pequeña y no compensa el coste computacional que conlleva, por lo que en lo sucesivo no se considerarán parámetros dinámicos de tercer orden.

6.3.4.2. ENERGIA Y PARAMETROS DINAMICOS

Para incorporar la información de la energía logarítmica local también se plantea la dicotomía diccionarios múltiples-distancia compuesta.

En primer lugar se probó la impiementación de dos diccionarios, uno para el cepstrum y otro para la energía. Para ello, fue necesario optimizar el número de palabras-código del diccionario de energía. En la figura 6.27 se comparan las tasas de reconocimiento obtenidas con el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos, ventana cepstral rampa de orden 12, con las obtenidas incorporando un nuevo diccionario de energía de M palabras-código.

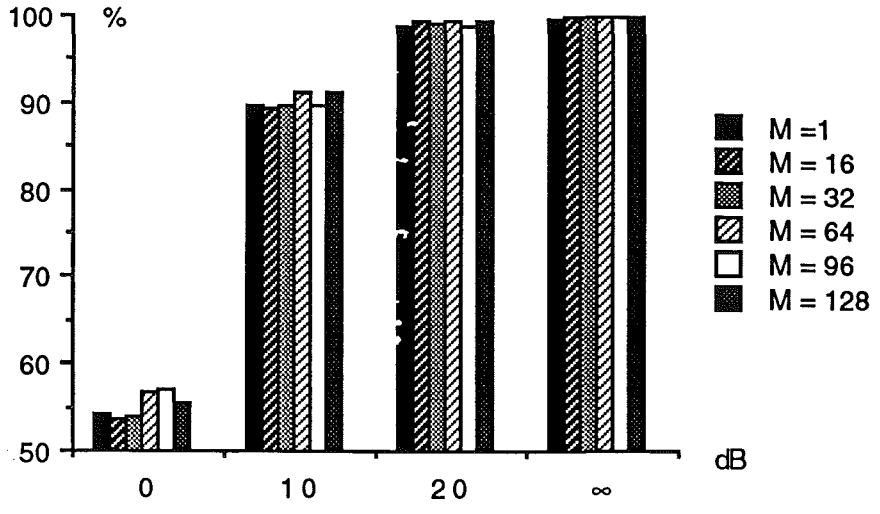

Fig. 6.27. Incorporación de un diccionario de energía de M palabras-código

En la figura puede observarse que la incorporación de la información de energía es beneficiosa para todas las condiciones consideradas. Esto parece contradecir lo afirmado en el apartado 4.5 de la memoria sobre la necesidad de normalizar la energía antes de su incorporación. La explicación es simplemente que la base de los dígitos catalanes con la que se han realizado estas pruebas ya está normalizada. En cuanto al número óptimo de palabras-código del diccionario de energía, se observa que la elección de un diccionario de igual tamaño que el de cepstrum, 64, es la que proporciona mejores resultados teniendo en cuenta todas la condiciones consideradas.

Seguidamente, se consideró la posibilidad de incorporar la información de energía mediante la estrategia de la distancia compuesta, que en este caso es de la forma

$$d_{WE}(T_1,T_2) = w_c \sum_{n=1}^{L} \left(w_c(n) \left(c_1(n)-c_2(n) \right) \right)^2 + w_E \left(E_1-E_2 \right)^2, \qquad (6.5)$$

donde T_1 y T_2 representan las dos tramas cuyas informaciones se comparan, $c_1(n)$ y $c_2(n)$ son los coeficientes cepstrales no ponderados de índice n correspondientes a las tramas T_1 y T_2, E_1 y E_2 son las energías logarítmicas correspondientes a las mismas

tramas y w_c, $w_c(n)$ y w_E son ponderaciones, que pueden elegirse según diferentes criterios.

Dos posibles ponderaciones en la expresión (6.2) son

$$(a)\quad w_c = 1 \qquad w_c(n) = \frac{1}{\sigma_{c(n)}} \qquad w_E = \frac{1}{\sigma_E} \tag{6.6}$$

$$(b)\quad w_c = \left(\prod_{n=1}^{L} \frac{1}{\sigma_{c(n)}} \right)^{\frac{1}{L}} \qquad w_c(n) = n \qquad w_E = \frac{\frac{1}{L} \sum_{n=1}^{L} n}{\sigma_E}, \tag{6.7}$$

donde $\sigma_{c(n)}$ y σ_E son las desviaciones típicas de c(n) y E, respectivamente.

En la figura 6.28 se comparan las tasas de reconocimiento mostradas en la figura 6.27 para M = 64 (dicc. múlt.) con las obtenidas utilizando el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y vector cepstral de longitud 12 sin ponderar, añadiendo a la información del cepstrum que consideraba el sistema básico la información de la energía mediante la estrategia de distancia compuesta (6.5), utilizando las ponderaciones (a) (6.6) y (b) (6.7).

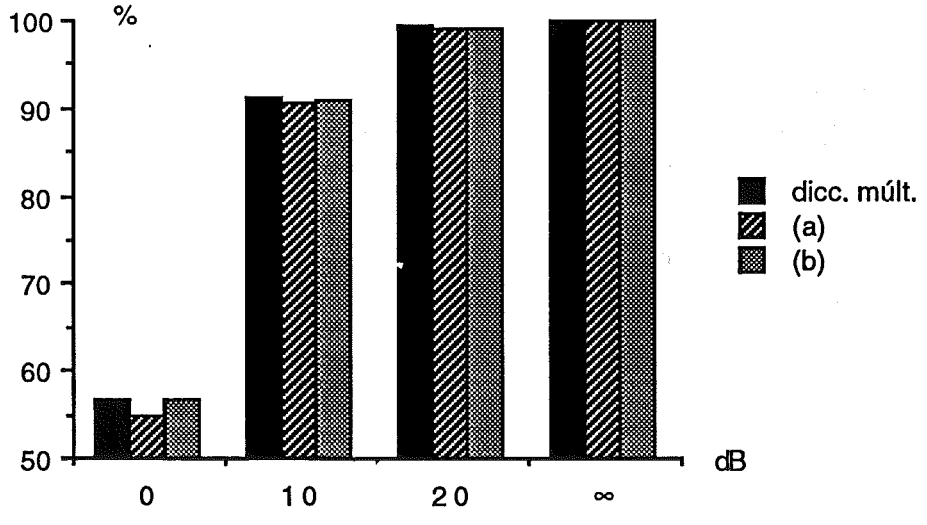

Fig. 6.28. Incorporación de la energía utilizando distancia compuesta

Se puede comprobar en esta figura que las mejores prestaciones se obtienen para la estrategia de múltiples diccionarios. Por tanto, esta será la técnica que se aplicará para incorporar parámetros dinámicos de orden superior.

En la tabla 6.21 se observa la mejora que supone la incorporación de la energía al sistema de reconocimiento básico, en la que sólo se usa la información de cepstrum.

Parám. /SNR(dB)	sin ruido	20	10	0
c	99.8	98.9	89.5	54.2
c, E	100	99.3	91.2	56.6

Tabla 6.21. Incorporación de la energía al sistema básico

Como se vio en el apartado 4.5 de la memoria también son útiles en reconocimiento del habla los parámetros dinámicos de la energía. Por analogía a los parámetros dinámicos usados en el caso del cepstrum, se utilizarán como parámetros dinámicos de la energía la delta-energía y la delta-delta-energía con los mismos valores de K. Los resultados obtenidos se muestran en las siguientes tablas de resultados.

En la tabla 6.22 se observa la importante mejora que se produce en un sistema que utiliza el cepstrum y el delta-cepstrum (c, Δc) al incorporar los parámetros energía y delta-energía (c, Δc, E, ΔE).

