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 5

RECONOCIMIENTO DEL HABLA MEDIANTE MODELOS OCULTOS DE MARKOV

Los modelos ocultos de Markov fueron descritos por primera vez por Baum [Bau72]. Poco después, fueron aplicados al reconocimiento automático del habla en CMU [Bak75] e IBM [Bak76] [Jel76]. En los últimos años se han convertido en la aproximación predominante en reconocimiento del habla, superando la técnica de comparación de patrones, debido a la simplicidad de su estructura algorítmica y a sus buenas prestaciones. Por ello será el sistema de reconocimiento utilizado en las pruebas experimentales realizadas en este trabajo.

En este capítulo, en primer lugar se presentará la estructura de los modelos ocultos de Markov. Seguidamente, se presentarán los algoritmos para la evaluación, decodificación y entrenamiento de la aproximación básica discreta; se tratarán aspectos prácticos de implementación, como la inicialización, el escalado y el suavizado de los parámetros; y se estudiará su aplicación al reconocimiento automático del habla. Por último, se extenderán estos resultados a las aproximaciones continua, semicontinua y con múltiple etiquetado, estudiando las posibles ventajas de estas en reconocimiento de habla en entornos adversos.

5.1. MODELOS OCULTOS DE MARKOV. DEFINICION Y TIPOS

Un modelo oculto de Markov es la representación de un proceso estocástico que consta de dos mecanismos interrelacionados: una cadena de Markov de primer orden subyacente, con un número finito de estados, y un conjunto de funciones aleatorias, cada una de las cuales asociada a un estado. En un instante discreto de tiempo se supone que el proceso está en un estado determinado y que genera una observación mediante la función aleatoria asociada. Al instante siguiente, la cadena subyacente de Markov cambia de estado siguiendo su matriz de probabilidades de transición entre estados, produciendo una nueva observación mediante la función aleatoria correspondiente. El observador externo sólo "ve" la salida de las funciones aleatorias asociadas a cada estado, siendo incapaz de observar directamente la secuencia de estados de la cadena de Markov. De ahí el nombre de modelo oculto.

Un modelo oculto de Markov queda, pues, caracterizado por los siguientes elementos:

1) El conjunto finito de N estados de la cadena de Markov de primer orden. Aunque los estados están ocultos, en muchas aplicaciones prácticas éstos tienen un significado físico que es preciso considerar. Se denotará como $S=\{S_i\}_{i=1...N}$ a este conjunto de estados y al estado en el tiempo t como q_t.

2) El conjunto de probabilidades de transición entre estados. Denotando los instantes de tiempo regularmente espaciados asociados a los cambios de estados como t = 1, 2, ... T una descripción probabilística completa de una cadena requeriría, en general, especificaciones sobre el estado actual en el instante t y de todos los estados predecesores. Para el caso especial de una cadena de Markov de primer orden, esta descripción probabilística se trunca en el estado actual y el último predecesor, es decir,

$$P (q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, ...) = P (q_t = S_j \mid q_{t-1} = S_i). \qquad (5.1)$$

Además, se considera que esta última probabilidad es independiente del tiempo (propiedad de homogeneidad temporal), lo cual da lugar a un conjunto de probabilidades de transición entre estados que se denotará con una matriz $A=\{a_{ij}\}_{i,j=1...N}$, donde

$$a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i) \qquad i, j = 1, ... N. \qquad (5.2)$$

Este conjunto de probabilidades determinará la topología del modelo. Así, para un modelo en que cada estado puede ser alcanzado desde cualquier otro en un sólo paso, $a_{ij} > 0$ i,j = 1...N. En general, los modelos pueden tener $a_{ij} = 0$ para una o más parejas de valores (i,j). En cualquier caso deben verificar

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad i = 1,\dots N \qquad (5.3)$$

$$a_{ij} \geq 0 \qquad i, j = 1,\dots N. \qquad (5.4)$$

3) La distribución de probabilidad de estados iniciales, que se denotará como $\Pi = \{\pi_i\}_{i=1\dots N}$, definida de la forma

$$\pi_i = P(q_1 = S_i) \qquad i = 1,\dots N. \qquad (5.5)$$

Como tales probabilidades, también deben verificar

$$\sum_{i=1}^{N} \pi_i = 1 \qquad i = 1,\dots N \qquad (5.6)$$

$$\pi_i \geq 0 \qquad i = 1,\dots N. \qquad (5.7)$$

4) Las probabilidades de generación de observaciones, que caracterizan el proceso asociado a cada uno de los estados del modelo y que se denotarán como $B = \{b_j(O_t)\}_{i=1\dots N}$, con

$$b_j(O_t) = P(O_t \mid q_t = S_j) \qquad j = 1,\dots N, \qquad (5.8)$$

en donde O_t representa el valor de la observación en el instante t, correspondiente a la secuencia de observaciones $O = \{O_t\}_{t=1\dots T}$. Se supone que el proceso de generación de observaciones es independiente del tiempo y que únicamente depende del estado actual del modelo. Hay que hacer notar que en algunas variantes de modelos ocultos de Markov estas probabilidades de observación están asociadas a las transiciones, en lugar de a los estados, caso que no se considerará en este trabajo.

De esta forma el modelo HMM queda definido por la especificación de los conjuntos Π, A y B, que implícitamente fijan el valor de N. Por ello, se suele utilizar la notación compacta

$$\lambda = (\Pi, A , B) \qquad (5.9)$$

para referirse a un determinado modelo λ.

Una representación esquemática de un modelo oculto de Markov de tres estados ergódico (con ningún elemento de la matriz de transiciones nulo) puede verse en la figura 5.1.

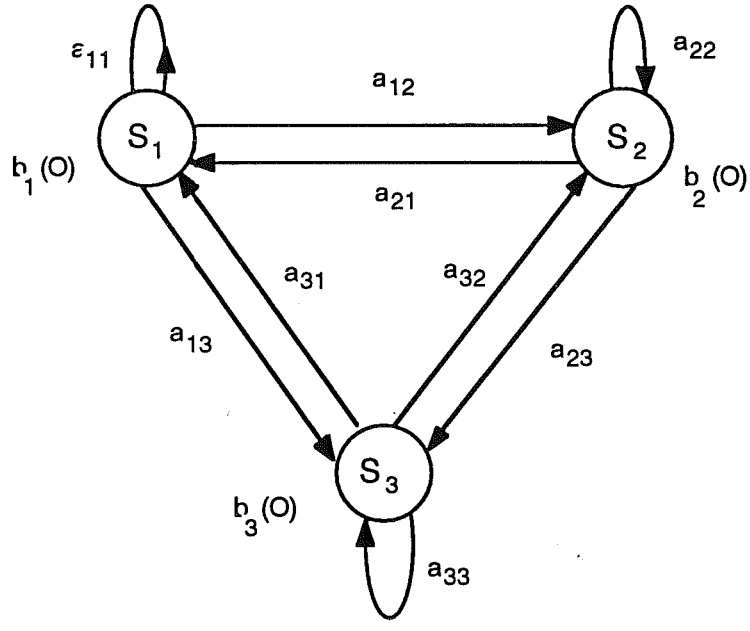

Fig. 5.1. Representación de un Modelo Oculto de Markov

La naturaleza de las probabilidades de generación de observaciones de cada estado $b_j(O_t)$ es la diferencia fundamental entre los distintos tipos de modelos. En los llamados modelos discretos (DHMM), estas probabilidades están representadas a través de distribuciones de probabilidad discretas, ya que las observaciones O_t toman valores dentro de un conjunto discreto y finito de símbolos llamado alfabeto $V = \{v_k\}_{k=1...M}$,

siendo M el tamaño del alfabeto. Las probabilidades de observación forman, pues, un conjunto que se denota como una matriz $B = \{b_j(k)\}_{j=1...N;\ k=1...M}$, donde

$$b_j(k) = P(\ v_k \text{ en } t \mid q_t = S_j\ ) \qquad j = 1,... N \qquad k = 1,... M. \qquad (5.10)$$

Por ser probabilidades estos parámetros deben verificar

$$\sum_{j=1}^{N} b_j(k) = 1 \qquad k = 1,... M \qquad (5.11)$$

$$b_j(k) \geq 0 \qquad j = 1, ... N \qquad k = 1,... M \qquad (5.12)$$

En el caso de la aplicación al reconocimiento del habla, los vectores espectrales de las tramas de voz son cuantificados vectorialmente y estos símbolos v_k se corresponden con las etiquetas de las palabras-código producto de dicha cuantificación vectorial.

En los modelos continuos (CHMM), las probabilidades de observación están representadas a través de funciones de densidad de probabilidad multivariadas, ya que las observaciones toman valores dentro de un espacio continuo multidimensional. En el caso del reconocimiento del habla, las observaciones consisten simplemente en los vectores espectrales de las tramas de voz sin cuantificación.

Además de estas dos aproximaciones básicas, en esta memoria también se trabajará con situaciones intermedias entre ambas como son los modelos con múltiple etiquetado y los semicontinuos (SCHMM), por haberse probado su buen comportamiento en reconocimiento de habla en entornos adversos. Para su mejor comprensión definiremos en detalle estos modelos después de haber revisado las aproximaciones básicas.

Es importante destacar que todos los tipos de modelos que se describirán en esta memoria y que han sido objeto de trabajo son clásicos, en el sentido de que están completamente caracterizados por los elementos anteriormente descritos. No se tratarán temas tales como el modelado temporal de la permanencia en los estados [Fer80] [Rab85a] [Lev86], el modelado paramétrico de las transiciones [Tak89] [Cha90] o la consideración de la correlación entre tramas [Ken90].

Sea cual sea la aproximación utilizada, el número de estados y transiciones permitidas entre ellos, así como las ligaduras entre estados y arcos y la posible

existencia de estados sin observaciones, son elegidos en general por el diseñador del sistema. No se entrará tampoco en esta memoria en el aprendizaje automático de la estructura de los modelos [Cas90].

Una vez definido el modelo para un determinado proceso, surgen tres problemas básicos de interés que deben resolverse de cara a posibles aplicaciones prácticas:

Problema de evaluación: Dada una secuencia de observaciones $O = \{O_t\}_{t=1...T}$ y un modelo $\lambda = (\Pi, A, B)$, cómo evaluar eficientemente $P(O|\lambda)$, la probabilidad de la secuencia de observación dado el modelo. Esta probabilidad se puede utilizar para clasificar las secuencias de observación, punto básico en la aplicación al reconocimiento.

Problema de decodificación: Dada una secuencia de observaciones $O = \{O_t\}_{t=1...T}$ y un modelo $\lambda = (\Pi, A, B)$, cómo elegir la correspondiente secuencia de estados $Q = \{q_t\}_{t=1...T}$ que es óptima en algún sentido, que mejor "explica" las observaciones. Su solución permite obtener información sobre el proceso oculto, por ejemplo, el significado de los estados del modelo. También puede utilizarse, como se verá, para obtener una aproximación eficiente al problema de evaluación.

Problema de entrenamiento: Dada una secuencia de observaciones $O = \{O_t\}_{t=1...T}$, cómo ajustar los parámetros del modelo $\lambda = (\Pi, A, B)$ de forma que se maximice $P(O|\lambda)$, la probabilidad de generación de dicha secuencia por el modelo. Su solución permite desarrollar un método para obtener los parámetros de un modelo en base a secuencias de observaciones que se pretenden modelar.

En el siguiente apartado, se describirán las soluciones a estos tres problemas para el caso de modelos discretos y su aplicación al reconocimiento automático del habla. Más tarde, se extenderán los resultados a otros tipos de modelos.

5.2. MODELOS OCULTOS DE MARKOV DISCRETOS

Como ya se ha indicado, en los modelos ocultos de Markov discretos (DHMM) las observaciones consisten en símbolos pertenecientes a un alfabeto discreto y finito y, por tanto, las probabilidades de observación forman un conjunto finito.

Para este caso sencillo, se formularán a continuación las soluciones a los problemas de evaluación, codificación y entrenamiento. Seguidamente, se estudiará su

aplicación al reconocimiento automático del habla, incluyendo algunos aspectos prácticos de implementación, como la necesidad de la cuantificación vectorial o los problemas de inicialización, escalado y suavizado de los parámetros.

5.2.1. SOLUCION A LOS TRES PROBLEMAS BASICOS

Para exponer la solución a los tres problemas básicos en el caso de modelos ocultos de Markov discretos se utilizará la notación introducida en el apartado 5.1.

5.2.1.1. EVALUACION

Deseamos calcular la probabilidad de que una secuencia de observaciones $O = \{O_t\}_{t=1...T}$ dado un modelo λ, es decir, $P(O|\lambda)$.

La forma más directa de estimar esta probabilidad consiste en enumerar cada secuencia posible de estados de longitud T. Para una secuencia de estados dada $Q = \{q_t\}_{t=1...T}$, la probabilidad de la secuencia de observaciones O es

$$P(O \mid Q, \lambda) = \prod_{t=1}^{T} P(O_t|q_t,\lambda), \qquad (5.13)$$

donde se ha supuesto independencia entre las observaciones. En términos de los parámetros del modelo, esta probabilidad puede escribirse de la forma

$$P(O \mid Q, \lambda) = b_{q1}(O_1)\, b_{q2}(O_2) \ldots b_{qT}(O_T). \qquad (5.14)$$

Por otro lado, la probabilidad correspondiente a la secuencia de estados Q puede escribirse como

$$P(Q \mid \lambda) = \pi_{q1}\, a_{q1q2}\, a_{q2q3} \ldots a_{qT-1qT}. \qquad (5.15)$$

La probabilidad del suceso conjunto de la secuencia de observaciones y estados es simplemente el producto de ambos términos

$$P(O, Q \mid \lambda) = P(O \mid Q, \lambda)\; P(Q \mid \lambda). \qquad (5.16)$$

Finalmente, la probabilidad de generación de observaciones dado el modelo puede obtenerse sumando esta probabilidad conjunta sobre todas la secuencias de estados posibles

$$P(O \mid \lambda) = \sum_{\forall Q} P(O|Q,\lambda)\, P(Q|\lambda) \qquad (5.17)$$

$$= \sum_{q_1 q_2 \ldots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \ldots a_{q_{T-1} q_T} b_{q_T}(O_T). \qquad (5.18)$$

Esta expresión conlleva del orden de $2TN^T$ cálculos, lo cual la hace inaceptable incluso para valores moderados de N y T. Afortunadamente, existe un algoritmo recursivo que permite obtener esta probabilidad de forma eficiente, el algoritmo *Forward-Backward* [Bau67] que se describirá a continuación. Aunque la parte *forward* del algoritmo es suficiente para resolver el problema de la evaluación, se presentará también en este apartado la parte *backward*, ya que se utilizará en la solución del problema de entrenamiento.

