Reconocimiento Automático de Voz basado en Modelos Ocultos de Markov

Autor: Dr. Juan Carlos Gómez. Presentación basada en las siguientes **Referencias**:

[1] Rabiner, L. & Juang, B-H.. *Fundamentals of Speech Recognition*, Prentice Hall, N.J., 1993.

[2] Rabiner, L.. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, Vol. 77, No. 2, pp: 257-286, 1989.

[3]Rabiner, L. & Juang, B.. An Introduction to Hidden Markov Models, *IEEE Acoustic, Speech and Signal Processing Magazine*, Vol. 3, No. 1, pp: 4-16, 1986.

Procesos de Markov Discretos

Cadenas de Markov

- Todas las variables aleatorias y eventos están definidos en un **espacio de probabilidad** Ω, con una σ-álgebra F, y una probabilidad P.
- Consideramos sistemas que en un instante determinado están en uno de entre un conjunto de N estados distintos: S_1 , S_2, ..., S_N . A tiempos discretos regularmente espaciados el sistema realiza un cambio de estado (o vuelve al mismo estado = permanece en el estado) de acuerdo a un conjunto de probabilidades asociadas con el estado.

- Denotamos
  - Instantes de tiempo asociados con los cambios de estados $t=1, 2, \ldots$
  - Estado en el tiempo t: q_t
- Una descripción probabilística completa del sistema requeriría especificar el estado actual (en el tiempo t) y todos los estados precedentes.
- Una secuencia de estados $S_1, S_2, \ldots$ es una **Cadena de Markov (MC: Markov Chain)**, si posee la propiedad de Markov

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

- Una cadena de Markov se denomina **homogénea** si

$$P(q_t = S_j \mid q_{t-1} = S_i)$$

no depende del instante *t*. En este caso quedan definidos un conjunto de **probabilidades de transición de estados** a_{ij} de la forma

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

con las (obvias) propiedades

$$a_{ij} \ge 0$$

$$\sum_{j=1}^{N} a_{ij} = 1$$

- Figura 1 representa una MC con 4 estados. La salida del proceso es el conjunto de estados en cada instante de tiempo, donde cada estado corresponde a un evento observable. El proceso de Markov se denomina entonces **observable.**

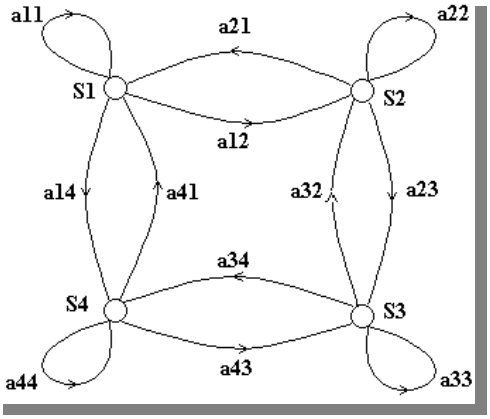

Figura 1: Cadena de Markov con 4 estados.

- **Ejemplo: Modelo del estado del tiempo con una MC de 3 estados.**

Una vez por día (por ejemplo al mediodia) se observa el estado del tiempo y éste puede ser uno de los siguientes

Estado S_1: lluvioso

Estado S_2: nublado

Estado S_3: soleado

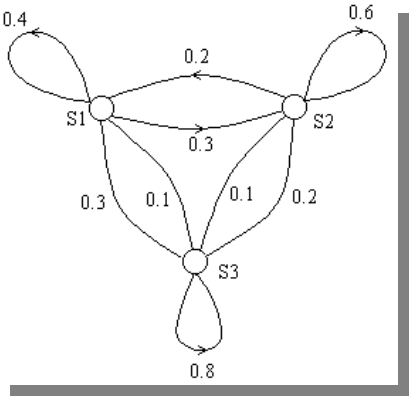

Figura 2: Modelo del estado del tiempo con una MC de tres estados.

Se postula que en el día t, el estado del tiempo puede estar en uno sólo de estos estados y que la matriz A de probabilidades de transición de estados es

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

Suponiendo que el estado en el día 1 ($t=1$) es soleado (estado S_3), queremos calcular la probabilidad de que la sucesión de estados en los próximos 5 días sea:

soleado-soleado-nublado-lluvioso-nublado

Más formalmente, podemos definir una secuencia de observación

$$\mathrm{O} = \{S_3, S_3, S_2, S_1, S_2\}$$

correspondiente a los tiempos $t = 1,2,3,4,5$, respectivamente, y **queremos calcular la probabilidad de la secuencia dado el modelo**. Asumiendo que el modelo posee la propiedad Markoviana, esta probabilidad puede calcularse como

$$\begin{aligned} P(\mathrm{O} \mid \text{modelo}) &= P(S_3, S_3, S_2, S_1, S_2 \mid \text{modelo}) \\ &= P(S_3)P(S_3 \mid S_3)P(S_2 \mid S_3)P(S_1 \mid S_2)P(S_2 \mid S_1) \\ &= \pi_3 a_{33} a_{32} a_{21} a_{12} = 1.(0.8).(0.1).(0.2).(0.3) \\ &= 0.0048 \end{aligned}$$

donde hemos usado la notación

$$\pi_i = P(q_1 = S_i) \qquad 1 \le i \le N$$

para denotar las **probabilidades de estado inicial**.