Parám. /SNR(dB)	sin ruido	20	10	0
c, Δc	99.9	99.9	94.4	60.2
c, Δc, E, ΔE	100	99.9	95.9	65.8

Tabla 6.22. Incorporación de E y ΔE en un sistema con c y Δc

Análogamente, en la tabla 6.23 se observa la mejora que se produce en un sistema que utiliza el cepstrum, el delta-cepstrum y el delta-delta-cepstrum (c, Δc, Δ_2c) al incorporar los parámetros energía, delta-energía y delta-delta-energía (c, Δc, Δ_2c, E, ΔE, Δ_2E).

Parám. /SNR(dB)	sin ruido	20	10	0
c, Δc, $\Delta_2 c$	99.9	99.9	95.2	61.4
c, Δc, $\Delta_2 c$, E, ΔE, $\Delta_2 E$	100	100	96.1	69.6

Tabla 6.23. Incorporación de E, ΔE y $\Delta_2 E$ a un sistema con c, Δc y $\Delta_2 c$

En la tabla 6.24 y en la figura 6.29 puede observa la gran mejora que se produce en reconocimiento de habla ruidosa a medida que se van añadiendo a la información del cepstrum parámetros dinámicos del cepstrum, energía y parámetros dinámicos de energía.

Parám. /SNR(dB)	sin ruido	20	10	0
c (a)	99.8	98.9	89.5	54.2
c, Δc (b)	99.9	99.9	94.4	60.2
c, Δc, $\Delta_2 c$ (c)	99.9	99.9	95.2	61.4
c, Δc, $\Delta_2 c$, E (d)	99.9	99.9	95.5	63.6
c, Δc, $\Delta_2 c$, E, ΔE (e)	100	99.9	96.0	66.6
c, Δc, $\Delta_2 c$, E, ΔE, $\Delta_2 E$ (f)	100	100	96.1	69.6

Tabla 6.24. Incorporación de energía y parámetros dinámicos

Algunos autores recomiendan la construcción de un cuantificador vectorial compartido por los parámetros de energía ponderados con las inversas de sus desviaciones típicas. En nuestra aplicación hemos probado esta estrategia, pero no ha supuesto ninguna mejora en reconocimiento de habla ruidosa.

En la tabla 6.25, se observa que esta estrategia (c, Δc, $\Delta_2 c$, E-ΔE-$\Delta_2 E$) da resultados muy similares al uso de la energía sin sus parámetros dinámicos (c, Δc, $\Delta_2 c$, E) y bastante peores que la estrategia que se ha tomado en esta tesis de separar todas las informaciones (c, Δc, $\Delta_2 c$, e, ΔE, $\Delta_2 E$).

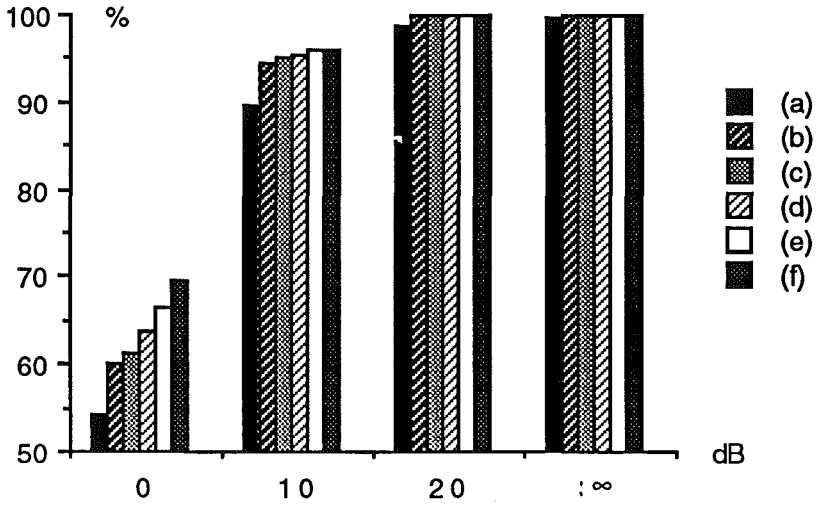

Fig. 6.29. Incorporación de energía y parámetros dinámicos

Parám. /SNR(dB)	sin ruido	20	10	0
c, Δc, Δ_2c, E	99.9	99.9	95.5	63.6
c, Δc, Δ_2c, E, ΔE, Δ_2E	100	100	96.1	69.6
c, Δc, Δ_2c, E-ΔE-Δ_2E	100	99.9	95.6	63.8

Tabla 6.25. Cuantificador vectorial de energía (E-ΔE-Δ_2E)

6.3.5. SUAVIZADO DE LOS MODELOS

En el apartado 5.2.2.4 de esta memoria se describieron algunos métodos de suavizado de la matriz de generación de observaciones de los modelos ocultos de Markov discretos. Además de la técnica de *floor-smoothig*, siempre utilizada, se revisaron tres métodos de suavizado basados en distancias: Parzen, distancias mutuas y correlaciones; y dos métodos basados en información mutua: coocurrencias y alineación de secuencias. En este apartado, se estudiará el efecto de estas técnicas en reconocimiento de habla ruidosa.

En la tabla 6.26 se muestran las tasas de reconocimiento obtenidas con el sistema básico de reconocimiento, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral de orden 12, en el

caso de aplicar solamente la técnica de *floor-smoothing* (-) y en el caso de utilizar, además, otras técnicas de suavizado más elaboradas.

Parám. /SNR(dB)	sin ruido	20	10	0
-	99.8	98.9	89.5	54.2
Parzen	99.3	98.6	96.0	69.8
Dist. mutuas	99.5	98.8	93.2	60.4
Correlaciones	98.8	97.0	95.6	63.2
Coocurrencias	99.6	98.9	95.9	62.2
Alineación sec.	99.6	98.6	95.8	61.8

Tabla 6.26. Efecto del suavizado

En la tabla puede observarse que, en general, todas las técnicas elaboradas de suavizado mejoran las prestaciones en condiciones de SNR bajas, 10 y 0 dB, mientras que producen una merma en los resultados en las condiciones menos ruidosas. El caso más claro es el de la técnica de Parzen, que alcanza una tasa de reconocimiento de casi el 70% a 0 dB de SNR y en ausencia de ruido da lugar a una tasa del 99.3%.

En el caso de las técnicas basadas en distancias la interpretación de este efecto es bastante sencilla. Para SNR altas, se degrada el comportamiento del sistema, pues aumenta la confusibilidad entre palabras-código. Para SNR bajas, aunque no se minimizan los errores de cuantificación debidos al ruido, estos errores tienen menor influencia en la estimación de la probabilidad a posteriori del modelo, ya que la técnica de suavizado impone semejanza a las probabilidades de las palabras-código cercanas y precisamente los errores de cuantificación debidos al ruido consisten en cambios de asignación entre estas palabras-código.

6.3.6. MODELOS SEMICONTINUOS Y MULTIPLE ETIQUETADO

En este apartado se compararán las prestaciones de los modelos ocultos de Markov discretos, semicontinuos y de múltiple etiquetado, que fueron descritos en los apartados 5.2, 5.4 y 5.5 de la memoria. respectivamente. Los modelos continuos no fueron probados por las razones comentadas en el apartado 5.3.

En la tabla 6.27 se comparan las tasas de reconocimiento expresadas en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos catalanes con el sistema básico de reconocimiento descrito en el apartado 6.2, sin preénfasis, utilizando un diccionario de 64 palabras-código, modelos de 10 estados sin saltos y ventana cepstral de orden 12, con las obtenidas sustituyendo en el mismo sistema los modelos ocultos de Markov discretos (DHMM) por modelos semicontinuos (SCHMM) y de múltiple etiquetado (MLHMM).

Además de los modelos semicontinuos con funciones de densidad gaussianas en el cuantificador vectorial (SCHMM-Gauss), se han probado modelos semicontinuos con laplacianas (SCHMM-Laplace). Por otro lado, también se han probado las prestaciones de dos tipos de modelos de múltiple etiquetado: los MLHMM-A son los que no utilizan en la reestimación de las probabilidades de generación de observaciones de los modelos los valores anteriores de dichas probabilidades, expresión (5.87); los MLHMM-B son los que utilizan la estimación de máxima probabilidad, expresión (5.86). En todo estos modelos se ha tomado K, el número de palabras-código consideradas en la cuantificación de cada vector de características, igual a 5.