Evaluación forward

Se define la variable *forward* como

$$\alpha_t(i) = P(O_1\ O_2 \ldots O_t,\ q_t = S_i \mid \lambda), \qquad (5.19)$$

es decir, la probabilidad de generar la secuencia parcial de observaciones, $O_1\ O_2 \ldots O_t$ de manera que el modelo queda en el estado S_i en el instante t, dado el modelo. Es fácil ver que para $\alpha_t(i)$ puede establecerse la siguiente recursión temporal

1) Inicialización. Se calcula $\alpha_1(i)$ como la probabilidad conjunta de generar la primera observación O_1 y terminar en el estado S_i, para cada uno de los N estados.

$$\alpha_1(i) = \pi_i\, b_i(O_1) \qquad i = 1, \ldots N \qquad (5.20)$$

2) Inducción. Se calcula $\alpha_{t+1}(j)$, para t = 1... T-1 y j = 1... N, multiplicando la probabilidad de generación de la observación O_{t+1} en el estado S_j, $b_j(O_{t+1})$, por la suma de las probabilidades de generar la secuencia parcial de las t observaciones previas finalizando en cada estado S_i, $\alpha_t(i)$, multiplicadas por las probabilidades de transición entre este estado y S_j, a_{ij}. Este proceso puede verse esquematizado en la Fig. 5.2.

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}) \qquad t = 1,\ldots T\text{-}1 \quad j = 1,\ldots N, \tag{5.21}$$

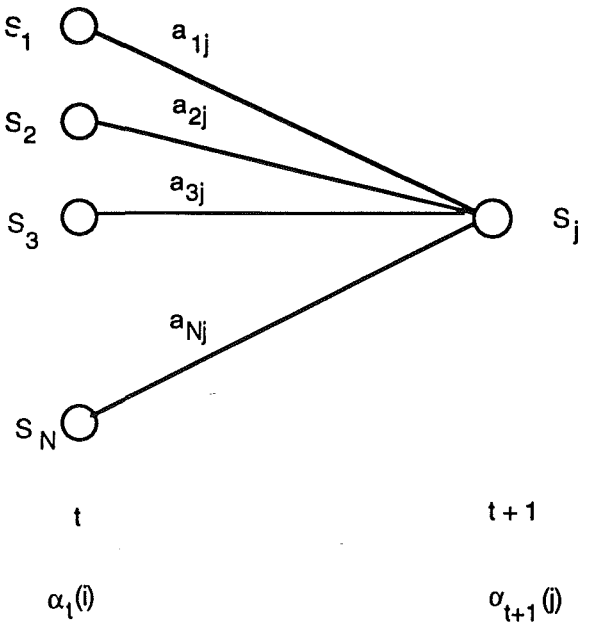

Fig. 5.2. Cálculo de la variable forward

3) Terminación. La suma de las variables forward terminales nos proporciona precisamente P(O|λ), ya que la última observación puede producirse en cualquiera de los estados.

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{5.22}$$

Los cálculos de este algoritmo pueden realizarse eficientemente si se considera una celosía de observaciones y estados por la que se avanza y se van considerando las probabilidades de observación y transición, como es muestra en la Fig. 5.3.

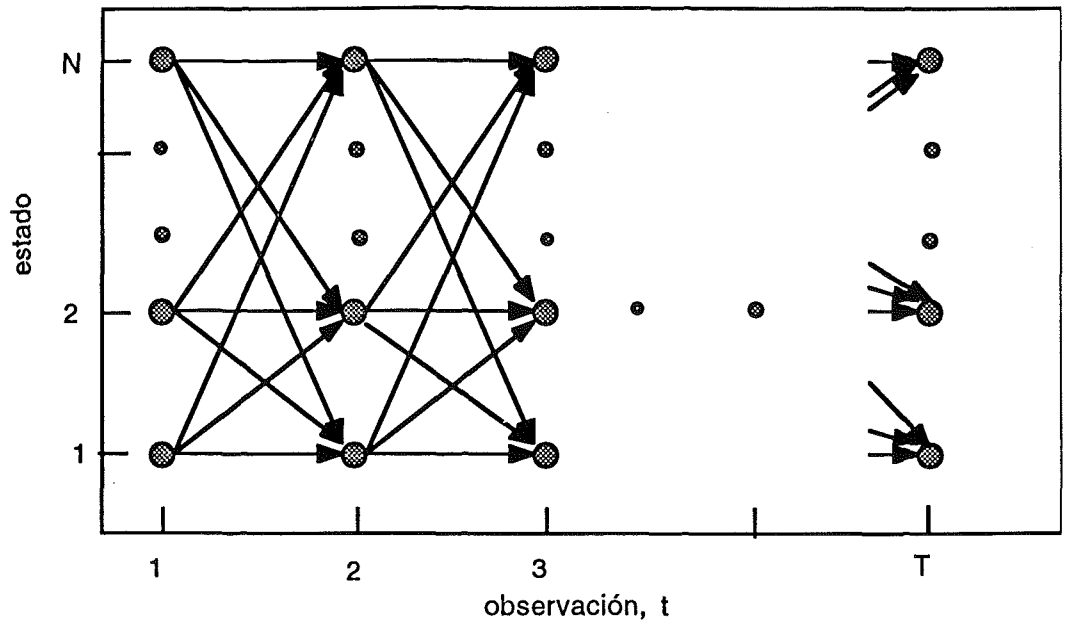

Fig. 5.3. Celosía de cálculo del algoritmo Forward

La complejidad de este algoritmo es del orden de N^2T cálculos frente a los $2TN^T$ de la evaluación directa, lo cual lo hace aceptable para valores moderados de N y T.

Evaluación backward

Se define la variable *backward* como

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots O_T \mid q_t = S_i, \lambda), \qquad (5.23)$$

es decir, la probabilidad de la secuencia parcial de observaciones desde t + 1 hasta el final, dados el estado S_i en el instante t y el modelo. También puede realizarse el cálculo de $\beta_t(i)$ recursivamente como sigue:

1) Inicialización. Arbitrariamente se define

$$\beta_T(i) = 1 \qquad i = 1 \ldots N. \qquad (5.24)$$

2) Inducción. Se calcula $\beta_t(i)$, para t = T-1, T-2, ... 1 e i = 1,... N, de forma análoga al cálculo recursivo de las α's, teniendo en cuenta que del estado S_i puede pasarse a cualquiera de los N estados. Este proceso está ilustrado en la Fig. 5.4.

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \qquad t = T-1, T-2, \ldots 1 \qquad i = 1, \ldots N \tag{5.25}$$

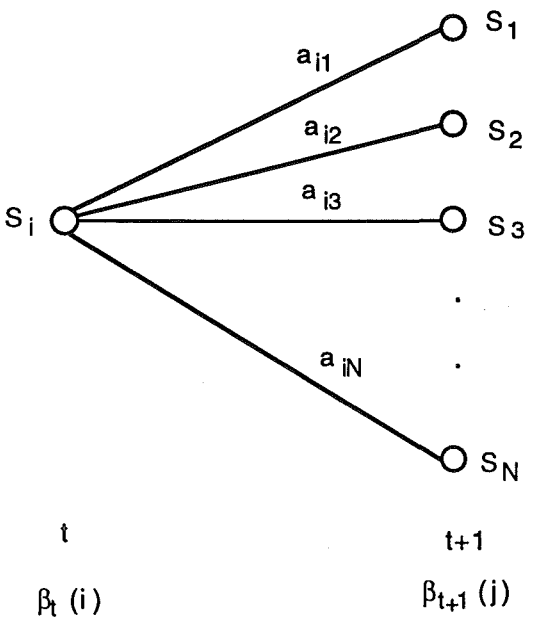

Fig. 5.4. Cálculo de la variable backward

3) Terminación. Considerando que la primera observación puede producirse en cualquier estado,

$$P(O \mid \lambda) = \sum_{i=1}^{N} \pi_i b_i(O_1) \beta_1(i). \tag{5.26}$$

El coste computacional es del mismo orden que el de la evaluación forward y también pueden realizarse los cálculos en una estructura en celosía, análoga a la de la Fig. 5.3.

5.2.1.2. DECODIFICACION

A diferencia del problema de evaluación, no existe una solución única al problema de la obtención de la secuencia óptima de estados dada una secuencia de observaciones y el modelo. Depende del criterio con que se defina esta secuencia óptima.

Así, un posible criterio es el de extraer la secuencia de estados que verifica que las probabilidades de cada uno de los estados que la componen es máxima. Este criterio, aunque conduce a una solución sencilla, presenta varios inconvenientes. El primero de estos es que, al realizar una optimización local de estados, es posible obtener una secuencia de estados imposible para el modelo si este contiene transiciones prohibidas. Además, no está garantizado que la secuencia de estados obtenida sea la de máxima probabilidad de generación de la secuencia de símbolos.

Por estos motivos, se suele utilizar el criterio de seleccionar la secuencia de estados para la que la probabilidad de generación condicionada es máxima:

$$Q^* = \underset{Q}{\operatorname{argmax}} \{ P (Q \mid O, \lambda) \} = \underset{Q}{\operatorname{argmax}} \{ P (Q, O \mid \lambda) \}. \qquad (5.27)$$

De nuevo, el cálculo directo de (5.27) presenta una complejidad que la hace inaplicable incluso para valores razonables de N y T. En su lugar, se utiliza un algoritmo recursivo análogo al *Forward-Backward*, basado en técnicas de programación dinámica, denominado algoritmo de Viterbi.

Algoritmo de Viterbi

Para encontrar la secuencia de estados $Q = \{q_t\}_{t=1...T}$, dada una secuencia de observaciones $O = \{O_t\}_{t=1...T}$, se define la variable

$$\delta_t (i) = \max_{q_1, q_2, \ldots, q_t} \{ P (q_1 q_2 \ldots q_t=i, O_1 O_2 \ldots O_t \mid \lambda) \}, \qquad (5.28)$$

es decir, la probabilidad máxima de generación de las primeras t observaciones sobre cualquier secuencia de estados cuyo estado final sea el S_i, dado el modelo. Fácilmente, se puede demostrar que esta variable verifica una recursión de la forma

$$\delta_{t+1}(j) = \max_i \{ \delta_t(i)\ a_{ij} \}\ b_j(O_{t+1}). \qquad (5.29)$$

Para recuperar la secuencia de estados de probabilidad máxima es necesario almacenar los valores del argumento que maximizan (5.29), para cada t y j. Para ello se utiliza la matriz $\psi_t(j)$.

El algoritmo de Viterbi consta, pues, de los siguientes pasos:

1) Inicialización:

$$\delta_t(i) = \pi_i\ b_i(O_1) \qquad i = 1,\ldots N \qquad (5.30)$$

$$\psi_t(i) = 0. \qquad (5.31)$$

2) Recursión:

$$\delta_t(j) = \max_{i=1\ldots N} \{ \delta_{t-1}(i)\ a_{ij}) \}\ b_j(O_t) \qquad t = 2,\ldots T \quad j = 1,\ldots N \qquad (5.32)$$

$$\psi_t(j) = \operatorname{argmax}_{i=1\ldots N} \{ \delta_{t-1}(i)\ a_{ij}) \} \qquad t = 2,\ldots T \quad j = 1,\ldots N. \qquad (5.33)$$

3) Terminación:

$$P^* = \max_{i=1\ldots N} \{ \delta_T(i) \} \qquad (5.34)$$

$$q_T^* = \operatorname{argmax}_{i=1\ldots N} \{ \delta_T(i) \} \qquad (5.35)$$

4) Recursión para obtener la secuencia de estados:

$$q_t^* = \psi_{t+1}(q^*_{t+1}) \qquad t = T\text{-}1, T\text{-}2, \ldots 1 \qquad (5.36)$$

También en este caso la estructura en celosía implementa eficientemente los cálculos (Fig. 5.5). Hay que destacar que en este caso no se consideran todas las transiciones hasta cada estado, sino solamente aquellas que dan lugar a una probabilidad máxima.

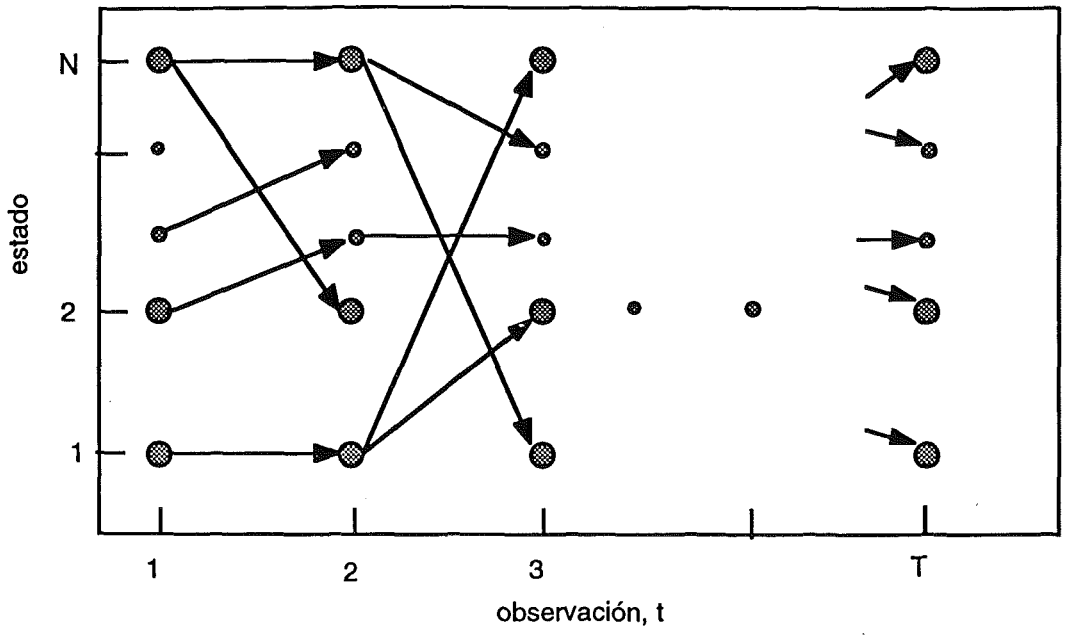

Fig. 5.5. Celosía de cálculo del algoritmo de Viterbi

El algoritmo de Viterbi no se utiliza tan sólo para determinar la secuencia de estados óptima, sino también para determinar la probabilidad de una secuencia de observaciones por el camino óptimo, ya que aunque es distinta de la obtenida por el método *Forward* proporciona una aproximación a la misma y es más rápida de calcular.