Otro problema interesante es: **Dado el modelo en un estado conocido, determinar la probabilidad de que permanezca en ese estado durante exactamente *d* días.** Es decir, determinar la probabilidad de la secuencia de observación

$$\mathrm{O} = \underset{1 \quad 2 \quad 3 \qquad d \quad d+1}{\{S_i, S_i, S_i, \cdots, S_i, \; S_j \neq S_i\}}$$

dado el modelo, que es

$$P(\mathrm{O} \mid \text{modelo}, q_1 = S_i) = (a_{ii})^{d-1}(1 - a_{ii}) = p_i(d)$$

La cantidad $p_i(d)$ es la **función de densidad de probabilidad (discreta) de duración *d* en el estado** S_i.

Esta densidad de probabilidad de duración de tipo exponencial es característica de la duración de estado en MCs.

Basados en $p_i(d)$, podemos calcular el **valor esperado del número de observaciones (duración) de un estado, condicionado a comenzar en ese estado**, como

$$\bar{d}_i = \sum_{d=1}^{\infty} d p_i(d)$$

$$= \sum_{d=1}^{\infty} d a_{ii}^{\;d-1}(1 - a_{ii}) = \frac{1}{1 - a_{ii}}$$

Es decir, de acuerdo al modelo tendríamos:

- Número esperado de días soleados consecutivos = $1/0.2 = 5$
- Número esperado de días lluviosos consecutivos = $1/0.6 = 1.67$
- Número esperado de días nublados consecutivos = $1/0.4 = 2.5$

Modelos Ocultos de Markov (HMM: Hidden Markov Models)

- **MC** → Modelos de Markov en los cuales cada estado corresponde a un evento observable → Modelos **demasiado restrictivos** para ser aplicados a muchos problemas de interés.
- **HMMs** → extensión de MCs donde la observación es una función probabilística del estado → proceso doblemente estocástico, donde hay un proceso estocástico subyacente que **no es observable** (es decir, es **oculto**), y sólo puede ser observado a través de otro conjunto de procesos estocásticos que producen la secuencia de observación.

❑ Ejemplo 1: Tirada de Monedas

- **Escenario:** Habitación con una cortina. De un lado de la cortina hay una persona (**oculta**) que ejecuta el experimento de tirar una (o varias) monedas. Sólo nos comunica el resultado del experimento sin decirnos como lo realiza. La secuencia de observación consistirá por ejemplo de

$$O = O_1\, O_2\, O_3 \ldots O_T = C\ S\ C \ldots S$$

donde **C: cara** **S: seca**

- **Problema de Interés:** Construir un **HMM** para explicar la secuencia de observación.
- Debemos **decidir** a qué corresponden los estados del modelo y cuántos estados son necesarios.

- **Modelo con dos estados observables**
  - Una única moneda (posiblemente "cargada") es arrojada.
  - Hay sólo dos estados, y cada estado corresponde a un lado de la moneda.
  - El modelo de Markov es **observable** (es decir es una MC y no un HMM) y para que quede completamente especificado se debe seleccionar el "mejor" valor de, por ejemplo, P(C). **(sólo 1 parámetro desconocido)**

$$O = C\ C\ S\ S\ C\ S\ C \cdots$$
$$S = 1\ 1\ 2\ 2\ 1\ 2\ 1 \cdots$$

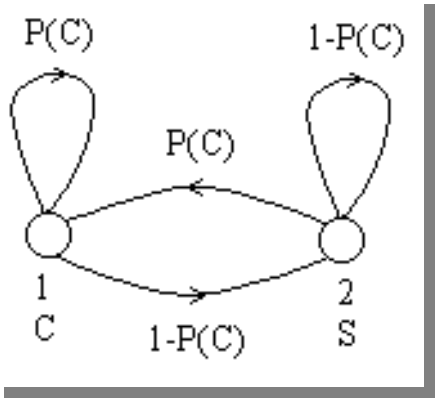

Figura 3: Modelo de Markov con 2 estados observables

- **HMM con dos estados**
  - Hay dos estados en el modelo, donde cada estado corresponde a una moneda (“posiblemente cargada") diferente que es arrojada.
  - Cada estado está caracterizado por una distribución de probabilidades de Caras y Secas, y las transiciones entre estados están caracterizadas por una matriz de transición de estados.
  - El mecanismo físico que explica cómo las transiciones entre estados son realizadas puede ser una tirada independiente de monedas o cualquier otro evento probabilístico.
  - Vemos que aquí hay **4 parámetros desconocidos**, que deben estimarse para que el modelo quede completamente especificado.