Modelos /SNR(dB)	sin ruido	20	10	0
DHMM	99.8	98.9	89.5	54.2
SCHMM-Gauss	99.8	98.9	96.4	72.8
SCHMM-Laplace	99.7	98.7	95.9	71.9
MLHMM-A	99.7	98.8	96.5	74.1
MLHMM-B	99.9	99.1	93.4	62.6

Tabla 6.27. Comparación de los distintos tipos de modelos

Puede observarse en la tabla que todos los tipos de modelos semicontinuos y de múltiple etiquetado superan notablemente en prestaciones a los modelos discretos en condiciones ruidosas y, además, no se produce un empeoramiento en condiciones libres de ruido.

Estas prestaciones superiores de los modelos de Markov semicontinuos y de múltiple etiquetado son debidas a una disminución de los errores de cuantificación. En el proceso de cuantificación vectorial de los modelos ocultos de Markov discretos (DHMM) se elige únicamente la palabra-código que dista menos del vector a cuantificar

y se descarta la información sobre el grado en que dicho vector se ajusta a otras palabras-código. Esta información puede ser especialmente importante en el caso de habla ruidosa, ya que la decisión tomada por el cuantificador del modelo discreto puede ser fácilmente modificada por el ruido añadido a la señal. Sin embargo, en los modelos semicontinuos (SCHMM) y de múltiple etiquetado (MLHMM) el cuantificador vectorial proporciona información sobre la distancia relativa a las K palabras-código más cercanas y, por tanto, se conserva parte de esta información.

En la figura 6.30, se ha intentado representar el efecto que puede suponer la presencia de ruido en la cuantificación. Las palabras-código se han representado mediante círculos y el vector a cuantificar, que es trasladado por efecto del ruido, mediante aspas. En el caso discreto (a), la presencia de ruido puede llegar a cambiar la elección de la palabra-código. En el caso semicontinuo o de múltiple etiquetado (b), la presencia de ruido, en el caso de la figura, sólo cambiaría ligeramente las ponderaciones de las palabras-código con las que se cuantifica. Como puede verse, los efectos en este último caso son menos drásticos.

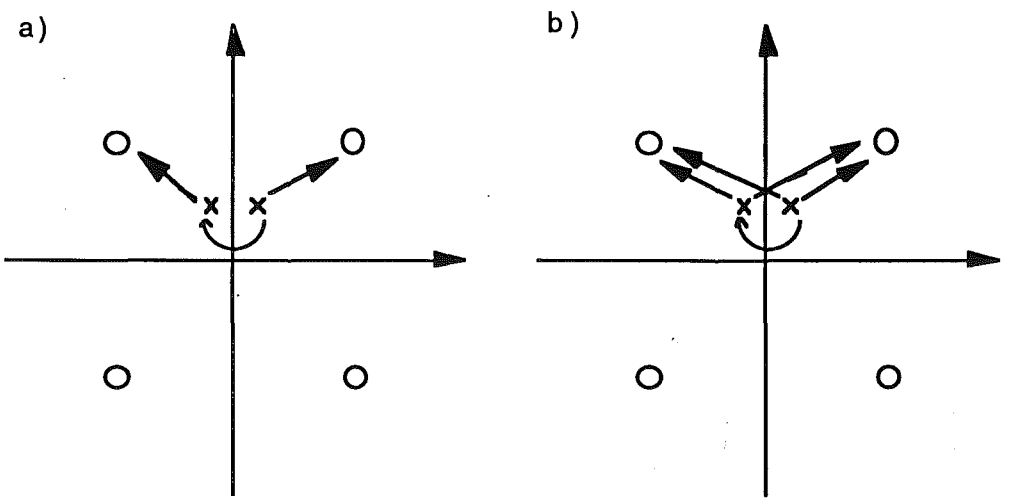

Fig. 6.30. Efecto del ruido en la cuantificación vectorial en el caso discreto (a) y en el caso semicontinuo o de múltiple etiquetado (b)

Puede verse en la tabla 6.27 que los resultados de los SCHMM son un poco peores que los correspondientes a los MLHMM. Además, el coste computacional de los MLHMM es mucho menor.

En cuanto a los modelos de múltiple etiquetado, los modelos MLHMM-A son mucho más eficientes computacionalmente que los MLHMM-B. En cuanto a sus prestaciones, aunque los resultados de los MLHMM-A son ligeramente peores para SNR altas, son muchísimo mejores en condiciones severas de ruido [Her93a]. Por todo ello, en la pruebas futuras se considerarán los modelos con múltiple etiquetado de tipo A.

En la figura 6.31 se observa la variación de las tasas de reconocimiento proporcionadas por los modelos con múltiple etiquetado de tipo A en función de K, el número de palabras-código consideradas en la cuantificación de un vector de características.

Como puede verse en la figura, el valor K = 5, que se ha utilizado en la tabla 6.27, parece ser una buena elección. Este valor es el que se considerará en pruebas futuras.

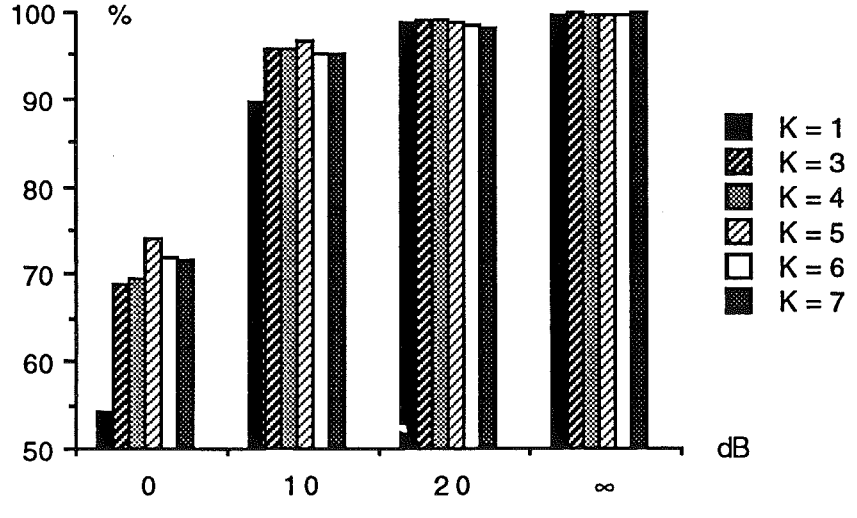

Fig. 6.31. Influencia del valor de K en los modelos MLHMM-A

Por último, en la figura 6.32 se observa la diferencia de prestaciones entre los modelos ocultos de Markov discretos (DHMM) y los de múltiple etiquetado de tipo A (MLHMM-A).

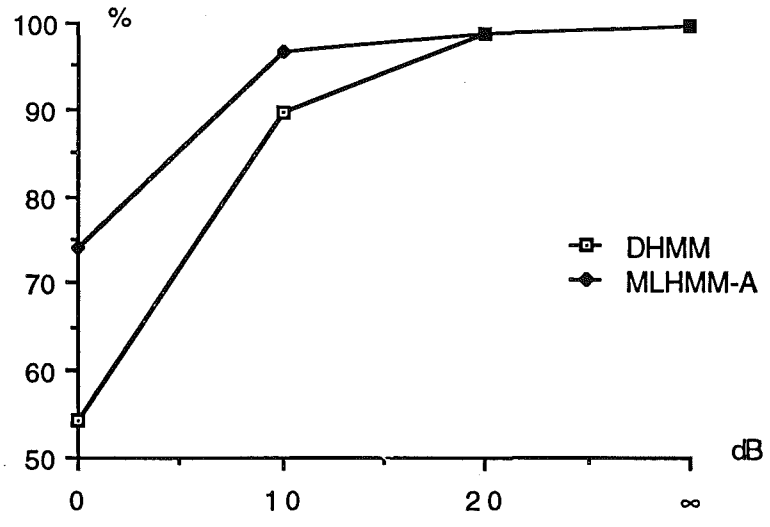

Fig. 6.32. Comparación de los modelos DHMM y los MLHMM-A

En las pruebas futuras se utilizara el múltiple etiquetado de tipo A, en lugar del suavizado de los modelos, los modelos semicontinuos y los modelos de múltiple etiquetado de tipo B, por sus mejores prestaciones.