5.2.1.3. ENTRENAMIENTO

El tercer problema relacionado con el modelado HMM es el de ajustar los parámetros del modelo $\lambda = (\Pi, A, B)$ para maximizar la probabilidad de generación de una secuencia de observaciones $O = \{O_t\}_{t=1...T}$, dado el modelo. Este problema es con mucho el más difícil. De hecho, dada una secuencia finita de observaciones no es posible estimar de forma óptima los parámetros del modelo. Sin embargo, se pueden elegir los parámetros del modelo de forma que se maximice localmente la probabilidad $P(O|\lambda)$ mediante un procedimiento iterativo como el de Baum-Welch [Bau72] (o equivalentemente el método EM [Dem77]) o técnicas de gradiente [Lev83].

A continuación, se expondrá el método de reestimación de Baum-Welch. Para ello, conviene definir $\xi_t(i,j)$, la probabilidad de que el modelo se encuentre en el estado S_i en el instante t y se produzca una transición de forma que en el instante t+1 el estado sea el S_j dada la secuencia de observaciones y el modelo, es decir,

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid O, \lambda) = \frac{P(q_t=S_i, q_{t+1}=S_j, O|\lambda)}{P(O|\lambda)}. \tag{5.37}$$

Este valor puede expresarse en función del las probabilidades *forward* y *backward*, en la forma (ver Fig. 5.6)

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}. \tag{5.38}$$

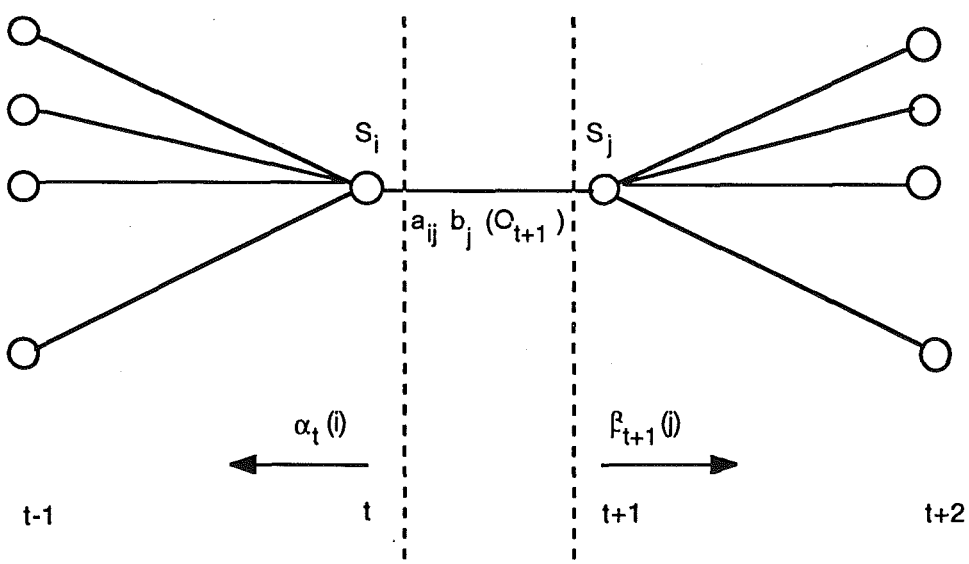

Fig. 5.6. Cálculo de $\xi_t(i,j)$

También es conveniente definir la variable $\gamma_t(i)$ como la probabilidad de estar en el estado S_i en el instante t, dada la secuencia de observaciones y el modelo, es decir,

$$\gamma_t(i) = P(q_t = S_i \mid O, \lambda) = \frac{P(q_t=S_i, O|\lambda)}{P(O|\lambda)}, \tag{5.39}$$

que también puede expresarse en función de las probabilidades *forward* y *backward*:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}. \qquad (5.40)$$

Sumando $\gamma_t(i)$ para t = 1,... T, se obtiene el número esperado (en el tiempo) de veces que el modelo se encuentra en el estado S_i; y sumando para t = 1,... T-1, se obtiene el número esperado de veces que el modelo realiza un transición desde el estado S_i. Además, sumando $\gamma_t(i)$ para t = 1,... T con la restricción de que el símbolo observado sea v_k, se obtiene el número esperado de veces que el modelo genera el símbolo v_k en el estado S_i. Por último, sumando $\xi_t(i,j)$ para t = 1,... T-1, se obtiene el número esperado de veces que se produce una transición entre los estados S_i y S_j.

Con estas definiciones y estos resultados, se pueden establecer las siguientes fórmulas de reestimación de los parámetros, que dan lugar a nuevos parámetros que verifican automáticamente las restricciones estocásticas:

π_i' = nº esperado de veces en el estado S_i en t = 1 =

$$= \gamma_1(i) = \frac{\alpha_1(i)\beta_1(i)}{\sum_{i=1}^{N} \alpha_1(i)\beta_1(i)} \qquad i = 1,... N \qquad (5.41)$$

$$a_{ij}' = \frac{\text{nº esperado de transiciones desde el estado } S_i \text{ al estado } S_j}{\text{nº esperado de transiciones desde el estado } S_i} =$$

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} = \frac{\sum_{t=1}^{T-1} \alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} \qquad i, j = 1,... N$$

(5.42)

$$b_j'(k) = \frac{\text{nº esperado de veces en el estado } S_j \text{ y observando el símbolo } v_k}{\text{nº esperado de veces en el estado } S_j} =$$

$$= \frac{\sum_{t=1,O_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} = \frac{\sum_{t=1,O_t=v_k}^{T} \alpha_t(j)\beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)} \qquad j = 1,\ldots N \quad k = 1,\ldots M.$$

(5.43)

Se demuestra que si a partir de un modelo $\lambda = (\Pi, A, B)$, utilizando estas fórmulas de reestimación, se obtiene un nuevo modelo $\lambda' = (\Pi', A', B')$, la probabilidad de generación de la secuencia de observaciones dado el modelo λ' es siempre mayor que la obtenida para el modelo inicial λ, excepto cuando se alcanza un valor crítico de la función probabilidad, en cuyo caso las dos probabilidades coinciden [Bau68]. Esta prueba garantiza una convergencia uniforme de este método de reestimación hacia este punto límite, llamado Estimación de Máxima Probabilidad.

Hay que destacar que este algoritmo sólo conduce a máximos locales y que en muchos problemas de interés la superficie de optimización es muy compleja y tiene muchos máximos locales. De ahí que en muchos casos sea importante el problema de la inicialización de los parámetros.

Las ecuaciones de reestimación (5.41)-(5.43) pueden obtenerse también de forma directa maximizando la función auxiliar de Baum

$$Q(\lambda, \lambda') = \sum_{Q} P(Q|O,\lambda) \log[P(O,Q|\lambda')] \qquad (5.44)$$

y también pueden interpretarse como una implementación del algoritmo estadístico EM (*Expectation-Modification*) [Dem77]. Alternativamente, se pueden obtener también las mismas relaciones maximizando directamente la función probabilidad $P(O|\lambda)$, sujeta a las restricciones estocásticas de los parámetros, mediante un método tradicional como el de los multiplicadores de Lagrange.

Por último, indicar que debido a que todo el problema puede plantearse como un problema de optimización pueden aplicarse las técnicas clásicas de gradiente. Esta aproximación permite utilizar otros criterios de optimización como el MMI (o de Máxima Información Mutua) [Mer88], que se basa en maximizar la información mutua promedio entre el conjunto de secuencias de entrenamiento y el conjunto de modelos a diseñar, cuando se desea diseñar conjuntamente una serie de modelos que se utilizarán con propósitos discriminativos.

También se han utilizado para la estimación de los parámetros de los modelos de Markov algoritmos menos formalizados como el llamado entrenamiento correctivo (ver capítulo 2). No obstante, en este trabajo sólo se ha considerado el entrenamiento de los modelos mediante el algoritmo clásico de Baum-Welch.

5.2.2. IMPLEMENTACION DE LOS MODELOS

En el apartado anterior se ha tratado la teoría básica relacionada con los modelos ocultos de Markov discretos (DHMM). En este apartado se discutirán cuestiones relacionadas con la implementación de dicho modelado como los problemas derivados del rango de valores de valores de las probabilidades de observación, la elección de los valores iniciales de los parámetros, la estimación de los parámetros del modelo con múltiples secuencias de observaciones y los efectos de la existencia de un número finito de observaciones.

5.2.2.1. ESCALADO DINAMICO Y COMPRESION LOGARITMICA

Considerando que los parámetros π_i, a_{ij} y $b_j(k)$ tienen valores inferiores a la unidad (frecuentemente muy inferiores), las definiciones de las variables $\alpha_t(i)$ y $\beta_t(i)$ vistas en el apartado 5.2.1 dan lugar a valores que decaen exponencialmente a cero con el tiempo. Para un valor de T moderadamente alto (p.e. 100), el rango dinámico de estas variables excede la precisión de cualquier máquina. Esta situación no es particular de los modelos discretos sino que es general a todo tipo de modelos ocultos de Markov. Se han propuesto al menos dos soluciones a este problema, el escalado dinámico y la compresión logarítmica de las probabilidades.

En los modelos ocultos de Markov discretos usados en este trabajo se ha realizado un escalado dinámico de las variables en el algoritmo de reestimación de Baum-Welch, lo cual proporciona una solución exacta al problema. Para ello, se multiplican las variables *forward y backward* por unos coeficientes de escalado que son independientes de i (es decir, sólo dependen de t), de manera que el valor escalado de estas variables se mantiene dentro del rango dinámico de la máquina para t = 1,... T y al finalizar los cálculos los coeficientes de escalado se cancelan exactamente. Las expresiones exactas de estos coeficientes de escalado pueden encontrarse en [Rab89]. En la realización del algoritmo de Viterbi se ha utilizado la compresión logarítmica de las probabilidades, que en este caso proporciona una solución exacta al problema sin necesidad de escalado dinámico.

5.2.2.2. INICIALIZACION DE LOS PARAMETROS

El algoritmo de Baum-Welch para la reestimación iterativa de los parámetros del modelo sólo garantiza la obtención de un máximo local de la función probabilidad de generación del modelo. Una cuestión importante es, pues, cómo elegir estimaciones iniciales de los parámetros de manera que el máximo local se corresponda con el máximo global de la función probabilidad.

No existe una solución exacta a esta cuestión. La experiencia demuestra que la elección aleatoria (sujeta a las restricciones estocásticas y sin permitir valores nulos para aquellos parámetros que no desean fijarse a cero) o uniforme para los valores de las probabilidades iniciales de los estados Π y para las probabilidades de transición A resulta adecuada para obtener parámetros útiles en casi todos los casos. Sin embargo, en el caso de las probabilidades de observación B se ha observado que estimaciones iniciales buenas resultan de gran ayuda en el caso de los modelos discretos y son esenciales en los modelos continuos, semicontinuos y de múltiple etiquetado.

En los modelos ocultos de Markov discretos usados en este trabajo se ha usado inicialización aleatoria de las matrices A y B. En cuanto a Π, se ha forzado por razones físicas que la secuencia comience en el estado 1; por tanto, $\pi_1 = 1$ y $\pi_i = 0$ para $i \neq 1$.

5.2.2.3. MULTIPLES SECUENCIAS DE OBSERVACIONES

En muchas situaciones no se dispone de una secuencia de observaciones de la suficiente duración como para poder estimar adecuadamente los parámetros del modelo. También suele ocurrir que existe una gran variabilidad entre las secuencias de observaciones que se desean hacer corresponder con el mismo modelo. En cualquiera de estos casos (la pronunciación de una palabra o un fonema es un ejemplo claro de ambas situaciones), la naturaleza del proceso a modelar hace necesario utilizar varias secuencias de observaciones en la estimación de dichos parámetros.

Por tanto, es necesario modificar las fórmulas de estimación (5.41)-(5.43), que están desarrolladas para una secuencia simple de observaciones, para tener en cuenta un conjunto de secuencias. Dado que estas fórmulas están basadas en términos de frecuencias de ocurrencia de determinados sucesos y suponiendo independencia estadística y equiprobabilidad en las diferentes secuencias de observaciones, se puede simplemente acumular los valores de dichas frecuencias para todas las secuencias del

conjunto de entrenamiento, con lo que las fórmulas de reestimación finales pueden expresarse como

$$\pi_i' = \frac{\sum_{l=1}^{L} \alpha_1^{(l)}(i)\beta_1^{(l)}(i)}{\sum_{l=1}^{L}\sum_{i=1}^{N} \alpha_1^{(l)}(i)\beta_1^{(l)}(i)} \qquad i = 1,\dots N \qquad (5.45)$$

$$a_{ij}' = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T_l-1} \alpha_t^{(l)}(i)\, a_{ij} b_j(O_{t+1}^{(l)})\beta_{t+1}^{(l)}(j)}{\sum_{l=1}^{L}\sum_{t=1}^{T_l-1} \alpha_t^{(l)}(i)\beta_t^{(l)}(i)} \qquad i, j = 1,\dots N \qquad (5.46)$$

$$b_j'(k) = \frac{\sum_{l=1}^{L}\sum_{t=1,\, O_t^{(l)}=v_k}^{T_l} \alpha_t^{(l)}(j)\beta_t^{(l)}(j)}{\sum_{l=1}^{L}\sum_{t=1}^{T_l} \alpha_t^{(l)}(j)\beta_t^{(l)}(j)} \qquad j = 1,\dots N \quad k = 1,\dots M, \qquad (5.47)$$

donde el índice l indica que los valores han sido obtenidos para la secuencia de observaciones l, perteneciente al conjunto de L secuencias $O = [O^{(1)}, O^{(2)},\dots O^{(L)}]$, con $O^{(l)} = [O_1^{(l)}, O_2^{(l)},\dots O_{T_l}^{(l)}]$.

5.2.2.4. INSUFICIENCIA DE DATOS DE ENTRENAMIENTO. SUAVIZADO

La secuencia de observaciones usadas en el entrenamiento es necesariamente finita, lo cual da lugar a menudo a un número insuficiente de ocurrencias de los diferentes eventos del modelo para proporcionar buenas estimaciones de los parámetros, especialmente las probabilidades de observación. En el caso de los modelos discretos, aunque los símbolos más frecuentes serán bien entrenados, si un símbolo no aparece nunca en un estado durante el entrenamiento se dará valor nulo a su probabilidad en la distribución correspondiente a dicho estado; y si luego aparece en el reconocimiento esta probabilidad cero puede ser atribuida a todo el modelo.