$$O = C\,C\,S\,S\,C\,S\,C\,C\cdots$$
$$S = 2\;1\;1\;2\;2\;2\;1\;2\cdots$$

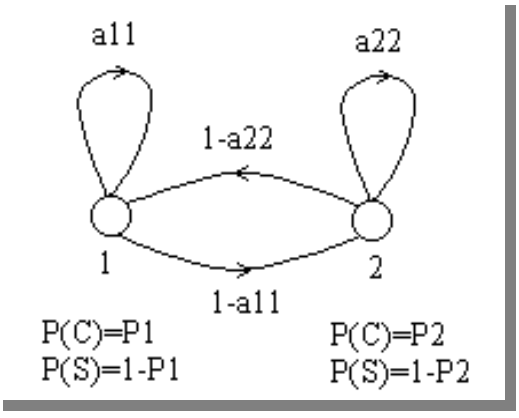

Figura 4: HMM con 2 estados ocultos

- **HMM con tres estados**
  - Hay tres estados en el modelo, donde cada estado corresponde a una moneda (“posiblemente cargada") diferente que es arrojada. La elección del estado está gobernada por algún evento probabilístico.
  - El HMM tiene **9 parámetros desconocidos**, que deben estimarse para que el modelo quede completamente especificado.

	1	2	3
$P(C)$	$P1$	$P2$	$P3$
$P(S)$	$1-P1$	$1-P2$	$1-P3$

$$O = C\,C\,S\,S\,C\,S\,C\,C\cdots$$
$$S = 3\;1\;2\;3\;3\;1\;3\;2\cdots$$

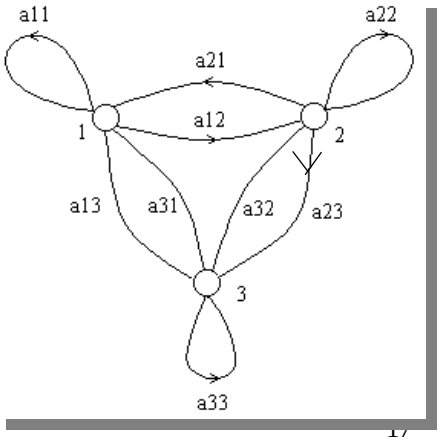

Figura 5: HMM con 3 estados

❑ Ejemplo 2: Bolillero con bolillas de colores

- Hay N bolilleros en una habitación, cada uno conteniendo un gran número de bolillas de colores.
- Asumimos que hay M colores distintos de bolillas.
- Proceso de obtención de las observaciones.
  - Por algún proceso aleatorio, alguien que está en la habitación selecciona un bolillero inicial.
  - De este bolillero es extraída aleatoriamente una bolilla, y su color es registrado como una observación.
  - La bolilla es devuelta al bolillero del cual fue extraída.
  - Un nuevo bolillero es entonces elegido de acuerdo al proceso aleatorio de selección asociado al presente bolillero, y el proceso de selección de la bolilla es repetido.

- Este proceso genera una secuencia finita de observación de colores que quisiéramos modelar como la salida observable de un HMM.
- El HMM más simple corresponde a uno en el cual cada estado corresponde a un bolillero específico, y por el cual están definidas probabilidades de color (bolillas) para cada estado. La elección de los bolilleros está determinada por la matriz de probabilidades de transición de estado del HMM.

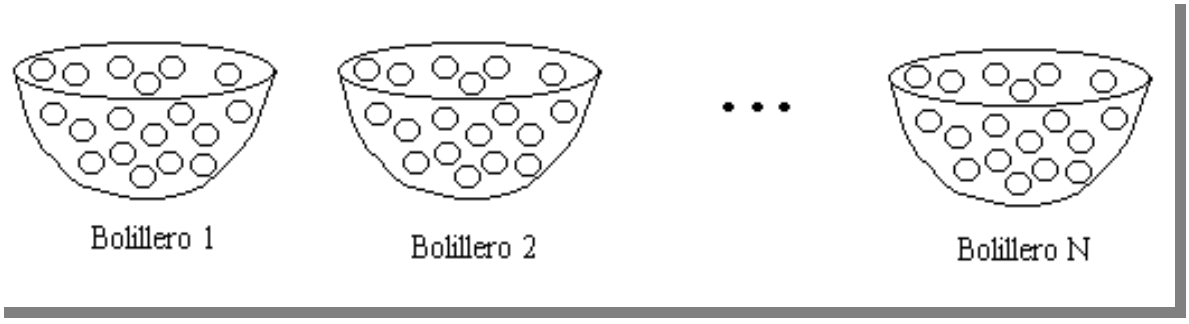

Figura 6: Modelo de Bolilleros con Bolillas con un HMM con N estados

$$\begin{cases} P(\text{rojo}) = b_j(1) \\ P(\text{azul}) = b_j(2) \\ P(\text{verde}) = b_j(3) \\ \vdots \\ P(\text{blanco}) = b_j(M) \end{cases}$$

❑ Elementos de un HMM

- ***N*:** número de estados del modelo.