6.3.7. COMPARACION Y COMBINACION DE TECNICAS

En los apartados anteriores se han obtenido buenos resultados en reconocimiento robusto del habla sustituyendo la parametrización clásica LPC por la nueva parametrización OSALPC2; la distancia euclídea, por la distancia de proyección; la información espectral instantánea, por las informaciones espectral y de energía instantáneas y dinámicas; y la cuantificación vectorial clásica, por el múltiple etiquetado. Sin embargo, estas técnicas sólo se han aplicado por separado.

En este apartado, se realiza un estudio comparativo de estas técnicas y se investigan posibles combinaciones de las mismas que puedan proporcionar mejoras todavía mayores a las obtenidas mediante su aplicación de forma aislada.

A efectos de claridad en la exposición, en primer lugar se presentan por separado los resultados obtenidos con las parametrizaciones LPC y OSALPC2 y, al final del apartado, se comparan los resultados obtenidos con ambas técnicas.

En la tabla 6.28 se comparan las tasas de reconocimiento en tanto por ciento correspondientes a la parametrización LPC en función de la distancia del cuantificador vectorial -euclídea o proyección-, el tipo de modelos - discretos o múltiple etiquetado- y el número de informaciones. Todas las pruebas se han realizado sin preénfasis, con modelos de izquierda a derecha de 10 estados sin saltos, diccionarios de 64 palabras-código y ventana cepstral rampa de orden 12.

En el caso de la distancia de proyección con varias informaciones, se ha utilizado esta distancia no sólo en el cuantificador del cepstrum sino también en los cuantificadores correspondientes a sus parámetros dinámicos. Con ello se consiguen resultados ligeramente superiores a los obtenidos utilizando distancia euclídea en los cuantificadores de los parámetros dinámicos del cepstrum.

Dist.	Mod.	Nº inf. /SNR(dB)	sin ruido	20	10	0
euc.	DHMM	c	99.8	98.9	89.5	54.2
"	"	c, Δc, Δ_2c, E, ΔE, Δ_2E	100	100	96.1	69.6
"	MLHMM	c	99.7	98.8	96.5	74.1
"	"	c, Δc, Δ_2c, E, ΔE, Δ_2E	100	99.9	97.2	78.8
proy.	DHMM	c	99.8	98.6	89.2	65.6
"	"	c, Δc, Δ_2c, E, ΔE, Δ_2E	100	99.9	96.8	77.4
"	MLHMM	c	99.7	98.9	95.4	72.8
"	"	c, Δc, Δ_2c, E, ΔE, Δ_2E	100	99.9	97.0	78.2

Tabla 6.28. Comparación y combinación de técnicas con parametrización LPC

Puede observarse en esta tabla que la utilización de varias informaciones, en lugar de sólo el cepstrum, mejora los resultados en todos los casos [Her92a]. Por otro lado, si no se tienen en cuenta los resultados en ausencia de ruido, cuyas pequeñas diferencias en estas pruebas tienen poca fiabilidad estadística, el múltiple etiquetado obtiene excelentes resultados en comparación con la cuantificación vectorial clásica. Por último, puede observarse que la distancia de proyección solamente proporciona mejoras con respecto a la euclídea en las condiciones más severas de ruido y siempre que no se realice múltiple etiquetado. Por esta última razón, no se realizarán más pruebas combinando distancia de proyección y múltiple etiquetado.

En la figura 6.33 se muestran, para el caso de ruido blanco y parametrización LPC, las prestaciones de la energía y los parámetros dinámicos y del múltiple etiquetado. La etiqueta (a) corresponde al uso único del cepstrum con cuantificación clásica, la etiqueta (b) corresponde al uso de las seis informaciones (c, Δc, Δ_2c, E, ΔE, Δ_2E) con cuantificación clásica, la etiqueta (c) corresponde al uso único del cepstrum con múltiple etiquetado y la etiqueta (d) corresponde al uso de las seis informaciones con múltiple etiquetado.

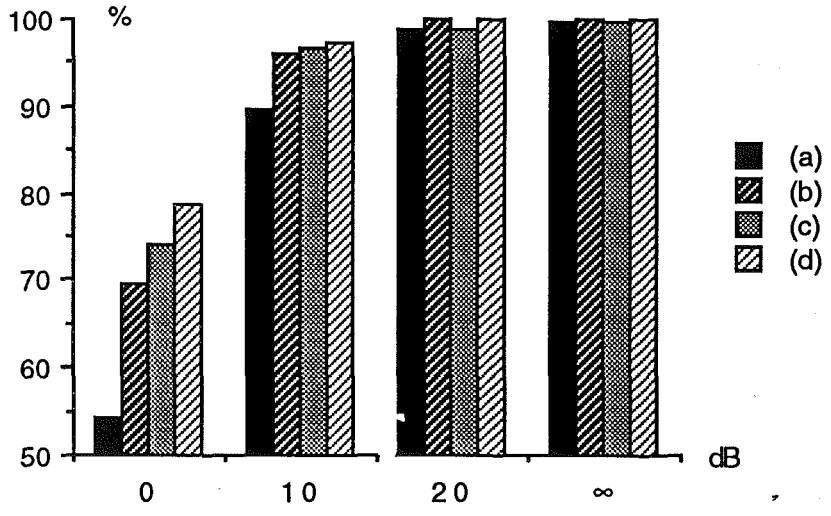

Fig. 6.33. Incorporación de informaciones y el múltiple etiquetado con parametrización LPC

Por lo que respecta a la parametrización OSALPC2, en primer lugar se plantea la necesidad de averiguar si el uso de los parámetros dinámicos del cepstrum OSALPC2, la energía y sus parámetros dinámicos contribuye a una mejora de las prestaciones que se consiguen utilizando únicamente el cepstrum OSALPC2.

No se han considerado en este trabajo los parámetros dinámicos de segundo orden en el caso de la parametrización OSALPC2. En cuanto a los parámetros de primer orden, en la tabla 6.29 se muestran los resultados obtenidos mediante varias combinaciones del cepstrum (c), la energía (E), el delta-cepstrum (Δc) y la delta-energía (ΔE).

Parám. / SNR(dB)	sin ruido	20	10	0
c	99.4	98.4	94.7	72.2
c, Δc	98.9	98.3	96.0	75.5
c, E	98.6	98.6	93.2	65.9
c, ΔE	99.1	98.7	96.2	81.4
c, Δc, ΔE	98.9	99.0	97.4	80.4

Tabla 6.29. Incorporación de parámetros dinámicos de primer orden con par. OSALPC2

Los mejores resultados en condiciones ruidosas se obtienen utilizando c, Δc, ΔE, a costa de un empeoramiento de la prestaciones en ausencia de ruido. En la figura 6.34 se comparan gráficamente los resultados obtenidos utilizando sólo cepstrum (a) y utilizando cepstrum, delta-cepstrum y delta-energía (b).