Se puede combatir este problema aumentando el conjunto de entrenamiento, reduciendo el número de parámetros del modelo o ligando los valores de algunos de los parámetros [Jel80]. Sin embargo, esto no suele ser posible por razones prácticas de memoria y tiempo, en el primer caso, o por razones físicas, que pueden obligar a utilizar un determinado modelo, que no puede ser reducido.

La solución a este problema son las técnicas de suavizado (*smoothing*, en la literatura inglesa), que consisten en un procesado de las distribuciones de probabilidad de observación posterior al entrenamiento.

El método de suavizado más simple es el llamado *floor smoothing*, que consiste en asegurarse que ninguna probabilidad de observación esté por debajo de un determinado umbral δ (en este trabajo se ha utilizado como umbral $\delta = 10^{-3}$), es decir, imponer

$$b_j(k) \geq \delta \qquad j = 1,\ldots N \qquad k = 1,\ldots M. \qquad (5.48)$$

Cuando esta restricción es violada, se realiza la corrección manualmente modificando la probabilidad en cuestión y reescalando el resto de las probabilidades para que se cumplan las restricciones estocásticas. Esta técnica resuelve eficientemente el problema mencionado de la probabilidad nula y es suficiente para obtener modelos razonablemente entrenados, por lo cual su uso es general. Sin embargo, no puede distinguir los símbolos improbables de los imposibles, lo cual crea problemas cuando los modelos no están bien entrenados y muchos símbolos no son observados.

Para combatir este problema, las técnicas de suavizado más elaboradas establecen una relación entre los símbolos de forma cuantitativa, considerando las características y el comportamiento de los mismos, y posteriormente recombinan sus probabilidades de observación teniendo en cuenta las relaciones previamente establecidas. Suponen que si un símbolo tiene una probabilidad alta de aparecer en un estado de un modelo otro símbolo que esté muy relacionado con él no podrá tener una probabilidad mucho más baja.

La forma de estimar cuantitativamente la relación entre símbolos es la que distingue las distintas técnicas de suavizado. En cualquier caso, dicha relación se expresa matemáticamente mediante una probabilidad condicionada $p(v_k|v_l)$, que se interpreta como la probabilidad de que aparezca la observación v_k suponiendo que haya aparecido la v_l.

Las probabilidades de observación suavizadas se obtienen como una combinación lineal de las probabilidades de observación originales utilizando como pesos dichas probabilidades condicionadas

$$b_j(k)^{suav} = \sum_{l=1}^{M} p(v_k|v_l) b_j(l)^{orig} \qquad j = 1,\ldots N \quad k = 1,\ldots M. \qquad (5.49)$$

La nueva probabilidad de observación de un símbolo recibe entonces la influencia de los demás, influencia que será tanto mayor cuanto mayor sea su relación con cada uno de ellos. De este modo, cuanto mayor sea la relación entre dos símbolos más parecidos serán los valores suavizados de sus probabilidades de observación. Se obtiene así un nuevo modelo con unas probabilidades de observaciones de valores menos extremos que los originales (de ahí el nombre de suavizado). Naturalmente, este nuevo modelo no cumple la propiedad de independencia estadística de las observaciones que cumplía el modelo original.

Matricialmente, puede expresarse (5.49) como

$$B^{suav} = B^{orig} T, \qquad (5.50)$$

donde T es la matriz de suavizado, cuyos elementos son las probabilidades condicionadas $p(v_k|v_l)$.

Una cuestión práctica a considerar es que la matriz de observaciones suavizada resultante de aplicar directamente (5.50) no cumple las restricciones estocásticas. Por ello, normalmente se normaliza convenientemente la matriz de suavizado antes realizar el producto matricial.

El método de estimación de la matriz T es el que diferencia las distintas técnicas de suavizado. En este trabajo se han realizado pruebas experimentales en reconocimiento de habla con cinco técnicas distintas de suavizado: Parzen, distancias mutuas, correlaciones, coocurrencias y alineación de secuencias. Seguidamente se describirán estas técnicas. Por su aplicación posterior al reconocimiento del habla se supondrá que los símbolos del modelo de Markov se corresponden con las etiquetas de las palabras-código producto de la cuantificación vectorial de los vectores espectrales de las tramas de voz.

Al hablar de las distintas técnicas de suavizado se suele distinguir entre técnicas de distancia y técnicas de información mutua. De las técnicas mencionadas, las tres primeras son técnicas de distancia y las dos últimas de información mutua. En las técnicas de distancia la relación entre símbolos se establece a priori en función de la semejanza de las palabras-código correspondientes. En las técnicas de información mutua, sin embargo, la relación entre símbolos se establece a posteriori en base a diferentes criterios. Consideración aparte merece la técnica de *floor smoothing*, que no utiliza matriz de suavizado y se utiliza siempre de forma general independientemente del posible uso de otra técnica de suavizado más elaborada.

Seguidamente se describirán estas técnicas, junto con otra técnica que permite la combinación de modelos entrenados convencionalmente con modelos suavizados.

Método de Parzen

En el método de suavizado de Parzen, propuesto por R. Schwartz et al. en [Sch89] y basado en el trabajo de K. Fukunuga [Fuk72], se estima la relación entre símbolos como

$$p\ (v_k \mid v_l\ ) = \exp\left\{ - \left(\frac{d^2}{\sigma^2} \right)^{\alpha} \right\}, \qquad (5.51)$$

donde d es la distancia entre las palabras-código asociadas a los símbolos v_k y v_l, σ^2 es la varianza de esta distancia y α es un parámetro a elegir. Puede observarse que los coeficientes de relación disminuyen cuanto mayor es la distancia entre palabras código.

Naturalmente, se utiliza la misma definición de distancia que se ha usado previamente en la construcción del diccionario del cuantificador vectorial. En cuanto al valor de α, los autores recomiendan el valor 1. En este caso, tenemos una función proporcional a una gaussiana.

Método de distancias mutuas

Este método, propuesto por K. Sugawara [Sug85], consiste en considerar sólo la relación de cada símbolo a los L símbolos cuyas palabras palabras-código sean más cercanas, en términos de la distancia usada en el cuantificador vectorial, y cuantificar esta relación de forma constante para los L símbolos.

Lo más habitual es considerar sólo los 5 símbolos más cercanos y asignar las siguientes relaciones

$$p\,(\,v_k \mid v_k\,) = 0.9 \tag{5.52}$$

$$p\,(\,v_k \mid v_l\,) = 0.02. \tag{5.53}$$

Con estos valores la matriz T de suavizada queda ya normalizada para que B^{suav} cumpla directamente las restricciones estocásticas.

Método de correlaciones

En este caso, la relación entre símbolos se establece directamente a partir de la correlación entre las palabras-código correspondiente, definida como

$$p\,(\,v_k \mid v_l\,) = \frac{< w_k,\, w_l >}{\sqrt{|w_k|\,|w_l|}}, \tag{5.54}$$

donde w_k y w_l son las palabras-código correspondientes a los símbolos v_k y v_l y el símbolo $<\,,\,>$ indica el producto escalar ente vectores.

Método de coocurrencias

El objetivo de este método, propuesto por K.F. Lee [Lee88b], es suavizar las probabilidades promediando la información de todos los modelos de la aplicación, de forma que si en el resto de los modelos dos símbolos presentan una gran semejanza en cuanto a sus probabilidades de observación también en el modelo objeto de suavizado las probabilidades deberán ser parecidas.

Se define la probabilidad de coocurrencia del símbolo v_k dado el símbolo v_l como

$$p\,(v_k|v_l) = \frac{p(v_k,v_l)}{\sum_{m=1}^{M} p(v_m,v_l)} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{N(h)} p(v_k,v_l|S_i,\lambda_h)p(S_i|\lambda_h)p(\lambda_h)}{\sum_{m=1}^{M}\sum_{h=1}^{H}\sum_{i=1}^{N(h)} p(v_m,v_l|S_i,\lambda_h)p(S_i|\lambda_h)p(\lambda_h)}, \tag{5.55}$$

donde H es el número de modelos, N(h) es el número de estados del modelo λ_h, M es el número de símbolos, $p(S_i|\lambda_h)$ es la probabilidad del estado S_i del modelo λ_h, $p(\lambda_h)$ es

la probabilidad del modelo λ_h y $p(v_m,v_l|S_i,\lambda_h)$ es la probabilidad conjunta de v_m y v_l en el estado S_i del modelo λ_h.

Esta probabilidad de coocurrencia puede definirse grosso modo como: "cuando se observa el símbolo v_l, con qué frecuencia se observa el símbolo v_k en contextos similares" [Lee88b] y se usa en el suavizado como medida de relación entre ambos símbolos.

Además, se supone que las probabilidades de los símbolos dentro de un estado de un modelo son independientes. Por tanto, la probabilidad de coocurrencia puede expresarse en función de los parámetros de los modelos como

$$p\ (v_k|v_l) = \frac{\sum_{h=1}^{H}\sum_{i=1}^{N(h)} b_i^{(h)}(k)b_i^{(h)}(l)p(S_i|\lambda_h)p(\lambda_h)}{\sum_{m=1}^{M}\sum_{h=1}^{H}\sum_{i=1}^{N(h)} b_i^{(h)}(m)b_i^{(h)}(l)p(S_i|\lambda_h)p(\lambda_h)}. \qquad (5.56)$$

El problema principal de esta técnica es que requiere un elevado volumen de cálculo. Asimismo, no siempre resulta evidente la estimación de la probabilidad de un modelo $p(\lambda_h)$ o de un estado de un modelo $p(S_i|\lambda_h)$, a no ser que se hayan calculado previamente mientras se realizaba el entrenamiento. Si no se dispone de estos datos pueden aproximarse considerando una distribución uniforme.

Método de alineación de secuencias

Esta técnica, propuesta por K. Sugawara [Sug85], estima la relación entre símbolos basándose en la frecuencia con que las palabras-código correspondientes quedan emparejadas al realizar un alineamiento mediante programación dinámica entre diferentes realizaciones del proceso a modelar, por ejemplo, una palabra.

Combinación de modelos

Por último, el efecto de la insuficiencia de datos de entrenamiento puede combatirse interpolando un modelo λ_i en base a un modelo estimado convencionalmente λ y a otro modelo suavizado con cualquiera de las técnicas mencionadas λ_s. El proceso de interpolación está controlado por un parámetro ε en la forma

$$\lambda_i = \varepsilon\,\lambda + (1 - \varepsilon)\,\lambda_s \qquad \varepsilon \in [0,1]. \qquad (5.57)$$

El parámetro de interpolación ε puede obtenerse mediante un método de prueba y error, o bien de forma automática mediante un algoritmo de reestimación *Forward-Backward*, conocido como *Deleted Interpolation* [Jel80].

5.2.3. APLICACION AL RECONOCIMIENTO DEL HABLA

Como ya se ha comentado, en los últimos años los modelos ocultos de Markov se han convertido en la aproximación predominante en reconocimiento del habla, superando la técnica de comparación de patrones, debido a la simplicidad de su estructura algorítmica y a sus buenas prestaciones.

En el apartado siguiente 5.3.2.1, tras una breve discusión sobre el modelado HMM de la señal de voz, se describirá la estructura general de un sistema de reconocimiento del habla basado en modelos ocultos de Markov discretos. En particular, se abordará únicamente el problema de reconocimiento de palabras aisladas, pues las pruebas experimentales realizadas en este trabajo son de este tipo. El problema concreto de la discretización del espacio de características de la señal de voz se tratará por separado en el apartado 5.3.2.2. Por último, el apartado 5.3.2.3 tratará las diversas posibilidades de incorporar varias informaciones, en lugar de utilizar únicamente la información espectral instantánea, a un sistema de reconocimiento como el descrito.

5.2.3.1. DESCRIPCION GENERAL DEL SISTEMA DE RECONOCIMIENTO

En el capítulo 3 de esta memoria se ha visto que, debido a la inercia inherente a los órganos articulatorios, es posible suponer que las características de la señal no varían apreciablemente en un intervalo suficientemente corto de tiempo (del orden de 20 ms) y, por tanto, es posible realizar un análisis espectral cuasi-estacionario sobre segmentos de señal de esta duración temporal. La evolución temporal de las características espectrales se obtiene repitiendo el análisis sobre segmentos consecutivos de la señal, que suelen tomarse con un cierto solapamiento temporal. De esta forma, a partir de una señal de voz se obtiene una secuencia de espectros, que suele denominarse espectrograma.

En los sistemas de reconocimiento del habla mediante técnicas de comparación de patrones se aborda el proceso de reconocimiento sin realizar un modelado de la evolución temporal de esta secuencia de espectros. Los patrones de referencia y de test

consisten simplemente en secuencias de espectros y el proceso de comparación se limita a calcular la distancia acumulada entre dichos patrones a lo largo del camino óptimo dado por el algoritmo de programación dinámica (ver capítulo 2). Salvo en el caso de aplicar técnicas de agrupamiento para obtener los patrones de referencia a partir de varias pronunciaciones, la variabilidad de la señal de voz sólo es tenida en cuenta en el alineamiento temporal no lineal de los patrones que realiza el citado algoritmo de programación dinámica.

En los sistemas de reconocimiento del habla basados en los modelos ocultos de Markov, se modela la evolución temporal de la secuencia de espectros obtenida de la señal de voz mediante un HMM con el fin de contemplar estocásticamente las diversas fuentes de variabilidad de la señal. Este modelado consiste en la asociación de los estados del HMM a los diferentes tramos de la señal, de forma que las probabilidades de generación de observaciones modelan la variabilidad estadística de las características espectrales de cada tramo, mientras que las probabilidades de transición modelan su secuenciamiento y duración.

Se suele motivar este tipo concreto de modelado haciendo corresponder los estados del modelo con diferentes configuraciones de los órganos del tracto vocal. Sin embargo, el modelado HMM no requiere esta correspondencia y no suele hacerse ningún intento en la práctica para establecerla. Así, por ejemplo, en el reconocimiento de palabras aisladas el número de estados del modelo de cada palabra puede no corresponderse con el contenido fonético esperado de la palabra.