  Los estados (**ocultos**) tienen (usualmente) asociados algún significado físico. Denotamos los estados individuales como

$$S = \{S_1, S_2, \cdots, S_N\}$$

- *M*: número de símbolos de observación distintos por estado, i.e. dimensión del alfabeto discreto. Denotamos los símbolos individuales como

$$V = \{V_1, V_2, \cdots, V_M\}$$

- Matriz de distribución de probabilidades de transición de estados

$$A = \{a_{ij}\}$$

$$a_{ij} = \mathrm{P}(q_{t+1} = S_j \mid q_t = S_i) \qquad 1 \le i, j \le N$$

- Distribución de probabilidad del símbolo de observación en el estado '*j*'

$$B = \{b_j(k)\}$$

donde

$$b_j(k) = \mathrm{P}(V_k \text{ en } t \mid q_t = S_j)\ ,\ 1 \le j \le N\ ,\ 1 \le k \le M$$

- Distribución de probabilidad del estado inicial $\pi = \{\pi_i\}$ donde

$$\pi_i = \mathrm{P}(q_1 = S_i),\ \ 1 \le i \le N$$

- Para que el HMM quede completamente especificado deben especificarse los parámetros M y N, los símbolos de observación V, y las tres medidas de probabilidad A, B, y π. Al HMM generalmente se lo denota

$$\lambda = (A, B, \pi)$$

- Dado un HMM, éste puede usarse para generar una secuencia de observación

$$O = O_1 O_2 \cdots O_T$$

donde cada observación O_t es uno de los símbolos en V, y T es el número de observaciones en la secuencia.

Generación de una secuencia de observación

1. Seleccionar un estado inicial $q_1 = S_i$ de acuerdo a la distribución de estado inicial π.
2. Fijar $t=1$.
3. Seleccionar $O_t = V_k$ de acuerdo a la distribución de probabilidad de símbolo en el estado S_i, es decir, $b_i(k)$.
4. Transicionar al nuevo estado $q_{t+1} = S_j$ de acuerdo a la distribución de probabilidad de transición del estado S_j, es decir, a_{ij}.
5. Fijar $t=t+1$ y volver al paso 3. si $t < T$, en caso contrario, finalizar.

❑ Los tres problemas básicos en HMMs

Problema 1: Dada una secuencia de observación $O=O_1O_2 \ldots O_T$, y un modelo $\lambda = (A, B, \pi)$, como computar eficientemente la probabilidad de que la secuencia de observación haya sido generada por el modelo $P(O \mid \lambda)$. La solución de este problema permite seleccionar de entre un conjunto de modelos, el que mejor se ajusta a la secuencia de observación. La solución del problema es provista por el **Algoritmo Forward-Backward** (debido a Baum y colaboradores).

Problema 2: Dada la secuencia de observación $O=O_1O_2 \ldots O_T$, y un modelo HMM , $\lambda = (A, B, \pi)$, como seleccionamos la secuencia de estados $Q=q_1 q_2 \ldots q_T$ correspondiente que mejor explique la secuencia de observación. (Determinación del estado oculto del modelo). La solución a este problema es provista por el **Algoritmo de Viterbi.**

Problema 3: Como ajustamos los parámetros del modelo HMM $\lambda = (A, B, \pi)$, de manera de maximizar la probabilidad de que la observación haya sido generada por el modelo $P(O \mid \lambda)$. La solución de este problema es provista por el **Algoritmo de Baum-Welch**, equivalente al **Algoritmo EM (Expectation-Maximization).**

Problema 1: Dada una secuencia de observación $O=O_1O_2 \ldots O_T$, y un modelo $\lambda = (A, B, \pi)$, se desea computar eficientemente la probabilidad de que la secuencia de observación haya sido generada por el modelo $P(O \mid \lambda)$.

La forma más directa de realizar esto sería enumerando todas las posibles secuencias de estados de longitud T (igual al número de observaciones), y calculando la probabilidad conjunta de la secuencia de observación O y la secuencia de estado Q , dado el modelo, es decir $P(O, Q| \lambda)$, para todas las posibles secuencias de estado de longitud T, y luego calcular $P(O \mid \lambda)$ como

$$P(O \mid \lambda) = \sum_{Q} P(O, Q \mid \lambda)$$

Este cálculo directo involucra del orden de $2\ T\ N^{\ T}$ operaciones, siendo N el número de estados del modelo HMM. Este número de operaciones es computacionalmente prohibitivo aún para el caso de N y T pequeños. Por ejemplo para $N=5$ (estados) y $T=100$ (observaciones), resulta $2\ T\ N^{\ T} = 2 \times 100 \times 5^{100} = 10^{\ 72}$.