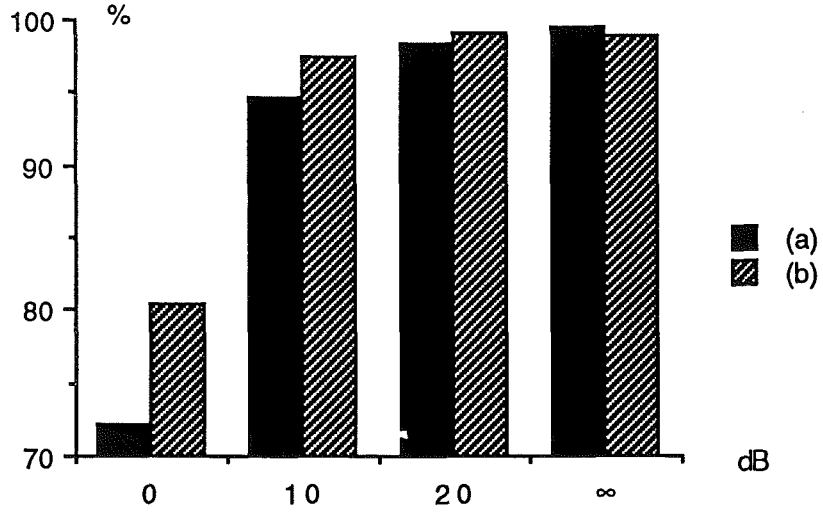

Fig. 6.34. Comparación de resultados de c (a) y c-Δc-ΔE (b) con par. OSALPC2

En cuanto a los resultados obtenidos utilizando distancia de proyección y múltiple etiquetado, estos se han resumido en la tabla 6.30 junto con los resultados anteriores más importantes. No se presentan las pruebas combinando distancia de proyección y múltiple etiquetado por las razones comentadas al abordar la técnica LPC.

Dist.	Mod.	Nº inf. /SNR(dB)	sin ruido	20	10	0
euc.	DHMM	c	99.4	98.4	94.7	72.2
"	"	c, Δc, ΔE	98.9	99.0	97.4	80.4
"	MLHMM	c	97.5	97.1	94.2	82.6
"	"	c, Δc, ΔE	98.1	97.7	97.1	85.4
proy.	DHMM	c	98.8	98.2	95.6	74.8
"	"	c, Δc, ΔE	99.3	99.3	97.1	72.6

Tabla 6.30. Comparación de técnicas con parametrización OSALPC2 y ruido blanco

En esta tabla puede observarse que la utilización de varias informaciones mejora siempre las prestaciones del sistema, salvo en el caso de ausencia de ruido. En cuanto al múltiple etiquetado, proporciona excelentes resultados pero únicamente en las condiciones más severas de ruido consideradas, 0 dB de SNR. Por último, la distancia de proyección sólo proporciona una mejora muy ligera de los resultados en condiciones severas de ruido, 10 y 0 dB de SNR, si se utiliza únicamente información de cepstrum y cuantificación vectorial clásica .

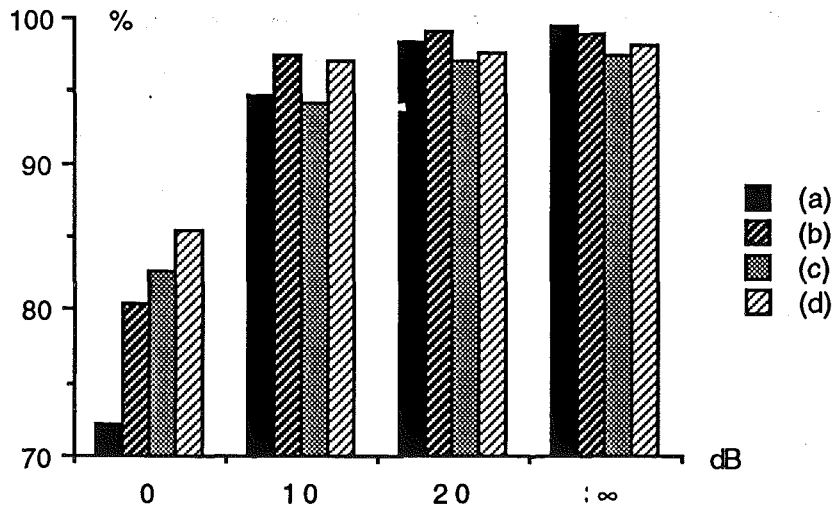

Fig. 6.35. Utilización de varias informaciones y múltiple etiquetado con parametrización OSALPC2 y ruido blanco

En la figura 6.35 se muestran, para el caso de ruido blanco y parametrización OSALPC2, las prestaciones de los parámetros dinámicos de primer orden y del múltiple etiquetado. La etiqueta (a) corresponde al uso único del cepstrum con cuantificación clásica, la etiqueta (b) corresponde al uso de cepstrum, delta-cepstrum y delta-energía con cuantificación clásicacorresponde al uso único del cepstrum con múltiple etiquetado, la etiqueta (c) corresponde al uso único del cepstrum con múltiple etiquetado y la etiqueta (d) corresponde al uso de cepstrum, delta-cepstrum y delta-energía con múltiple etiquetado.

En la tabla 6.31 se comparan los resultados obtenidos con ambas parametrizaciones, LPC y OSALPC2, utilizando sólo cepstrum o también energía y parámetros dinámicos de primer orden, siempre con distancia euclídea y cuantificación vectorial clásica.

Etiq.	Param..	Nº inf. /SNR(dB)	sin ruido	20	10	0
(a)	LPC	c	99.8	98.9	89.5	54.2
(b)	"	c, Δc, E, ΔE	100	99.9	95.9	65.8
(c)	OSALPC2	c	99.4	98.4	94.7	72.2
(d)	"	c, Δc, ΔE	98.9	99.0	97.4	80.4

Tabla 6.31. Comparación de las parametrizaciones LPC y OSALPC2 en ruido blanco variando el número de informaciones

En la figura 6.36 se representan gráficamente estos resultados (las etiquetas se corresponden con las de la tabla). En el apartado 6.3.2.3 de este capítulo se había observado y justificado que, en el caso de utilizar solamente la información de cepstrum, la parametrización OSALPC2 proporciona mejores resultados que la clásica LPC en condiciones severas de ruido, 10 y 0 dB de SNR. En esta figura puede observarse que este efecto se mantiene en el caso de utilizar energía y parámetros dinámicos de primer orden.

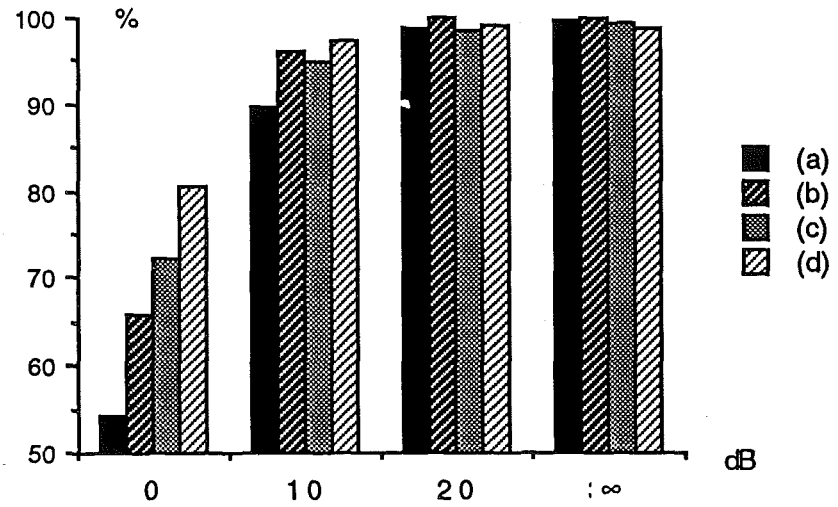

Fig. 6.36. Comparación de las parametrizaciones LPC y OSALPC2 en ruido blanco variando el número de informaciones

Teniendo en cuenta que la distancia de proyección en el caso de la técnica LPC sólo proporciona una mejora de resultados cuando no se realiza múltiple etiquetado y que en el caso de de la parametrización OSALPC2 estas mejoras son muy pequeñas, en la tabla 6.32 se resumen los resultados más importantes obtenidos hasta ahora en ruido blanco por orden de prestaciones en las condiciones más severas de ruido.