Debido a la correspondencia entre estados del modelo y tramos de la señal de voz, en los sistemas de reconocimiento del habla la topología usualmente elegida para los HMM es la denominada izquierda-derecha, en la que las probabilidades de transición son tales que

$$a_{ij} = 0 \qquad j < i, \qquad (j - i) > \Delta, \qquad (5.58)$$

es decir, sólo están permitidas las transiciones hacia adelante y el número de estados que pueden ser "saltados" por el modelo durante su evolución temporal está limitado por un parámetro Δ. Esta topología, que se muestra esquemáticamente en la figura 5.7, fue inicialmente propuesta por Bakis [Bak76] y se ha aplicado para el modelado de unidades constitutivas del habla tales como palabras [Rab85a] [Gup87], fonemas [Cho86] [Lee88a] o semisílabas [Mar90]. También será utilizada en las pruebas

experimentales realizadas en este trabajo. El número de estados del modelo es elegido por el diseñador.

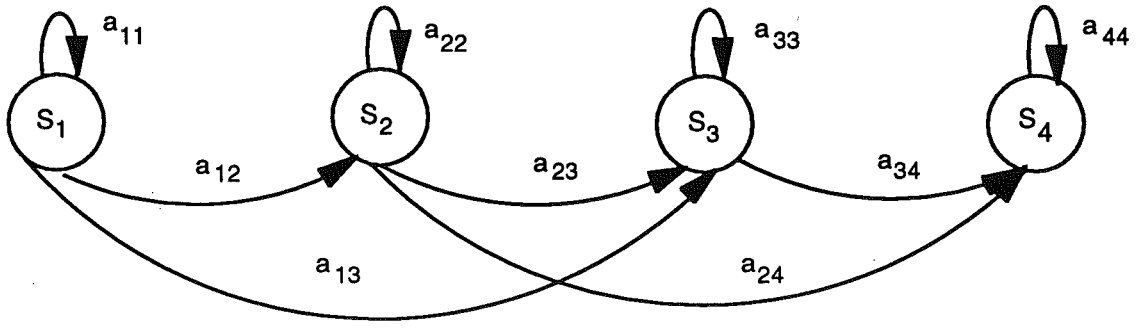

Fig. 5.7. Modelo oculto de Markov izquierda-derecha

Como ya se ha mencionado, en esta memoria se considerará únicamente el reconocimiento de palabras aisladas, pues las pruebas experimentales realizadas en este trabajo son de este tipo. El problema se reducirá, por tanto, a la construcción de un modelo, como mínimo, para cada palabra del diccionario y la selección del modelo más probable dada una palabra incógnita.

En el caso de los modelos discretos, tanto en la fase de construcción de los modelos como en la de selección del modelo más probable, es necesario obtener una secuencia de observaciones discretas a partir de la señal de voz. Esta tarea se realiza en dos etapas:

La primera etapa consiste en la extracción de las características de la señal de la voz. Para ello, normalmente se aplican métodos paramétricos de análisis espectral (ver capítulo 3) sobre segmentos consecutivos de señal, obteniéndose así una secuencia de vectores de parámetros espectrales.

La segunda etapa consiste en la discretización de este espacio vectorial de características espectrales para pasar de una secuencia de vectores a una secuencia de símbolos correspondiente a un alfabeto finito. Esto se consigue mediante la cuantificación vectorial de los vectores de características espectrales y la asociación de

las palabras-código del diccionario del cuantificador con los símbolos del alfabeto del modelo. Este proceso de discretización será descrito en el apartado 5.2.3.2. Una vez obtenida la secuencia de símbolos a partir de la señal de voz se pueden aplicar los algoritmos de modelado HMM vistos en el apartado 5.2.

La fase de entrenamiento, en que se construyen los modelos, se realiza en base a un conjunto de observaciones obtenidas para cada palabra y de una determinada inicialización de los parámetros. Usualmente, se utiliza un algoritmo de estimación de máxima probabilidad como el de Baum-Welch, cuyas fórmulas para el caso de múltiples observaciones son (5.45)-(5.47) y será utilizado en la pruebas experimentales de este trabajo. También se han propuesto otros algoritmos de entrenamiento, que ya han sido citados en el apartado 5.2.1.3.

Hay que hacer notar que la utilización de modelos ocultos de Markov con topología izquierda-derecha como el descrito en la figura 5.7, que ha sido el escogido en las pruebas experimentales de este trabajo, conlleva implícitamente varias condiciones sobre los parámetros del modelo que obligan a modificar las fórmulas de reestimación (5.45)-(5.47).

En primer lugar, al comenzar siempre el proceso en el estado 1, el vector de probabilidades iniciales tendrá un valor fijo que no habrá que entrenar: $\pi_1 = 1$ y $\pi_i = 0$, $i \neq 1$. No se utiliza, por tanto, la fórmula (5.45). Por otro lado, las restricciones estocásticas sobre una matriz de transiciones que cumpla (5.58) obligan a que $a_{NN} = 1$. Este valor nos llevaría a la conclusión de que la duración esperada del estado N del modelo es infinita, lo cual no es coherente con el hecho de que las secuencias de observaciones sean finitas. Para corregir esta incoherencia, se incorpora al modelo un estado terminal F, que no genera observaciones, al cual se produce una transición desde el estado N cuando se genera la última observación. A este respecto hay que mencionar que en los modelos usados en las pruebas experimentales de este trabajo, se ha obligado que la última observación se produzca en el estado N mediante una adecuada inicialización de la probabilidades *backward*: $\beta_T(N) = 1$ y $\beta_T(i) = 0$, $i \neq N$. Todo esto, naturalmente, requiere modificaciones en la fórmulas (5.24) y (5.46).

Posteriormente a la fase de entrenamiento propiamente dicha se procede al suavizado de los modelos, que puede consistir simplemente en la aplicación de la técnica de *floor smoothing* o en la de otras técnicas más elaboradas descritas en el apartado 5.2.2.4.

Una vez construidos los modelos, la fase de reconocimiento selecciona el modelo más probable λ_X, perteneciente al conjunto de modelos correspondientes a las diferentes palabras a reconocer $\Lambda = \{\lambda_i\}_{i=1,...L}$, dada la secuencia de observaciones correspondientes a la palabra de test $O = \{O_t\}_{t=1,...T}$. Es decir, se busca

$$\lambda_X = \max_{\lambda \in \Lambda} \{ P(\lambda \mid O) \}. \qquad (5.59)$$

Estas probabilidades a priori pueden calcularse en base a las probabilidades a posteriori de la secuencia de observaciones por parte de los modelos según la regla de Bayes

$$P(\lambda_i|O) = \frac{P(O|\lambda_i)P(\lambda_i)}{\sum_{j=1}^{L} P(O|\lambda_j)P(\lambda_j)}, \qquad i = 1,... L \qquad (5.60)$$

donde $P(\lambda_i)$ es la probabilidad de ocurrencia del modelo λ_i. Dado que el denominador de (5.60) es constante, si suponemos que las probabilidades $P(\lambda_i)$ de todos los modelos son iguales, la expresión (5.59) es equivalente a

$$\lambda_X = \max_{\lambda \in \Lambda} \{ P(O \mid \lambda) \}. \qquad (5.61)$$

Por tanto, para identificar la palabra de test basta con calcular las probabilidades a posteriori con un algoritmo eficiente y seleccionar el modelo de mayor probabilidad. Estas probabilidades a posteriori pueden calcularse mediante el algoritmo *Forward-Backward* (ver apartado 5.2.1.1). Sin embargo, en muchos sistemas de reconocimiento, como el usado en este trabajo, se utiliza por su mayor eficiencia el algoritmo de Viterbi (ver apartado 5.2.1.2), que proporciona una aproximación a dichas probabilidades considerando únicamente el camino óptimo.

Como resumen, en el esquema de la figura 5.8 se han intentado reflejar las principales fases del funcionamiento de un sistema de reconocimiento de palabras aisladas mediante modelos ocultos de Markov discretos.

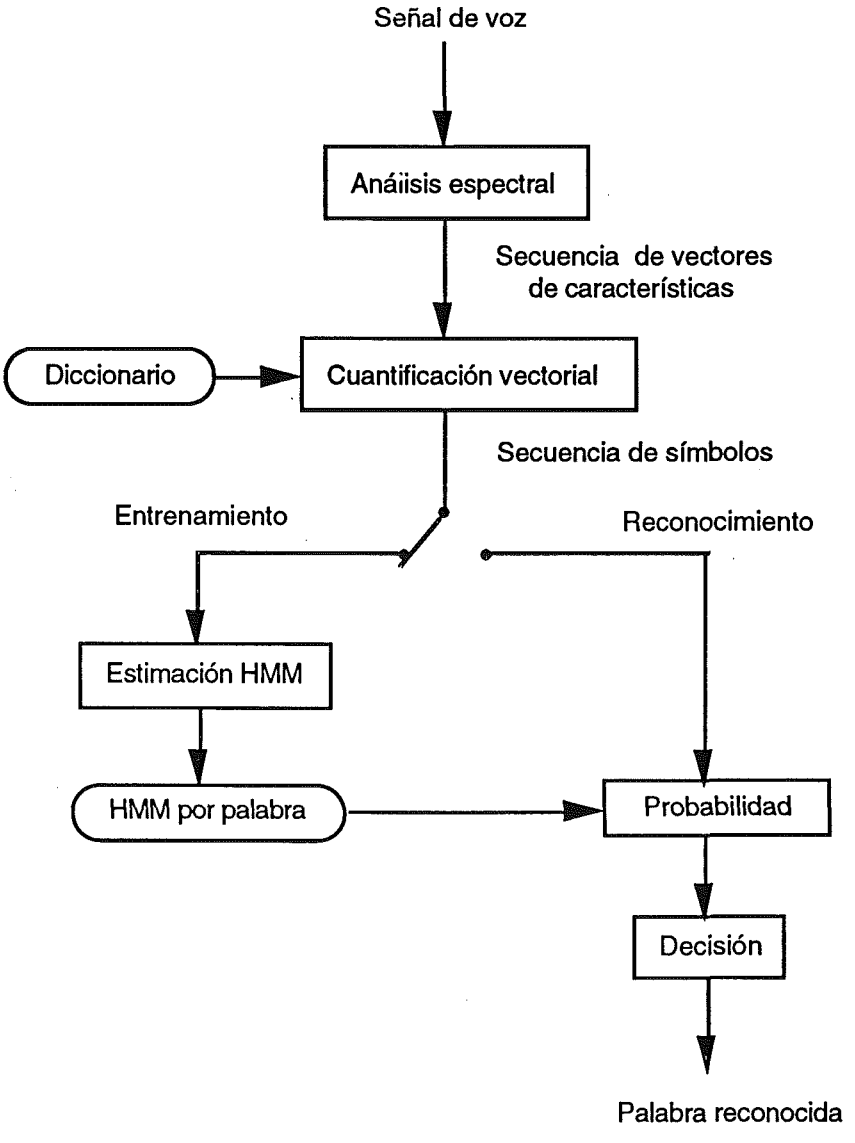

Fig. 5.8. Sistema de reconocimiento de palabras aisladas mediante HMM discretos

5.2.3.2. DISCRETIZACION DEL ESPACIO DE CARACTERISTICAS

Como ya se ha mencionado, la aplicación de los modelos discretos de Markov al reconocimiento del habla requiere la obtención de una secuencia de símbolos correspondiente a un alfabeto discreto y finito a partir de una secuencia de vectores que representan las características espectrales de la señal de voz.

Para ello, es necesario realizar una discretización del espacio de características de la señal de voz, la cual puede llevarse a cabo mediante técnicas de cuantificación

vectorial (*Vector Quantization* o VQ, en la literatura inglesa) [Gra84] [Mak85], que serán comentadas en este apartado. Dicha cuantificación vectorial consiste en establecer una partición del espacio vectorial en un conjunto finito de clases, de forma que quede unívocamente definida la clase a la que pertenece cada vector del espacio, y sustituir cada vector por un representante de cada clase. Los representantes de cada clase reciben el nombre de palabras-código y al conjunto de ellas se le conoce como diccionario del cuantificador. Una vez realizada la cuantificación vectorial de los vectores de parámetros espectrales de la señal de voz, se obtiene directamente la secuencia de símbolos requerida por los modelos ocultos de Markov discretos estableciendo una correspondencia entre las palabras-código del diccionario del cuantificador y los símbolos del alfabeto del modelo.

Dado un conjunto de finito de vectores $\{x_i\}_{i=1...NV}$, que constituye una representación estadísticamente significativa de los posibles valores de los vectores de observación, el problema de la selección de un conjunto de clases $\{y_i\}_{i=1...M}$ que represente adecuadamente el espacio de características necesita la especificación de un criterio de agrupamiento, que a su vez se formula a través de una medida de distancia entre vectores del espacio.

La medida de distancia entre vectores que se usará en el caso de la aplicación a reconocimiento del habla será alguna de las medias de distorsión espectral vistas en el capítulo 4 de esta memoria. En las pruebas experimentales realizadas en este trabajo se han usado las distancias cepstrales ponderadas euclídea y de proyección.

Una vez definida la medida de distancia, el criterio de agrupamiento más sencillo y comúnmente utilizado es el de minimizar la distancia media entre el conjunto de vectores de entrenamiento $\{x_i\}_{i=1...NV}$ y el conjunto de clases $\{y_i\}_{i=1...M}$, en base a la suma de las distancias entre el conjunto de vectores y el conjunto de representantes de las clases o palabras-código $\{v_i\}_{i=1...M}$ (nótese que se ha usado la misma notación para la palabras-código del cuantificador que para los símbolos del modelo) en la forma

$$D = \frac{1}{NV} \sum_{i=1}^{M} \sum_{x \in y_i} d(x, v_i), \qquad (5.62)$$

donde $d(x,v_i)$ es la distancia entre el vector de observación y la palabra-código v_i y el valor de D se conoce como la distorsión media del cuantificador.

La minimización de la distorsión media (5.62) impone como criterio de asignación de vectores a clases la selección de la clase cuya palabra-código dista menos del vector considerado y, además, obliga a escoger como palabra-código de cada clase el vector del espacio que minimiza la distancia media entre dicho vector, también llamado centroide, y los vectores de observación asignados a dicha clase. En el caso particular de utilizar distancia euclídea, los centroides son simplemente las medias aritméticas de los vectores que pertenecen a cada clase. En general, el cálculo del centroide depende del tipo de distancia utilizada y puede conllevar un importante coste de cálculo.

Evidentemente, el valor de la distorsión media depende de la forma en que se elija la partición en M clases del conjunto de vectores y, por tanto, la partición óptima será aquella que minimice dicho valor. Desafortunadamente, no existe una solución analítica conocida a este problema salvo la prueba exhaustiva de todas las posibles particiones, que conllevaría un cálculo inabordable para un valor moderado de M.