Claramente es necesario un procedimiento más eficiente. Este procedimiento existe y es el denominado **Algoritmo Forward-Backward**, propuesto por Baum y colaboradores.

❑ Algoritmo Forward-Backward

Se define la **variable forward** como

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

Es decir, la probabilidad de la secuencia de observación parcial O_1 , O_2 , ..., O_t , hasta el instante t, siendo el estado en t igual a S_i, dado el modelo λ. La variable $\alpha_i(t)$ puede calcularse en forma recursiva con el siguiente algoritmo (por inducción).

1. Inicialización

$$\alpha_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N$$

2. Recursión (inducción)

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

$$1 \le j \le N$$

3. Terminación

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

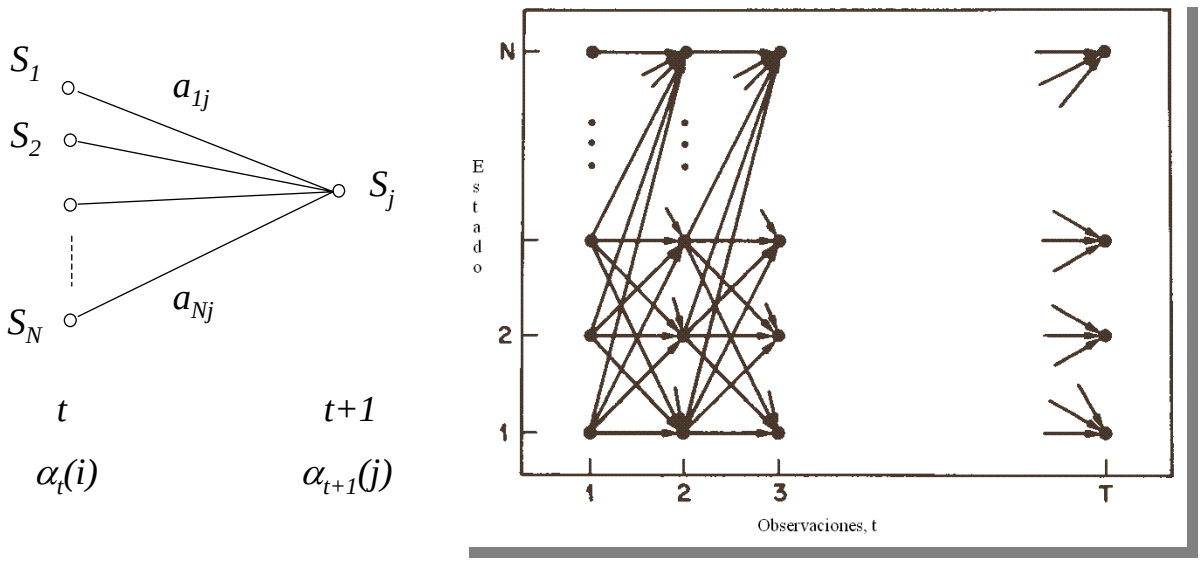

Figura 7: Cálculo de la variable forward

El algoritmo requiere en el orden de $N^2 T$ operaciones, en lugar de las $2 T N^T$ requeridas por el algoritmo directo. Para el caso de $N=5$ (estados) y $T=100$ (observaciones), resultan 2500 operaciones, frente a las 10^{72} operaciones requeridas por el algoritmo directo (69 órdenes de magnitud menos).

En forma similar puede definirse la **variable backward**.

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

Es decir, la probabilidad de la secuencia de observación parcial desde el instante $t+1$ hasta el final (T), dado el estado S_i en t , y el modelo λ.

Si bien esta variable no es necesaria para la resolución del Problema 1, la introducimos aquí ya que es necesaria para la solución de los Problemas 2 y 3.

La variable backward puede calcularse en forma recursiva como sigue.

1. Inicialización

$$\beta_T(i) = 1 \qquad 1 \le i \le N$$

2. Recursión (inducción)

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

$$1 \le i \le N$$

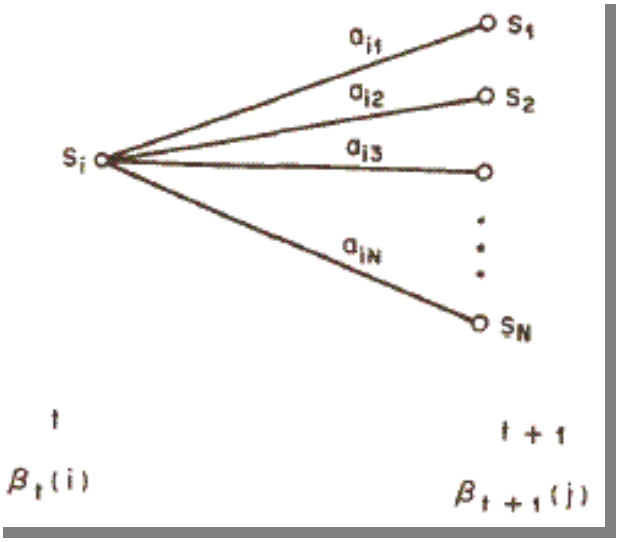

Figura 8: Cálculo de la variable backward (recursión)

Problema 2: Dada la secuencia de observación $O=O_1O_2 \ldots O_T$, y un modelo HMM , $\lambda = (A, B, \pi)$, como seleccionamos la secuencia de estados $Q=q_1 q_2 \ldots q_T$ correspondiente que mejor explique la secuencia de observación. (Determinación del estado oculto del modelo). A diferencia del Problema 1, para el que puede encontrarse una solución exacta, no existe una única solución para el Problema 2, ya que la secuencia óptima de estados va a depender del criterio de optimalidad que se adopte.