Etiq.	Par.-dist.	Mod.	Nº inf. / SNR(dB)	∞	20	10	0
(a)	LPC-euc.	DHMM	c	99.8	98.9	89.5	54.2
(b)	LPC-proy.	DHMM	c	99.8	98.6	89.2	65.6
(c)	LPC-euc.	DHMM	c, Δc, Δ_2c, E, ΔE, Δ_2E	100	100	96.1	69.6
(d)	OSALPC2	DHMM	c	99.4	98.4	94.7	72.2
(e)	LPC-euc.	MLHMM	c	99.7	98.8	96.5	74.1
(f)	LPC-proy.	DHMM	c, Δc, Δ_2c, E. ΔE, Δ_2E	100	99.9	96.8	77.4
(g)	LPC-euc.	MLHMM	c, Δc, Δ_2c, E, ΔE, Δ_2E	100	99.9	97.2	78.8
(h)	OSALPC2	DHMM	c, Δc, ΔE	98.9	99.0	97.4	80.4
(i)	OSALPC2	MLHMM	c	97.5	97.1	94.2	82.6
(j)	OSALPC2	MLHMM	c, Δc, ΔE	98.1	97.7	97.1	85.4

Tabla 6.32 Comparación de técnicas en ruido blanco

En la figura 6.37 se representan gráficamente estos resultados (las etiquetas se corresponden con las de la tabla). Puede observarse que las mejores prestaciones en condiciones severas de ruido se obtienen con parametrización OSALPC2, varias informaciones y múltiple etiquetado. Sin embargo, se observa de nuevo que el uso de la parametrización OSALPC2 conlleva una merma de las prestaciones del sistema en ausencia de ruido.

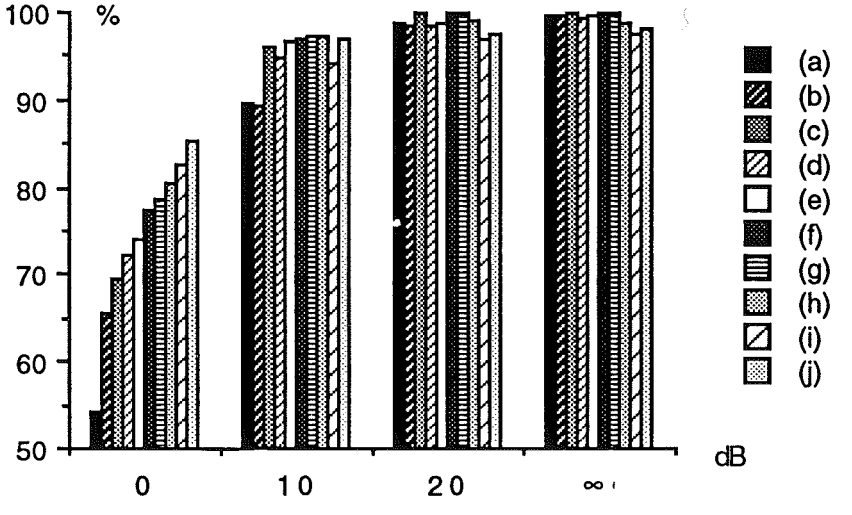

Fig. 6.37. Comparación de técnicas en ruido blanco

6.4. RESULTADOS OBTENIDOS CON RUIDO DE COCHE

En este apartado, las técnicas estudiadas y optimizadas hasta ahora en el caso de ruido blanco se aplicarán al caso de ruido de coche utilizando el mismo sistema básico de reconocimiento descrito en el apartado 6.2 y la base de datos descrita en el apartado 6.1.2.

El apartado se ha dividido en tres subapartados: un dedicado a la parametrización clásica LPC (6.4.1), otro dedicado a la parametrización propuesta en esta tesis OSALPC2 (6.4.2) y en el último se comparan los resultados correspondientes a ambas parametrizaciones.

6.4.1. PARAMETRIZACION LPC

En la tabla 6.33 se muestran las tasas de reconocimiento en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos italianos con el sistema básico de reconocimiento descrito en el apartado 6.2, sin préenfasis, utilizando modelos de izquierda a derecha de 10 estados sin saltos, diccionario de 64 palabras-código y los órdenes de predicción lineal y ventanas cepstrales indicados. Los resultados se presentan en función de la velocidad del coche en km/h.

Orden	Vent. cep. / veloc.	v = 0	v = 70	v = 130
8	Rectangular	88.7	81.2	52.0
	Seno	93.0	89.0	59.0
	Inv. desv. típica	91.0	85.4	59.5
	Rampa	90.7	87.1	53.2
12	Rectangular	95.5	84.6	60.2
	Seno	96.2	89.2	72.0
	Inv. desv. típica	95.7	86.1	62.5
	Rampa	95.0	89.4	69.7
16	Rectangular	91.7	68.1	51.2
	Seno	92.7	85.5	69.7
	Inv. desv. típica	95.5	86.6	81.5
	Rampa	93.5	79.6	61.5

Tabla 6.33. Influencia del orden de predicción y la ventana cepstral sin preénfasis

En la tabla 6.34 se muestran los resultados de las mismas pruebas, pero utilizando preénfasis. El parámetro a del filtro de preénfasis se ha tomado igual a 0.95, que es el valor más usado en los sistemas automáticos de reconocimiento del habla cuando no se aborda el problema del ruido ambiente

Orden	Vent. cep. / veloc.	v = 0	v = 70	v = 130
8	Rectangular	91.2	80.4	41.7
	Seno	93.7	88.9	58.2
	Inv. desv. típica	93.0	84.5	61.2
	Rampa	93.2	85.1	59.7
12	Rectangular	95.5	78.5	75.2
	Seno	96.7	93.9	71.0
	Inv. desv. típica	95.2	84.1	60.0
	Rampa	93.2	91.4	75.2
16	Rectangular	88.7	69.5	56.2
	Seno	90.7	86.1	66.2
	Inv. desv. típica	97.5	92.1	79.0
	Rampa	92.7	85.6	72.0

Tabla 6.34. *Influencia del orden de predicción y la ventana cepstral con preénfasis*

Los mejores resultados se han obtenido con la ventana inversa de la desviación típica de orden 16 y con preénfasis. Se trata también en este caso de una ventana cepstral creciente de orden alto, como ocurrió en el caso de ruido blanco (ventana rampa de orden 12). En cuanto a la preénfasis, en este tipo de ruido ha resultado conveniente utilizar preénfasis, mientras que en el caso de ruido blanco se obtenían mejores resultados sin realizar preénfasis. Esto es debido a que el ruido del coche tiende a disminuir con la frecuencia y el realce de las zonas de alta frecuencia no provoca los resultados no deseados que se producen en ruido blanco. En las pruebas futuras de parametrización LPC con ruido de coche se utilizara preénfasis y ventana cepstral inversa de la desviación típica de orden 16.

Seguidamente se realizaron pruebas para optimizar la longitud del intervalo de derivación para calcular el delta-cepstrum e incorporarlo utilizando la estrategia de diccionarios múltiples. En la figura 6.38 se muestran gráficamente los resultados obtenidos en función del valor de K, que define el intervalo de derivación desde t-K a t+K. El valor de K=0 se corresponder con el caso de usar sólo el cepstrum.