Sin embargo, existen métodos iterativos que permiten alcanzar una partición subóptima. El más conocido, y usado en las pruebas experimentales realizadas en este trabajo, es el algoritmo de K-medias [Tou74], que comienza con una determinada partición en clases del conjunto de vectores de entrenamiento e iterar un proceso en el que se asignan los vectores a las clases y, posteriormente, se recalculan los representantes de las clases hasta que se ha cumplido un determinado criterio de convergencia. En las pruebas experimentales de este trabajo este criterio consiste en una disminución del 1% de la distorsión media.

El problema principal de estos algoritmos es que el diccionario obtenido finalmente depende fuertemente de la partición inicial considerada y no existe forma conocida de determinar la partición inicial óptima. Para seleccionar una partición inicial de forma que el valor final obtenido para la distorsión media sea cercano al mínimo absoluto se han propuesto diversos algoritmos jerárquicos [Dud73].

En las pruebas experimentales realizadas en este trabajo se ha utilizado un método iterativo de construcción jerárquica de diccionarios similar al propuesto para el algoritmo LBG [Lin80]. En este método la partición inicial consiste en una única clase que agrupa a todos los vectores del conjunto de entrenamiento. A partir de esta configuración inicial, en cada iteración se divide en dos cada una de las clases y se aplica el algoritmo K-medias para obtener iterativamente los valores de los representantes de las clases que componen la partición. El procedimiento se itera hasta obtener el número deseado de clases (ver Fig. 5.9).

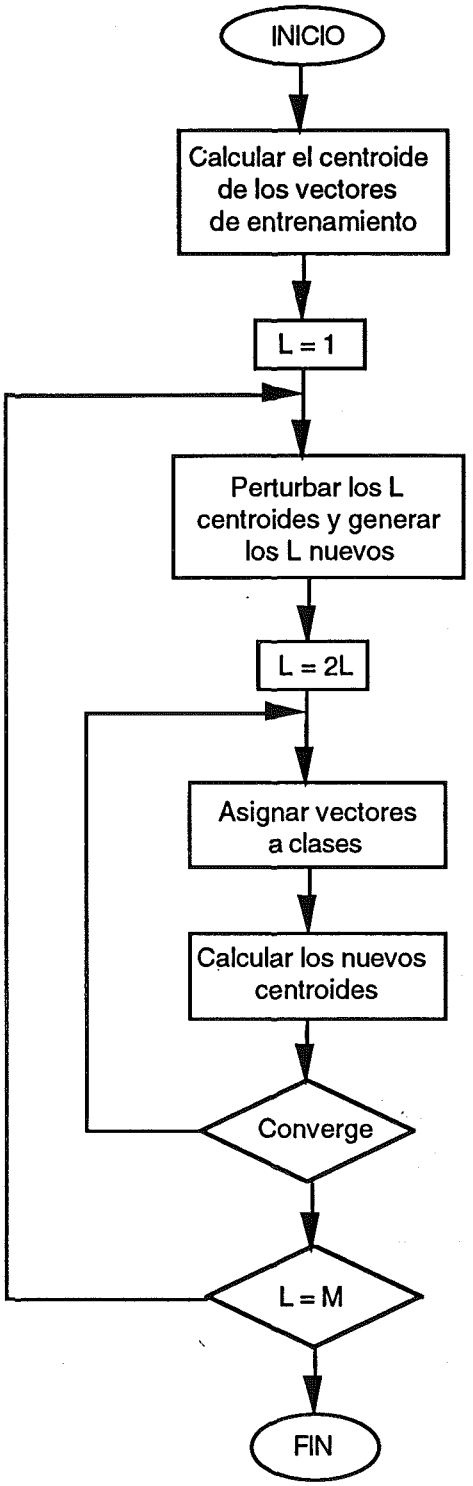

Fig. 5.9. Algoritmo jerárquico de construcción del diccionario

La división de las clases se realiza aplicando al representante $v_i^{(l)}$ de la clase $y_i^{(l)}$, de índice i de la iteración l, las expresiones (5.63) y (5.64) para obtener los representantes de las nuevas clases $y_i^{(l+1)}$ e $y_{i+l}^{(l+1)}$.

$$v_i^{(l+1)} = v_i^{(l)} \qquad (5.63)$$

$$v_{i+l}^{(l+1)} = (1+\mu)\, v_i^{(l)}. \qquad (5.64)$$

El valor de μ se elige pequeño para asegurar que los vectores pertenecientes a la clase $y_i^{(l)}$ se distribuyan efectivamente entre una de las nuevas clases $y_i^{(l+1)}$ e $y_{i+l}^{(l+1)}$.

5.2.3.3. UTILIZACION DE VARIAS INFORMACIONES

Hasta ahora, se ha descrito el caso de un sistema de reconocimiento del habla con HMM discretos en el que sólo se caracteriza la señal de voz por una información, en concreto el espectro instantáneo correspondiente a cada segmento de señal. Sin embargo, es cada vez más usual en reconocimiento del habla el uso de varias informaciones, como las características dinámicas del espectro (ver capítulo 3), la energía y las derivadas de esta. Esta incorporación de varias informaciones a un sistema como el descrito anteriormente puede realizarse de dos maneras fundamentalmente.

Una posibilidad es 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 en el proceso de cuantificación vectorial. En este caso, salvo en lo que respecta a la longitud del nuevo vector de características, el sistema de reconocimiento es idéntico al descrito para el caso del uso de una información.

La segunda posibilidad consiste en cuantificar por separado cada una de las informaciones y considerar independencia estadística de las mismas en el entorno de los modelos ocultos de Markov. A continuación se describirán las modificaciones que hay que introducir en este caso en la notación de los elementos de los modelos y en los algoritmos.

Cuando se utilizaba una sola información, la observación discreta O_t tomaba valores dentro de un alfabeto finito de símbolos $V = \{v_k\}_{k=1...M}$, siendo M el tamaño

del alfabeto. Ahora, al utilizar un número C de informaciones, cada observación discreta O_t estará compuesta por C valores correspondientes cada uno de ellos a C alfabetos de símbolos finitos y distintos $V^c = \{v_k^c\}_{k^c=1...Mc}$, para c = 1,... C, siendo M_c el tamaño del alfabeto V_c. Por tanto, donde antes se escribía $O_t = v_k$ ahora se escribirá $O_t = \{v_k^c\}_{c=1...C}$.

En cuanto a las probabilidades de generación de observaciones, en el caso de utilizar una sola información la probabilidad de generar una observación O_t en un estado j $b_j(O_t)$ era idéntica, por definición, a la probabilidad de observar el símbolo correspondiente v_k $b_j(k)$, es decir,

$$b_j(O_t) \mid_{O_t=v_k} = b_j(k), \qquad (5.65)$$

con

$$b_j(k) = P(\, v_k \text{ en } t \mid q_t = S_j \,) \qquad j = 1 \ldots N \qquad k = 1 \ldots M. \qquad (5.10)$$

Sin embargo, cuando se utilizan varias informaciones estadísticamente independientes la probabilidad de observar la observación O_t será el producto de las probabilidades de observación de los símbolos v_k^c correspcndientes a cada una de las informaciones. Es decir,

$$b_j(O_t) \mid_{O_t=\{v_k^c\}_{c=1...C}} = \prod_{c=1}^{C} b_j(k^c), \qquad (5.66)$$

con

$$b_j(k^c) = P(\, v_k^c \text{ en } t \mid q_t = S_j \,) \qquad j = 1 \ldots N \qquad k^c = 1 \ldots M_c \qquad c = 1,\ldots C. \qquad (5.67)$$

En este caso, las fórmulas de los algoritmos *Forward-Backward* y de Viterbi serán las mismas que las vistas en los apartados 5.2.1.1 y 5.2.1.2, respectivamente considerando la nueva expresión de $b_j(O_t)$ (5.66). En cuanto al algoritmo de entrenamiento de los parámetros del modelo, las fórmulas de π_j y a_{ij} tendrán las mismas expresiones que las vistas en 5.2.2.3 para varias secuencias de observaciones. Sin embargo, en el caso de las probabilidades de generación de observaciones $b_j(k)$ se

tendrá que sustituir la expresión (5.47), para el caso de una secuencia de observaciones, por

$$b_j'(k^c) = \frac{\sum_{l=1}^{L} \sum_{t=1, v_k^c \in O_t^{(l)}}^{T_l} \alpha_t^{(l)}(j)\beta_t^{(l)}(j)}{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \alpha_t^{(l)}(j)\beta_t^{(l)}(j)} \qquad j = 1,\dots N \quad k^c = 1,\dots M_c \quad c = 1,\dots C$$

(5.68)

5.3. MODELOS OCULTOS DE MARKOV CONTINUOS

Hasta este punto, se ha descrito solamente el caso en que las observaciones tomaban valores dentro de un alfabeto discreto y finito de símbolos y en este caso se utilizaban distribuciones de probabilidad discretas para modelar las probabilidades de generación de símbolos. En el caso de la aplicación al reconocimiento del habla, los vectores espectrales de las tramas de voz eran cuantificados vectorialmente y las observaciones discretas se correspondían con las etiquetas de las palabras-código producto de dicha cuantificación vectorial.

Como ya se apuntó en el apartado 5.1, dentro de la formulación general de los modelos ocultos de Markov puede suponerse que las observaciones toman valores dentro de un espacio continuo multidimensional, lo que fuerza a modelar las probabilidades de observación a través de funciones de densidad de probabilidad multivariadas. Esta aproximación da lugar a los llamados modelos ocultos de Markov continuos (CHMM). En el caso del reconocimiento del habla, las observaciones consisten simplemente en los vectores espectrales de las tramas de voz sin cuantificación.

En este caso, es necesario seleccionar una forma paramétrica para estas funciones de densidad de probabilidad que permitan establecer unas fórmulas de reestimación de los parámetros de dichas funciones de forma consistente. Un modelo usual es la función de densidad gaussiana multivariada [Pau86] o una combinación lineal finita o mezcla de ellas [Rab86a] [Jua86], aunque también se ha propuesto el uso de una mezcla de laplacianas [Ney88] o el modelado autorregresivo de las observaciones [Jua85].

En el caso de la utilización de una mezcla de funciones de densidad gaussianas multivariadas, cuya formulación es bastante general dado que puede aproximar arbitrariamente cualquier densidad de probabilidad sin más que tomar un número suficientemente elevado de términos en la combinación lineal, la probabilidad de generación de observaciones adopta la forma general

$$b_j(O_t) = \sum_{m=1}^{M} c_{jm}\, N(O_t, \mu_{jm}, \Sigma_{jm}) \qquad j = 1,\ldots N, \qquad (5.69)$$

donde N es una función de densidad de probabilidad gaussiana de vector media μ_{jm} y matriz de covarianza Σ_{jm}, que forma parte de la combinación lineal de M gausssianas con peso relativo c_{jm}. Estos pesos, que se denominan coeficientes de mezcla, verifican las relaciones

$$\sum_{m=1}^{M} c_{jm} = 1 \qquad j = 1,\ldots N \qquad (5.70)$$

$$c_{jm} \geq 0 \qquad m = 1,\ldots M \quad j = 1,\ldots N, \qquad (5.71)$$

de forma que las densidades de probabilidad así obtenidas verifican la propiedad de normalización

$$\int_{-\infty}^{+\infty} b_j(x)\, dx = 1 \qquad j = 1,\ldots N. \qquad (5.72)$$

Las fórmulas de reestimación para los coeficientes de la mezcla, así como para los vectores media y matrices de covarianza pueden encontrarse en [Jua86]. No se transcribirán en esta memoria, pues este tipo de modelado HMM no se ha sido utilizado en las pruebas experimentales. En cuanto a la reestimación del resto de los parámetros del modelo, así como los algoritmos de evaluación y codificación expuestos en el apartado 5.2, no se ven afectados más que en lo que concierne a la evaluación de los valores de $b_j(O_t)$.

La implementación de estos modelos continuos presenta problemas similares a los apuntados en el apartado 5.2.2. El escalado temporal y la compresión logarítmica de las probabilidades, así como la reestimación de los parámetros del modelo en el caso de múltiples secuencias de observaciones, se resuelven de forma análoga al caso de los

modelos discretos. La cuestión del suavizado deja de tener sentido al dejar de haber distribuciones de probabilidad discretas. Sin embargo, se impone siempre un umbral mínimo para el valor de la funciones de densidad de probabilidad por razones análogas al uso de la técnica de *floor smoothing* en los modelos discretos.

Por lo que respecta a las estimaciones iniciales de los parámetros del modelo, la experiencia demuestra que la elección aleatoria o uniforme para los valores de las probabilidades iniciales de los estados Π y para las probabilidades de transición A resulta adecuada para obtener parámetros útiles en casi todos los casos, como en el caso de los modelos discretos. Sin embargo, en el caso de las probabilidades de observación se ha observado que estimaciones iniciales buenas resultan esenciales en los modelos continuos. En la bibliografía pueden encontrarse diversas formas de obtener estimaciones iniciales adecuadas, obtenidas siempre a partir de una segmentación inicial de las secuencias de entrenamiento entre los estados del modelo, la cual puede ser manual, lineal o de máxima probabilidad. Un método iterativo muy usado consiste en partir de un inicialización burda de los modelos u en realizar en cada paso una segmentación de máxima probabilidad mediante el algoritmo de Viterbi y una reestimación de los parámetros de las funciones de densidad de probabilidad de cada estado a partir del agrupamiento de los vectores de observación mediante el algoritmo de K-medias [Rab86b].

Por otro lado, en caso de usar varias informaciones de la señal de voz, al no existir etapa de cuantificación vectorial, siempre se opta por construir un supervector concatenando con una ponderación adecuada los vectores y/o las componentes escalares que se desean utilizar.

La principal ventaja de usar modelos continuos es la capacidad de modelar directamente los parámetros del habla, sin preocuparse de los errores de cuantificación vectorial ni de la definición de una medida de distancia vectorial en el cuantificador. Además, el uso de modelos continuos conduce a una modesta reducción del número de parámetros a entrenar en el caso de utilizar distribuciones de probabilidad sencillas como una única gaussiana o matrices de covarianza diagonales.

Sin embargo, los modelos continuos requieren un tiempo considerablemente más largo para el entrenamiento y el reconocimiento que los modelos discretos. Por ejemplo, obtener la probabilidad de observación usando modelos discretos consiste simplemente en mirar una tabla, mientras que usando modelos continuos se necesitan muchas operaciones incluso en el caso más simple de una única gaussiana y matriz de

covarianza diagonal. Esta diferencia de coste computacional es todavía más evidente si se considera que una sola gaussiana con matriz de covarianza diagonal no representa adecuadamente los parámetros del habla [Rab86a] y que el uso de mezclas o matrices de covarianza completas incrementa considerablemente la complejidad del entrenamiento y el reconocimiento con modelos continuos.