Por ejemplo, un posible criterio de optimalidad es seleccionar los estados q_t que individualmente sean los más probables. Este criterio, entonces maximizaría el número esperado de estados individuales correctos. Esto, sin embargo, puede ocasionar problemas con la secuencia de estados resultante, sobre todo en los casos de HMM en los que la matriz de probabilidades de transición tiene elementos nulos (transiciones no permitidas), ya que la secuencia resultante podría no ser ni siquiera una secuencia válida.

Esto es debido al hecho de que la solución determina simplemente el estado más probable en cada instante, sin tener en cuenta la probabilidad de ocurrencia de secuencias de estados.

Una alternativa a esto, que es la que ha sido más ampliamente adoptada, es encontrar la ***única*** mejor secuencia de estados (o camino), dada las observaciones y el modelo, i.e., la secuencia que maximiza P($Q|O, \lambda$), que es equivalente a maximizar P($Q,O|\lambda$). Una técnica formal para encontrar esta única mejor secuencia de estados existe, basada en métodos de programación dinámica, y es el denominado **Algoritmo de Viterbi**.

❑ Algoritmo de Viterbi

Para encontrar la *única* mejor secuencia de estados $Q=q_1 q_2 \ldots q_T$ para una dada secuencia de observación $O=O_1O_2 \ldots O_T$, es necesario definir la cantidad

$$\delta_t(i) = \max_{q_1,q_2,\cdots,q_{t-1}} \mathrm{P}(q_1 q_2 \cdots q_t = S_i, O_1 O_2 \cdots O_t \mid \lambda) \qquad (1)$$

que es la más alta probabilidad a lo largo de un simple camino, que tiene en cuenta las observaciones parciales hasta el tiempo *t* alcanzando el estado S_i en ese tiempo. Por inducción se tiene que

$$\delta_{t+1}(j) = \left[\max_i \ \delta_t(i) a_{ij}\right] b_j(O_{t+1}) \qquad (2)$$

Para recuperar la secuencia de estados es necesario mantener un registro del argumento que maximiza la probabilidad en (2), para cada *t* y para cada *j*. Esto puede hacerse con la variable auxiliar (arreglo) $\psi_t(j)$. El procedimiento completo es el siguiente.

1. Inicialización

$$\delta_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N$$
$$\psi_1(i) = 0$$

2. Recursión (inducción)

$$\delta_t(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right] b_j(O_t) \qquad 2 \le t \le T$$
$$1 \le j \le N$$

$$\psi_t(j) = \operatorname*{argmax}_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right] \qquad 2 \le t \le T$$
$$1 \le j \le N$$

3. Terminación

$$p^* = \max_{1 \le i \le N} \left[\delta_T(i)\right]$$
$$q_T^* = \operatorname*{argmax}_{1 \le i \le N} \left[\delta_T(i)\right]$$

4. Backtracking de la secuencia de estados

$$q_t^* = \psi_{t+1}\left(q_{t+1}^*\right) \qquad t = T-1, T-2, \cdots, 1$$

Problema 3: Como ajustamos los parámetros del modelo HMM $\lambda = (A, B, \pi)$, de manera de maximizar la probabilidad de que la observación haya sido generada por el modelo $P(O \mid \lambda)$. La solución de este problema es provista por el **Algoritmo de Baum-Welch**, equivalente al **Algoritmo EM (Expectation-Maximization)**

Para describir el algoritmo de estimación de Baum-Welsh denotemos con $\xi_t(i, j)$ a la probabilidad de estar en el estado S_i en el tiempo t, y en el estado S_j en el tiempo $t+1$, dado el modelo y la secuencia de observación, es decir:

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

Esta probabilidad puede escribirse en función de las variables back-ward y forward de la siguiente forma

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)}$$

$$= \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}$$

Donde el numerador es $P(q_t = S_i, q_{t+1} = S_j, O \mid \lambda)$, y la normalización por el factor $P(O|\lambda)$ se realiza para que $\xi_t(i, j)$ sea una medida de probabilidad.