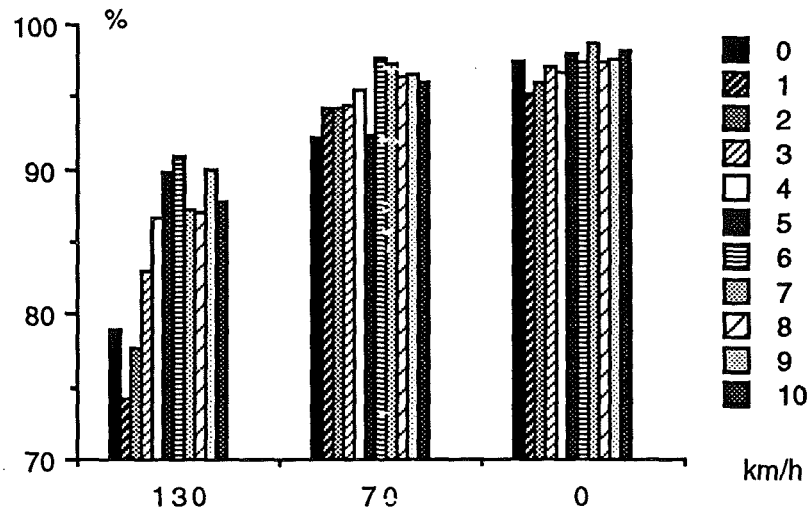

Fig. 6.38. Incorporación del delta-cepstrum

Como puede observarse, en el caso de ruido de coche el valor de K ha de ser mayor que 2 (valor elegido en el caso de ruido blanco) para obtener buenos resultados. En concreto se ha elegido para realizar las pruebas siguientes el valor K = 7.

Las pruebas que se realizaron usando los parámetros dinámicos de segundo orden proporcionaron mejoras bastante pequeñas en los resultados de reconocimiento, por lo que se decidió prescindir de ellos.

En la tabla 6.35 se muestran los resultados obtenidos mediante varias combinaciones del cepstrum (c), la energía (E), el delta-cepstrum (Δc) y la delta-energía (ΔE).

Parám. / veloc.	v = 0	v = 70	v =130
c	97.5	92.1	79.0
c, Δc	98.7	97.2	87.2
c, E	95.2	91.9	82.2
c, ΔE	97.5	93.6	87.2
c, Δc, ΔE	98.5	96.6	92.0

Tabla 6.35. Incorporación de parámetros dinámicos de primer orden

Puede observarse en esta tabla que la utilización de la energía logarítmica no proporciona buenos resultados. Esto es debido a que la base de datos utilizada en estas pruebas no está normalizada y para utilizar la energía haría falta una normalización previa de la misma.

Así pues, los mejores resultados se obtienen utilizando c, Δc, ΔE. En la figura 6.39 se comparan gráficamente los resultados obtenidos utilizando sólo cepstrum (a) y utilizando cepstrum, delta-cepstrum y delta-energía (b). Puede observarse una mejora notoria de resultados.

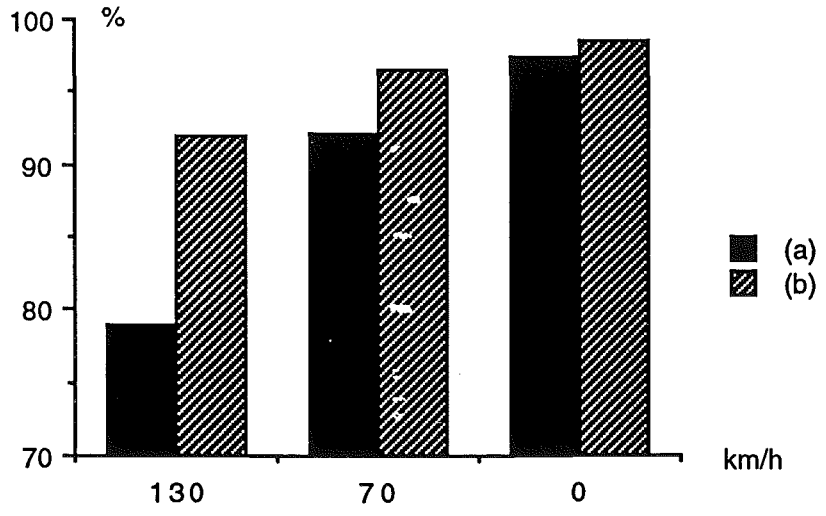

Fig. 6.39. Incorporación de parámetros dinámicos de primer orden

En cuanto a los resultados obtenidos utilizando distancia de proyección y modelos con múltiple etiquetado de tipo A, estos se han resumido en la tabla 6.36 junto con los resultados anteriores más importantes. En esta tabla puede observarse que la distancia de proyección no mejora el comportamiento del sistema de reconocimiento. Sin embargo, con el múltiple etiquetado se obtienen excelentes resultados.

Dist.	Mod.	Nº inf. / veloc.	0	70	130
euc.	DHMM	c	97.5	92.1	79.0
"	"	c, Δc, ΔE	98.5	96.6	92.0
"	MLHMM	c	98.2	94.9	81.7
"	"	c, Δc, ΔE	99.2	96.6	94.0
proy.	DHMM	c	95.2	88.9	67.2
"	"	c, Δc, ΔE	99.0	97.1	87.5

Tabla 6.36. Comparación de técnicas con parametrización LPC y ruido de coche

En la figura 6.40 se muestran, para el caso de ruido de coche y parametrización LPC, las prestaciones de los parámetros dinámicos de primer orden y del múltiple etiquetado. La etiqueta (a) corresponde al uso único del cepstrum con cuantificación clásica, la etiqueta (b) corresponde al uso único del cepstrum con múltiple etiquetado, la etiqueta (c) corresponde al uso de cepstrum, delta-cepstrum y delta-energía con cuantificación clásica y la etiqueta (d) corresponde al uso de cepstrum, delta-cepstrum y delta-energía con múltiple etiquetado.

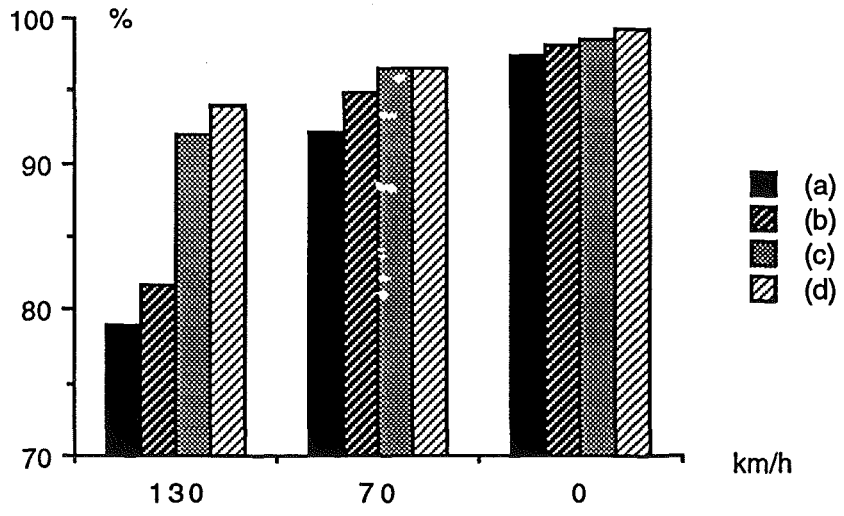

Fig. 6.40. Utilización de varias informaciones y múltiple etiquetado con parametrización LPC y ruido de coche

6.4.2. PARAMETRIZACION OSALPC2

En la tabla 6.37 se muestran las tasas de reconocimiento en tanto por ciento obtenidas en la aplicación del reconocimiento multilocutor de los dígitos italianos con el sistema básico de reconocimiento descrito en el apartado 6.2, sin préenfasis, utilizando modelos de izquierda a derecha de 10 estados sin saltos, diccionario de 64 palabras-código y los órdenes de predicción lineal y ventanas cepstrales indicados, sustituyendo en el sistema básico de reconocimiento la predicción lineal clásica por la parametrización OSALPC2 propuesta en esta tesis. Los resultados se presentan en función de la velocidad del coche en km/h. No se han presentado los resultados con ventana rectangular porque ya se ha comprobado que da lugar a tasas de reconocimiento bajas.

Orden	Vent. cep. / veloc.	v = 0	v = 70	v = 130
8	Seno	88.0	78.43	70.75
	Inv. desv. típica	80.7	79.93	69.5
	Rampa	89.0	82.36	75.0
12	Seno	87.7	81.9	74.5
	Inv. desv. típica	91.0	79.6	73.5
	Rampa	87.7	79.2	66.5
16	Seno	90.7	86.1	66.2
	Inv. desv. típica	86.7	80.0	68.0
	Rampa	89.0	75.3	73.0

Tabla 6.37. Influencia del orden de predicción y la ventana cepstral sin preénfasis

En la tabla 6.38 se muestran los resultados de las mismas pruebas, pero utilizando preénfasis con a = 0.95.