Por otro lado, hay bastantes discrepancias en la literatura sobre el tipo de modelado que obtiene mejores resultados en reconocimiento del habla. Según Brown [Bro87], el comportamiento de la estimación de máxima probabilidad sólo es predecible si (1) las distribuciones supuestas son correctas, (2) las distribuciones son bien comportadas y (3) el tamaño de las muestras es lo suficientemente grande. Cuando se usan modelos discretos, debido a que las distribuciones no son paramétricas, al menos (1) y (2) no se violan. Sus resultados también sugieren que para usar estimación de máxima probabilidad y modelos continuos se necesitan distribuciones complejas, como mezclas de gaussianas o matrices de covarianza completas, lo cual requiere un considerable coste computacional, como ya se ha mencionado, y una gran cantidad de datos de entrenamiento para una estimación fiable del gran número de parámetros libres correspondientes a estas distribuciones complejas.

Por tanto, a pesar de las ventajas de los modelos continuos, las razones aludidas han llevado que en este trabajo no se hayan realizado pruebas con este tipo de modelos. A todo esto hay que añadir el hecho de que no se disponía de todos los recursos necesarios para realizar pruebas exhaustivas con este tipo de modelos.

Aunque los modelos discretos no pueden evitar los errores de cuantificación vectorial, son muy eficientes y pueden representar cualquier tipo de distribución, ya que no hacen suposiciones sobre la distribución subyacente de los símbolos observados. De ahí que la mayoría de las pruebas realizadas en este trabajo hayan sido realizadas con este tipo de modelos. Hay que hacer notar, no obstante, que los modelos continuos constituyen actualmente una importante área de investigación y que muchas de las ideas propuestas en esta memoria pueden también aplicarse a ellos.

A continuación, se describirán aproximaciones intermedias al modelado de Markov entre las aproximaciones clásicas discreta y la continua, que intentan minimizar los errores de cuantificación vectorial que se producen en los modelos discretos y evitar, a la vez, los problemas de ineficiencia, suposiciones incorrectas sobre las distribuciones subyacentes de los parámetros de la voz y entrenamiento costoso que se producen en los modelos continuos.

Se trata de los modelos ocultos de Markov semicontinuos (SCHMM) y los modelos ocultos de Markov con múltiple etiquetado, con los cuales también se han realizado pruebas experimentales en este trabajo. Como se verá en el capítulo 6, gracias a ciertas modificaciones introducidas en la cuantificación vectorial, los resultados obtenidos con ellos en condiciones limpias y ruidosas superan claramente los obtenidos con modelos discretos.

5.4. MODELOS OCULTOS DE MARKOV SEMICONTINUOS

Como ya se ha visto, en los modelos ocultos de Markov discretos (DHMM) la cuantificación vectorial hace posible el uso de distribuciones discretas de probabilidad no paramétricas, capaces de modelar adecuadamente cualquier estadística subyacente. El problema más importante de este tipo de modelado es que dicha cuantificación vectorial divide el espacio de características en regiones totalmente separadas correspondientes a cada una de las palabras-código, de forma que cada vector a cuantificar es asociado a una sola de ellas sin tener en cuenta su proximidad a otras palabras-código. Esta estricta regla de decisión puede causar una seria pérdida de información para el modelado siguiente. Otra desventaja del modelado discreto HMM, no mencionada hasta ahora, es que el cuantificador vectorial y el modelo discreto son construidos de forma independiente, lo cual puede no ser una solución óptima para el proceso de clasificación.

Los modelos ocultos de Markov semicontinuos (SCHMM), propuestos inicialmente por Huang y Jack [Hua89], intentan paliar el problema de la cuantificación vectorial antes mencionado modelando el diccionario del cuantificador vectorial mediante una familia de funciones de densidad de probabilidad gaussianas solapadas. Cada palabra-código está representada por una de estas funciones de probabilidad gaussianas. Gracias al solapamiento entre dichas funciones no se produce la partición del espacio de características y, por tanto, cada palabra-código puede usarse conjuntamente con otras palabras-código para modelar los vectores de características. Como consecuencia, se minimizan los errores debidos a la cuantificación vectorial. Además, usando esta formulación el cuantificador vectorial y el modelo de Markov pueden unificarse dentro del mismo entorno probabilístico para obtener una combinación VQ/HMM optimizada.

Los elementos del modelo oculto de Markov semicontinuo son los mismos que los descritos para un modelo discreto, incluso en el caso de la matriz B de distribuciones probabilidad discretas. Sin embargo, en el caso de los modelos semicontinuos las observaciones no son símbolos v_k sino vectores de características O_t. Por tanto, el término $b_j(k)$ no corresponderá a la probabilidad de observación del símbolo v_k en el estado S_j sino a una probabilidad asignada a la palabra-código de índice k del diccionario del nuevo cuantificador vectorial, que se denotará también como v_k.

Nótese que en el entorno de los modelos discretos v_k denotaba un símbolo observable de un alfabeto finito, que se correspondía con una palabra-código de un cuantificador vectorial convencional, es decir, un determinado vector del espacio de características. Mientras, en el entorno de los modelos semicontinuos no se hablará de símbolos observables y v_k denotará directamente una palabra-código del nuevo cuantificador vectorial, que ya no consistirá en un vector del espacio de características sino que estará representado por una función de densidad de probabilidad.

La probabilidad de una observación O_t en el caso semicontinuo se calculará en base a estas probabilidades asignadas a cada palabra-código v_k y a las funciones de densidad de probabilidad gaussianas asociadas a cada palabra-código. Para un estado dado S_j del modelo, la función densidad de probabilidad de que se genere una observación O_t, que se corresponde con un vector del espacio de características, se expresa como

$$b_j(O_t) = \sum_{k=1}^{M} f(O_t|v_k, S_j) P(v_k|S_j), \tag{5.73}$$

donde M denota el número de palabras-código del diccionario del cuantificador vectorial, v_k denota la k-ésima palabra-código y $f(O_t|v_k,S_j)$ es la función densidad de probabilidad correspondiente a la palabra-código v_k evaluada en O_t, dado el modelo S_j. Como se supone que $f(O_t|v_k,S_j)$ es independiente del estado S_j, (5.73) puede escribirse en función de los elementos de la matriz de observaciones B como

$$b_j(O_t) = \sum_{k=1}^{M} f(O_t|v_k) b_j(k). \tag{5.74}$$

La probabilidad de observación semicontinua formulada en (5.74) puede considerarse una aproximación intermedia entre las aproximaciones clásicas discreta y continua, que cuenta con gran número de ventajas. Desde el punto de vista de los

modelos discretos, los modelos semicontinuos minimizan los errores debidos a la cuantificación vectorial de una manera bastante eficiente. Desde el punto de vista de los modelos continuos, los modelos semicontinuos pueden considerarse como una forma especial de modelos continuos con mezclas de M funciones de densidad gaussianas compartidas por todos los modelos. En este caso, los términos $b_j(k)$ corresponderían a los M coeficientes de la mezcla de cada modelo S_j. Al ser compartidas las funciones de densidad gaussianas, el número de parámetros libres y la complejidad computacional se reducen en comparación con los modelos continuos con mezclas de M gaussianas. La disminución de parámetros libres permite entrenar estos modelos con una cantidad de datos significativamente menor. En cuanto a la complejidad computacional, esta resulta comparable en el proceso de decodificación con la de unos modelos continuos con una única gaussiana por modelo.

Además, en la práctica, suele reducirse la complejidad computacional de los modelos semicontinuos aproximando (5.74) con los K valores más significativos de $f(O_t|v_k)$ para cada valor de O_t, lo cual no afecta las prestaciones del sistema. Denotando $\eta(O_t)$ el conjunto de palabras-código que dan lugar a estos K valores más significativos, (5.74) puede escribirse como

$$b_j(O_t) = \sum_{\eta(O_t)} f(O_t|v_k) b_j(k). \qquad (5.75)$$

En este caso, las fórmulas de los algoritmos de evaluación *Forward-Backward* y de decodificación de Viterbi serán las mismas que las vistas en los apartados 5.2.1.1 y 5.2.1.2 para el caso discreto, respectivamente, considerando la nueva expresión de $b_j(O_t)$ (5.75).

En cuanto al algoritmo de entrenamiento de los parámetros del modelo, las fórmulas de π_i y a_{ij} tendrán las mismas expresiones que las vistas en 5.2.2.3 para múltiples secuencias de observaciones. Sin embargo, en el caso de las probabilidades $b_j(k)$, utilizando el criterio de estimación de máxima probabilidad, se tendrá que sustituir la expresión (5.47),

$$b_j'(k) = \frac{\sum_{l=1}^{L} \sum_{t=1, O_t^{(l)}=v_k}^{T_l} \alpha_t^{(l)}(j)\beta_t^{(l)}(j)}{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \alpha_t^{(l)}(j)\beta_t^{(l)}(j)} \quad j = 1,\dots N \quad k = 1,\dots M \tag{5.47}$$

por

$$b_j'(k) = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \alpha_t^{(l)}(j)\beta_t^{(l)}(j) \frac{f(O_t^{(l)}|v_k) b_j(k)}{b_j(O_t^{(l)})}}{\sum_{l=1}^{L} \sum_{t=1}^{T} \alpha_t^{(l)}(j)\beta_t^{(l)}(j)} \quad j = 1,\dots N \quad k = 1,\dots M, \tag{5.76}$$

ya que el cociente

$$\frac{f(O_t^{(l)}|v_k) b_j(k)}{b_j(O_t^{(l)})} \tag{5.77}$$

expresa la contribución de la función de densidad de probabilidad asociada a la palabra-código v_k en la probabilidad de la observación $O_t^{(l)}$. El índice l indica que los valores han sido obtenidos para la secuencia de observaciones l (ver apartado 5.2.2.3).

Por otro lado, ya se ha mencionado que la formulación de los modelos semicontinuos permite la unificación del cuantificador vectorial y el modelo de Markov dentro del mismo entorno probabilístico para obtener una combinación VQ/HMM optimizada. Esta optimización conjunta implica que el cuantificador vectorial se ajusta junto con los parámetros del modelo, en lugar de diseñarse en base a una minimización de la distorsión de cuantificación.

La reestimación de los parámetros del cuantificador vectorial, es decir los vectores de medias μ_k y las matrices de covarianza Σ_k de las funciones de densidad de probabilidad gaussianas $f(O_t|v_k)$, se realiza mediante consideraciones de estimación de máxima probabilidad como en el algoritmo de Baum-Welch de entrenamiento de los modelos de Markov.

Para ello, es conveniente definir la variable $\zeta_t(k)$ como la probabilidad correspondiente a la palabra-código v_k en un instante t, dada la secuencia de observaciones O y el modelo λ, es decir,

$$\zeta_t(k) = P(v_k \text{ en } t \mid O, \lambda) = \sum_{j=1}^{N} P(q_t=S_j, v_k \text{ en } t|O,\lambda). \qquad (5.78)$$

Teniendo en cuenta las expresiones (5.39) y (5.40) correspondientes a la variable de ocupación de estados $\gamma_t(i)$, recogidas en el apartado de modelos discretos y válidas también en este caso, y el significado del cociente de probabilidades (5.77), se puede expresar $\zeta_t(k)$ como

$$\zeta_t(k) = \frac{1}{P(O|\lambda)} \sum_{j=1}^{N} \alpha_t(j)\ \beta_t(j)\ \frac{f(O_t|v_k)b_j(k)}{b_j(O_t)}. \qquad (5.79)$$

Considerando estimación de máxima probabilidad, las expresiones de reestimación de los vectores de medias μ_k y de las matrices de covarianza Σ_k en función de $\zeta_t(k)$, para el caso de L secuencias de observaciones denotadas con índice l, serán

$$\mu'_k = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \zeta_t^{(l)}(k)\ O_t^{(l)}}{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \zeta_t^{(l)}(k)} \qquad k = 1,\ldots M \quad (5.80)$$

$$\Sigma'_k = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \zeta_t^{(l)}(k)\ (O_t^{(l)}-\mu'_k)(O_t^{(l)}-\mu'_k)^T}{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \zeta_t^{(l)}(k)} \qquad k = 1,\ldots M. \quad (5.81)$$

Incluyendo el valor de $\zeta_t(k)$ en estas expresiones, se obtienen las fórmulas definitivas de reestimación de las funciones de densidad del cuantificador vectorial

$$\mu'_k = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\sum_{j=1}^{N} \alpha_t^{(l)}(j)\,\beta_t^{(l)}(j)\,\frac{f(O_t^{(l)}|v_k)\,b_j(k)}{b_j(O_t^{(l)})}\,O_t^{(l)}}{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\sum_{j=1}^{N} \alpha_t^{(l)}(j)\,\beta_t^{(l)}(j)\,\frac{f(O_t^{(l)}|v_k)\,b_j(k)}{b_j(O_t^{(l)})}} \qquad k = 1,\ldots M \qquad (5.82)$$

$$\Sigma'_k = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\sum_{j=1}^{N} \alpha_t^{(l)}(j)\,\beta_t^{(l)}(j)\,\frac{f(O_t^{(l)}|v_k)\,b_j(k)}{b_j(O_t^{(l)})}\,(O_t^{(l)}-\mu'_k)(O_t^{(l)}-\mu'_k)^T}{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\sum_{j=1}^{N} \alpha_t^{(l)}(j)\,\beta_t^{(l)}(j)\,\frac{f(O_t^{(l)}|v_k)\,b_j(k)}{b_j(O_t^{(l)})}} \qquad k = 1,\ldots M. \qquad (5.83)$$

Nótese que en las expresiones (5.81) y (5.83) la reestimación de las matrices de covarianza dependen del valor reestimado de los vectores de medias μ'_k.

Debido al elevado coste computacional de las expresiones (5.82) y (5.83), el proceso de entrenamiento en el modelado semicontinuo no suele realizarse mediante una reestimación iterativa en la que se ajustan en cada paso de la iteración los parámetros de los modelos y del cuantificador vectorial. Además, se ha observado que unas buenas inicializaciones de las distribuciones de probabilidad discretas de los modelos y de los parámetros de las funciones de densidad del cuantificador vectorial son esenciales para la obtención de una buenas estimaciones finales. Debido a estos dos factores, se han propuesto varias estrategias de inicialización y reestimación.