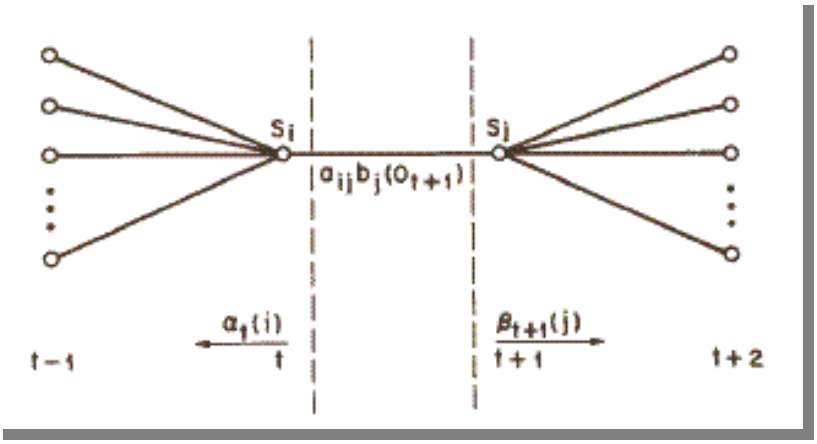

Figura 9: Algoritmo de Baum-Welch.

Definamos la variable

$$\gamma_t(i) = \mathrm{P}(q_t = S_i \mid O, \lambda)$$

Es decir, la probabilidad de estar en el estado S_i en el tiempo t, dada la secuencia de observación y el modelo. Esta probabilidad puede expresarse en función de las variables forward y backward, como

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\mathrm{P}(O \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$

El factor de normalización $\mathrm{P}(O|\lambda)$ hace que $\gamma_t(i)$ sea una medida de probabilidad, es decir se verifica

$$\sum_{i=1}^{N} \gamma_t(i) = 1$$

La variable $\gamma_t(i)$ se puede relacionar con la variable $\xi_t(i,j)$ según

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$$

Sumando $\gamma_t(i)$ sobre t se obtiene el valor esperado del número de veces que el estado S_i es visitado, o lo que es equivalente, el número de transiciones desde el estado S_i.

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{valor esperado nro. de transiciones desde } S_i$$

En forma similar, sumando $\xi_t(i,j)$ sobre t, se obtiene el valor esperado del número de transiciones desde el estado S_i hasta el estado S_j

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \text{valor esperado nro. de transiciones desde } S_i \text{ a } S_j$$

Basándose en estas expresiones se pueden obtener las siguientes fórmulas recursivas de estimación de los parámetros de un HMM.

$$\bar{\pi}_i = \text{valor esperado nro. veces en estado } S_i \text{ en } (t=1) = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\overbrace{\sum_{t=1}^{T-1} \xi_t(i,j)}^{(1)}}{\underbrace{\sum_{t=1}^{T-1} \gamma_t(i)}_{(2)}} \qquad \bar{b}_j(k) = \frac{\overbrace{\sum_{t=1,\, s.t.\, O_t = v_k}^{T} \gamma_t(j)}^{(3)}}{\underbrace{\sum_{t=1}^{T} \gamma_t(j)}_{(4)}}$$

Donde

(1) Valor esperado del nro. de transiciones desde el estado S_i al S_j

(2) Valor esperado del nro. de transiciones desde el estado S_i

(3) Valor esperado del nro. de veces en el estado S_j y observando el símbolo v_k

(4) Valor esperado del nro. de veces en el estado S_j

Partiendo de un modelo inicial $\lambda = (A, B, \pi)$, las expresiones anteriores permiten reestimar el modelo obteniendo un nuevo modelo $\overline{\lambda} = \left(\overline{A}, \overline{B}, \overline{\pi}\right)$, de manera que, o bien

1) el modelo inicial λ define un punto crítico de la función de verosimilitud, en cuyo caso $\overline{\lambda} = \lambda$, o

2) El modelo $\overline{\lambda}$ es más probable que el modelo λ en el sentido que

$$\mathrm{P}\left(O \mid \overline{\lambda}\right) > \mathrm{P}\left(O \mid \lambda\right)$$

Es decir, se ha encontrado un nuevo modelo $\overline{\lambda}$, que es más probable que haya producido la secuencia de observaciones.

Basado en este procedimiento, si iterativamente se usa $\overline{\lambda}$ en lugar de λ y se repite el proceso de re-estimación, se incrementa la probabilidad de que las observaciones hayan sido generadas por el modelo, hasta que se alcanza un punto límite. El valor final de las probabilidades que definen el HMM se denomina **estima de máxima verosimilitud del HMM**.

Cabe aclarar que el procedimiento descripto conduce a máximos locales y que en general la superficie de optimización es muy compleja presentando múltiples máximos locales.

❑ Estimas iniciales de los parámetros del HMM

Las ecuaciones de re-estimación conducen a máximos locales de la función de verosimilitud. Una pregunta que surge es cómo elegir las estimas iniciales de manera que los algoritmos de optimización conduzcan al máximo global de la función de verosimilitud.