Orden	Vent. cep. / veloc.	v = 0	v = 70	v = 130
8	Seno	91.2	83.4	71.7
	Inv. desv. típica	93.5	82.6	72.2
	Rampa	91.7	85.1	68.0
12	Seno	96.7	89.3	74.5
	Inv. desv. típica	95.5	87.9	77.2
	Rampa	91.0	87.1	76.2
16	Seno	92.7	85.5	69.7
	Inv. desv. típica	95.5	91.2	80.7
	Rampa	96.0	94.6	85.0

Tabla 6.38. Influencia del orden de predicción y la ventana cepstral con preénfasis

Como en el caso de la parametrización LPC, se han obtenido mejores resultados utilizando preénfasis. Concretamente, los mejores resultados se han obtenido utilizando ventana rampa de orden 16.

En la tabla 6.39 se muestran los resultados obtenidos mediante varias combinaciones del cepstrum (c), la energía (E), el delta-cepstrum (Δc) y la delta-energía (ΔE).

Parám. / veloc.	v = 0	v = 70	v =130
c	96.0	94.7	85.0
c, Δc	99.5	96.1	95.5
c, E	91.7	93.6	80.7
c, ΔE	97.2	93.0	87.2
c, Δc, ΔE	98.5	96.0	90.4

Tabla 6.39. Incorporación de parámetros dinámicos de primer orden

Puede observarse en esta tabla que la utilización de la energía logarítmica no proporciona buenos resultados, como en el caso de la parametrización LPC. Además, la

delta-energía tampoco aporta resultados satisfactorios. Así pues, los mejores resultados se obtienen utilizando c, Δc.

En cuanto a los resultados obtenidos utilizando distancia de proyección y modelos con múltiple etiquetado de tipo A, estos se han resumido en la tabla 6.40 junto con los resultados anteriores más importantes. En esta tabla puede observarse que la distancia de proyección no mejora el comportamiento del sistema. Por otro lado, mediante el múltiple etiquetado se obtienen excelentes resultados si se utiliza el delta-cepstrum.

Dist.	Mod.	Nº inf. / veloc.	0	70	130
euc.	DHMM	c	96.0	94.7	85.0
"	"	c, Δc	99.5	96.1	95.5
"	MLHMM	c	97.7	92.1	91.2
"	"	c, Δc	99.5	98.1	95.0
proy.	DHMM	c	95.5	93.3	77.2
"	"	c, Δc	98.2	97.1	93.7

Tabla 6.40. Comparación de técnicas con parametrización LPC y ruido de coche

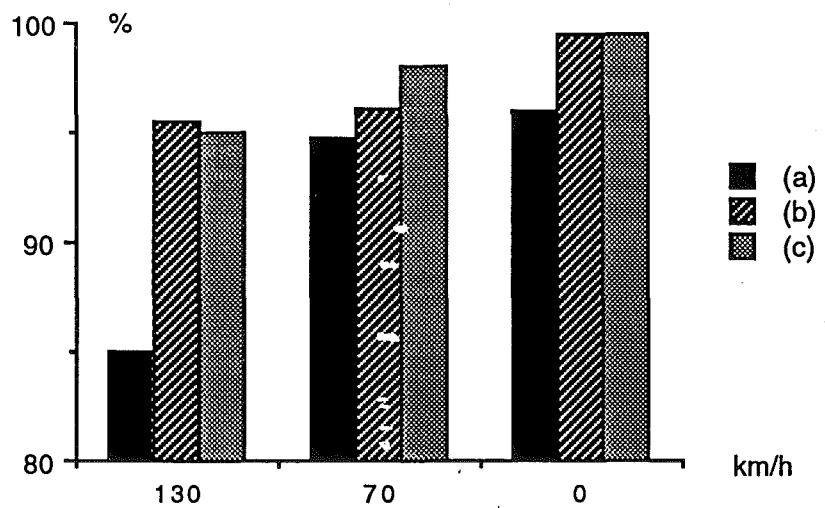

Fig. 6.41. Utilización de múltiple etiquetado con parametrización OSALPC2 y ruido de coche

En la figura 6.41 se muestran, para el caso de ruido de coche y parametrización OSALPC2, las prestaciones del delta-cepstrum y del múltiple etiquetado. La etiqueta (a) corresponde al uso único del cepstrum con cuantificación clásica, la etiqueta (b) corresponde al uso de cepstrum y delta-cepstrum con cuantificación clásica y la etiqueta (c) corresponde al uso de cepstrum y delta-cepstrum con múltiple etiquetado.

6.4.3. COMPARACION

En la tabla 6.41 se muestran los resultados más importantes obtenidos en los dos apartados anteriores con ruido de coche. Los resultados con distancia de proyección no se han incluido porque no proporcionan mejoras claras con respecto a la distancia euclídea.

Puede observarse que la técnica OSALPC2 sin utilizar delta cepstrum obtiene resultados excelentes en condiciones severas de ruido, pero sus prestaciones son algo peores que la técnica de predicción lineal clásica LPC en condiciones poco ruidosas. Sin embargo, utilizando delta-cepstrum la parametrización OSALPC2 es superior a a la LPC en todas las condiciones consideradas. Por otro lado, se observa que el múltiple etiquetado proporciona excelentes resultados combinado con el delta-cepstrum. Los mejores resultados se obtienen utilizando OSALPC2, delta-cepstrum y múltiple etiquetado.

Etiq.	Param.	Mod.	Nº inf. / veloc.	0	70	130
(a)	LPC	DHMM	c	97.5	92.1	79.0
(b)	LPC	MLHMM	c	98.2	94.9	81.7
(c)	OSALPC2	DHMM	c	96.0	94.7	85.0
(d)	OSALPC2	MLHMM	c	97.7	92.1	91.2
(e)	LPC	DHMM	c, Δc, ΔE	98.5	96.6	92.0
(f)	LPC	MLHMM	c, Δc, ΔE	99.2	96.6	94.0
(g)	OSALPC2	DHMM	c, Δc	99.5	96.1	95.5
(h)	OSALPC2	MLHMM	c, Δc	99.5	98.1	95.0

Tabla 6.41 Comparación de técnicas en ruido de coche

En la figura 6.42 se representan gráficamente estos resultados (las etiquetas se corresponden con las de la tabla).

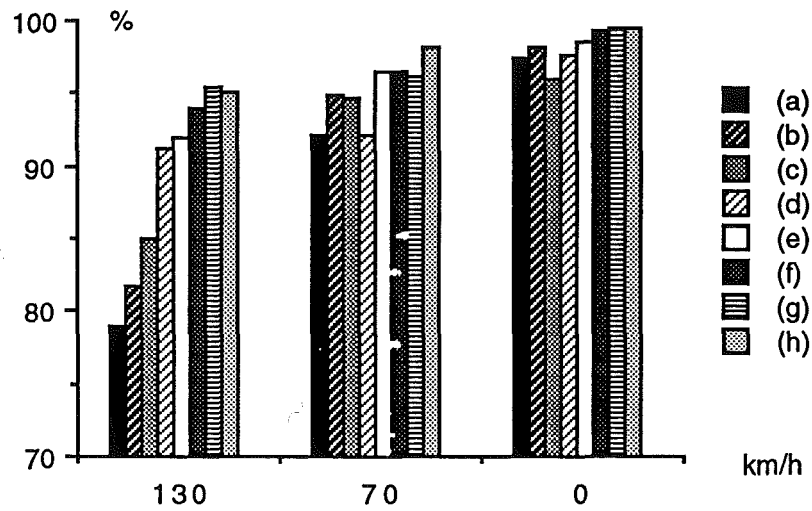

Fig. 6.42. Comparación de técnicas en ruido de coche