Una posible estrategia, llevada a cabo en las pruebas experimentales realizadas en este trabajo, consiste en realizar la inicialización en base a los parámetros de un modelado discreto y en reestimar separadamente los modelos y el cuantificador vectorial.

Pueden tomarse como valores iniciales de los parámetros de los modelos semicontinuos los parámetros de unos modelos discretos que tengan la misma estructura y hayan sido entrenados para representar el mismo proceso. Por otro lado, pueden inicializarse los parámetros de ia funciones de densidad de probabilidad del nuevo cuantificador vectorial utilizando el cuantificador vectorial discreto usado para crear los modelos discretos anteriores. Lo más razonable es tomar como medias μ_k de las nuevas funciones de densidad las palabras-código del cuantificador convencional y estimar las matrices de covarianza Σ_k a partir de la distribución estadística de los

vectores de características de entrenamiento asignados a cada una de estas palabras-código. La complejidad de cálculo ha llevado a considerar matrices de covarianza diagonales en las pruebas experimentales realizadas en este trabajo, hipótesis aceptable en cierta medida al utilizar como vectores de características de la señal de voz los coeficientes cepstrum LPC y emplear un número considerable de funciones de densidad gaussianas.

Una vez realizada la inicialización, pueden reestimarse separadamente los parámetros de los modelos y el cuantificador vectorial dividiendo el proceso de reestimación en dos etapas. En la primera etapa se reestiman iterativamente los parámetros de los modelos hasta maximizar la probabilidad de generación de las secuencias de entrenamiento utilizando el cuantificador vectorial inicial en todas las iteraciones. En la segunda etapa, se utilizan las expresiones (5.82) y (5.83) para obtener unos nuevos valores de las medias y las matrices de covarianza de las funciones de densidad gaussianas del cuantificador vectorial. Todo el proceso podría iterarse para obtener un refinamiento de todos los parámetros del sistema.

Sin embargo, en las pruebas experimentales presentadas en el capítulo 6 no se ha iterado el proceso y, además, no se ha llevado a cabo la segunda etapa de actualización de los parámetros de las funciones del cuantificador vectorial debido al elevado coste de cálculo que supone y al deseo de comparar el comportamiento de los modelos semicontinuos con los modelos discretos y los de múltiple etiquetado en condiciones similares de complejidad.

En cuanto a otros problemas de implementación de los modelos semicontinuos, el escalado temporal y la compresión logarítmica de las probabilidades se resuelven de forma análoga al caso de los modelos discretos, pero el uso de técnicas elaboradas de suavizado deja de tener sentido por el solapamiento existente entre las funciones de densidad gaussianas del cuantificador vectorial. Sin embargo, siempre se impone un umbral mínimo para los valores de las distribuciones de probabilidad discretas de los modelos (*floor smoothing*), como en el caso de los modelos discretos, y un umbral mínimo para los valores de las funciones de densidad gaussianas del cuantificador vectorial, como se hace en los modelos continuos para los valores de las funciones de densidad de probabilidad de observación.

Por otro lado, en caso de usar varias informaciones de la señal de voz, al igual que en el caso de los modelos discretos, se puede optar por construir un supervector concatenando con una ponderación adecuada los vectores y/o las componentes escalares

que se desean utilizar o bien construir cuantificadores independientes para cada información y considerar independencia estadística de las probabilidades correspondientes en cada una de ellas. En el primer caso, salvo en lo que respecta a la longitud del nuevo vector de características, el modelado combinado VQ/HMM es idéntico al descrito para el caso del uso de una información. En el segundo caso, habría que introducir las modificaciones análogas a las vistas en el caso de modelos discretos en el apartado 5.2.2.3.

5.5. MODELOS OCULTOS DE MARKOV CON MULTIPLE ETIQUETADO

Como ya se ha mencionado, otra aproximación intermedia entre las aproximaciones clásicas discreta y continua al modelado de Markov, que intenta minimizar los errores de cuantificación vectorial que se producen en los modelos discretos y evitar, a la vez, los problemas de ineficiencia, suposiciones incorrectas sobre las distribuciones subyacentes de los parámetros de la voz y entrenamiento costoso que se producen en los modelos continuos, la constituyen los modelos ocultos de Markov con múltiple etiquetado.

En la cuantificación vectorial tradicional, que se realiza como paso previo al modelado discreto de Markov, se asocia cada vector de características a una sola palabra-código. La palabra-código elegida es el vector del diccionario más próximo al vector de características a cuantificar, con lo cual se descarta la información acerca del grado de proximidad al resto de los vectores del diccionario. La cuantificación vectorial múltiple intenta minimizar los errores de cuantificación que se producen en el caso anterior utilizando esta información. Para ello, asocia a cada vector a cuantificar los K vectores del diccionario más próximos y asigna a cada una de estas palabras-código un coeficiente relacionado con su proximidad al vector a cuantificar. La utilización adecuada de toda esta información permite modelar el espacio de características de una manera más flexible que la mera partición del mismo de un modo similar, aunque más simple, que el modelado realizado en el caso semicontinuo.

La aplicación de esta cuantificación vectorial múltiple al entorno de los modelos de Markov conduce a un nuevo tipo de modelos [Nis87]. Debido a que la cuantificación vectorial equivale a asociar los vectores a cuantificar con las etiquetas de las palabras-código del diccionario del cuantificador, también suele referirse a este proceso con el

nombre de etiquetado. Por ello, en esta memoria se utilizará el nombre de modelos de Markov con múltiple etiquetado para referirse a este tipo de modelos.

Los elementos del modelo oculto de Markov con múltiple etiquetado son los mismos que los descritos para un modelo discreto, incluso en el caso de la matriz B de distribuciones de probabilidad discretas. Sin embargo, al igual que en el caso de los modelos semicontinuos, las observaciones no se considerarán símbolos sino vectores de características O_t. Por tanto, el término $b_j(k)$ no corresponderá a la probabilidad de observación del símbolo v_k en el estado S_j sino a una probabilidad asignada a la palabra-código de índice k del diccionario del cuantificador vectorial, que se denotará también como v_k.

La probabilidad de una observación O_t en un estado S_j se calculará en base a estas probabilidades asignadas a cada palabra-código y a los coeficientes asignados a cada palabra-código en función de su proximidad al vector de características. Para un estado S_j del modelo, la probabilidad de que se genere una observación O_t se expresa como

$$b_j(O_t) = \sum_{k=1}^{K} w(O_t,v_k)\ b_j(k), \qquad (5.84)$$

donde $w(O_t,v_k)$ es el coeficiente asignado asignado a la palabra-código v_k en el etiquetado múltiple.

Estos coeficientes $w(O_t,v_k)$ pueden estimarse de diferentes modos. En las pruebas experimentales realizadas en este trabajo, se ha utilizado la siguientes expresión:

$$w(O_t,v_k) = \frac{1/d(O_t,v_k)}{\sum_{j=1}^{K} 1/d(O_t,v_j)} \qquad k = 1,\ldots K, \qquad (5.85)$$

donde $d(O_t,v_k)$ es la distancia entre el vector de observación O_t y la palabra código v_k. Naturalmente, se utilizará la la definición de distancia usada en el cuantificador.

Puede observarse que en el caso particular K = 1 el coeficiente que se asocia a la única palabra-código considerada es 1. Por tanto, estos modelos degeneran en modelos discretos, lo cual no ocurre en el caso de los modelos semicontinuos.

La probabilidad de observación formulada en (5.84) es análoga a la expresión simplificada correspondiente a los modelos semicontinuos expresada en (5.75). Por tanto, las ventajas de los modelos semicontinuos con respectos a los discretos y a los continuos ya comentadas pueden extrapolarse al caso de modelos con múltiple etiquetado. Además, hay que destacar que en el caso de estos modelos se evita la complejidad de cálculo y la dificultad de entrenamiento que implican las funciones de densidad de probabilidad gaussianas del cuantificador vectorial de los modelos semicontinuos.

Respecto a los algoritmos de evaluación y decodificación de estos modelos, las fórmulas serán las mismas que las vistas en los apartados 5.2.1.1 y 5.2.1.2 para el caso discreto, respectivamente, considerando la nueva expresión de $b_j(O_t)$ (5.84). En cuanto al algoritmo de entrenamiento de los parámetros del modelo, las fórmulas de π_i y a_{ij} tendrán las mismas expresiones que las vistas en 5.2.2.3 para múltiples secuencias de observaciones en el caso discreto. Sin embargo, en el caso de las probabilidades $b_j(k)$, análogamente al caso de los modelos semicontinuos, se tendrá que sustituir la expresión (5.47) por

$$b_j'(k) = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T_l} \alpha_t^{(l)}(j)\beta_t^{(l)}(j) \frac{w(O_t^{(l)},v_k)b_j(k)}{b_j(O_t^{(l)})}}{\sum_{l=1}^{L}\sum_{t=1}^{T} \alpha_t^{(l)}(j)\beta_t^{(l)}(j)} \qquad j = 1,\dots N \quad k = 1,\dots M, \tag{5.86}$$

donde, como siempre, el índice l corresponde a los valores obtenidos para cada una de las L secuencias de observaciones. Obviamente, $w(O_t^{(l)},v_k)$ sólo tomará valores no nulos para las K palabras-código más próximas al vector de observación.

En las pruebas experimentales de este trabajo, se ha utilizado también una fórmula alternativa a la expresada en (5.86) para la reestimación de las probabilidades asignadas a cada palabra-código. La expresión alternativa (5.87)

$$b_j'(k) = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T_l} \alpha_t^{(l)}(j)\beta_t^{(l)}(j)\, w(O_t^{(l)}, v_k)}{\sum_{l=1}^{L}\sum_{t=1}^{T} \alpha_t^{(l)}(j)\beta_t^{(l)}(j)} \qquad j = 1,\ldots N \quad k = 1,\ldots M \tag{5.87}$$

se deriva de la fórmula de reestimación correspondiente a la estimación de máxima probabilidad (5.86) simplemente sustituyendo el término

$$\frac{w(O_t^{(l)}, v_k) b_j(k)}{b_j(O_t^{(l)})} \tag{5.88}$$

por $w(O_t^{(l)}, v_k)$. Con la utilización de esta nueva fórmula se pretende favorecer el valor de la probabilidad de aquellas palabras-código más próximas al vector de observación sin utilizar la información de las probabilidades obtenidas en la iteración anterior, que es considerada en la estimación de máxima probabilidad. Este hecho provoca saltos bruscos en los valores de las probabilidades de una iteración a la siguiente

Un inconveniente que presenta esta alternativa es que no garantiza la convergencia hacia una probabilidad máxima de generación de la secuencia de observaciones, ya que no se corresponde con una estimación de máxima probabilidad. Sin embargo, en las pruebas experimentales realizadas en este trabajo los modelos entrenados con la fórmula alternativa (5.87) han superado en rapidez de convergencia y tasa de reconocimiento a los entrenados usando criterios de máxima probabilidad (5.86). Esta mayor rapidez de convergencia puede justificarse por los saltos bruscos que se producen en los valores de las probabilidades de las palabras-código.

Por otro lado, análogamente al caso de los modelos semicontinuos, también se puede realizar una reestimación del cuantificador vectorial para obtener una combinación VQ/HMM optimizada. En este caso, este proceso afectaría simplemente a las palabras-código del cuantificador, para las que la fórmula de reestimación sería análoga a la descrita en el apartado anterior para los vectores de medias de las funciones de densidad de probabilidad del cuantificador vectorial de los semicontinuos, es decir,

$$v'_k = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \sum_{j=1}^{N} \alpha_t^{(l)}(j)\, \beta_t^{(l)}(j)\, \frac{w(O_t^{(l)}, v_k) b_j(k)}{b_j(O_t^{(l)})}\, O_t^{(l)}}{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \sum_{j=1}^{N} \alpha_t^{(l)}(j)\, \beta_t^{(l)}(j)\, \frac{w(O_t^{(l)}, v_k) b_j(k)}{b_j(O_t^{(l)})}} \qquad k = 1,\ldots M \qquad (5.87)$$

Al igual que en los modelos semicontinuos, el elevado coste computacional de la expresión (5.87) y la necesidad de una buena inicialización de las distribuciones de probabilidad discretas de observación de los modelos obliga al uso de estrategias adecuadas y eficientes de inicialización y reestimación. La estrategia seguida en las pruebas experimentales presentadas en el capítulo 6 de esta memoria ha sido análoga a la usada en el caso de los modelos semicontinuos debido al deseo de comparar el comportamiento de todos los tipos de modelos en condiciones similares de complejidad. Es decir, se han tomado como valores iniciales de los parámetros de los modelos los parámetros de unos modelos discretos, con idéntica estructura y entrenados para representar el mismo proceso, y como cuantificador vectorial inicial el mismo usado para crear estos modelos discretos. Posteriormente, se han reestimado iterativamente los parámetros de los modelos utilizando el cuantificador vectorial inicial sin proceder en ningún momento al refinamiento del cuantificador.

En cuanto a otros problemas de implementación, el escalado temporal y la compresión logarítmica de las probabilidades se resuelven de forma análoga al caso de los modelos discretos, pero el uso de técnicas elaboradas de suavizado deja de tener sentido ya que este tipo de modelado combate el problema de entrenamiento insuficiente al servir cada vector de características de la base de entrenamiento para reestimar las probabilidades de las K palabras-código más cercanas. Sin embargo, sí se utiliza la técnica de *floor smoothing*. Por último, en caso de usar varias informaciones de la señal de voz, lo dicho para el caso de los modelos discretos y semicontinuos también es aplicable en este caso.

Como se verá en el capítulo 6, los modelos semicontinuos y los modelos con múltiple etiquetado proporcionan prestaciones superiores a las de los modelos discretos en reconocimiento automático del habla. La causa principal, ya comentada a lo largo de este capítulo, es la minimización del error de cuantificación.

Los resultados han sido particularmente excelentes en el caso de habla ruidosa. Este hecho puede explicarse teniendo en cuenta que en presencia de ruido la cuantificación vectorial convencional usada por los modelos discretos introduce,

además de los errores debidos a la distorsión inherente a la cuantificación, errores de etiquetado que tienen graves repercusiones en la tasa de reconocimiento del sistema. Este efecto, sin embargo, se minimiza en el caso de la cuantificación vectorial usada por los modelos semicontinuos y con múltiple etiquetado.

También se observará que los modelos con múltiple etiquetado, con una complejidad y un coste de cálculo mucho menores, proporcionan prestaciones ligeramente superiores a los modelos semicontinuos. Una posible explicación es la mejor entrenabilidad de los modelos con múltiple etiquetado.