Lamentablemente, no existe una respuesta simple o directa a esta pregunta. La experiencia indica que estimas iniciales aleatorias (sujetas a las restricciones estocásticas y de no nulidad) o uniformes de π y A son adecuadas en la mayoría de los casos. Sin embargo, para los parámetros B es indispensable tener una buena estima inicial. Estas estimas iniciales se pueden obtener con diversas técnicas de segmentación de la secuencia de observaciones (ver [2]).

❑ Tipos de HMMs

- **HMM ergódicos:** Cada estado del modelo puede ser alcanzado en un sólo paso desde cualquier otro estado del modelo. También se los denomina HMM totalmente conectados. La matriz de probabilidades de transición de estados es una matriz llena con todos sus elementos positivos.

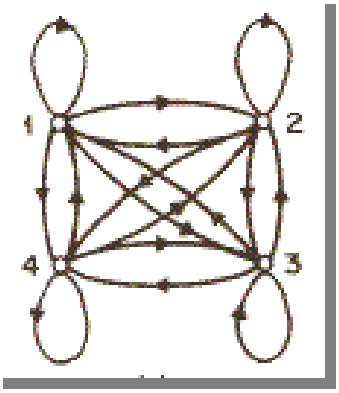

Fig 10: HMM ergódico de 4 estados.

- **HMM Izquierda-a-derecha:** La secuencia de estado asociada con el modelo tiene la propiedad de que a medida que el tiempo avanza el índice indicando el estado se incrementa (o permanece el mismo), i.e., el estado procede de izquierda a derecha. Estos modelos son aptos para modelar señales cuyas propiedades varían con el tiempo, como las señales de voz.

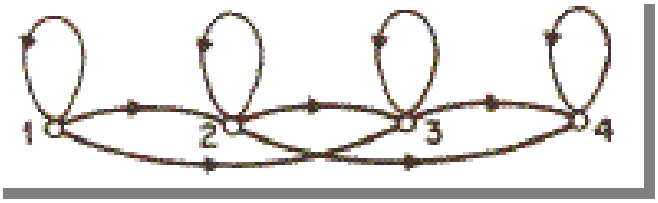

Fig 11: HMM Izquierda-a-Derecha de 4 estados.

Tienen la propiedad que los coeficientes de la matriz de probabilidades de transición de estados verifican

$$a_{ij} = 0 \qquad j < i$$

Es decir, no están permitidas las transiciones a estados cuyos índices son menores que el del estado presente. Además, las probabilidades de estado inicial verifican

$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases}$$

ya que la secuencia de estados debe comenzar en el estado 1 (y finalizar en el estado *N*).

A menudo se suelen imponer restricciones adicionales en las probabilidades de transición de estados para asegurar que no se producen cambios grandes en los índices en las transiciones. Por ejemplo se suele imponer una restricción de la forma

$$a_{ij} = 0 \qquad j > i + \Delta$$

En particular, para el caso de Figura 11 es $\Delta = 2$, es decir, no se permiten saltos de más de dos estados.

Resulta claro que para el último estado en el modelo izquierda-a-derecha se verifica

$$a_{NN} = 1$$
$$a_{Ni} = 0, \qquad i < N$$

❑ Densidades de probabilidad de observación continuas en HMMs

En lo desarrollado hasta el presente se asumió que las observaciones estaban caracterizadas por símbolos discretos dentro de un alfabeto finito, por lo que se podían usar densidades de probabili-dad discretas dentro de cada estado del modelo. El problema con este enfoque es que para la mayoría de las aplicaciones (y en particular para reconocimiento de palabra) las observaciones son señales (vectores) continuos. Si bien se pueden cuantizar usando un libro de códigos (**Vector Quantization**) esto trae aparejada una seria degradación de la performance de los dispositivos de reconocimiento.

Un enfoque muy usado, es representar la función de densidad de probabilidad continua como una combinación lineal de distribuciones Gaussianas (**mezclas Gaussianas**) de la forma

$$b_j(O) = \sum_{m=1}^{M} c_{jm}\ \mathbf{N}(O, \mu_{jm}, U_{jm}) \qquad 1 \le j \le N$$

Donde O es el vector de observación, c_{jm} , $m=1, ..., M$ son los coeficientes de la mezcla de M Gaussianas en el estado j, y $\mathbf{N}(O, \mu_{jm}, U_{jm})$ es una distribución Gaussiana con media μ_{jm} y matriz de covarianza U_{jm} correspondiente a la m-ésima componente de la mezcla en el estado j. Los coeficientes c_{jm} deben satisfacer la restricción estocástica

$$\sum_{m=1}^{M} c_{jm} = 1, \qquad 1 \le j \le N$$

$$c_{jm} \ge 0, \qquad 1 \le j \le N \ , \ 1 \le m \le M$$

de manera que resulte

$$\int_{-\infty}^{\infty} b_j(x)\ dx = 1, \qquad 1 \le j \le N$$

Esta representación puede ser usada para aproximar, con una precisión arbitraria, cualquier función de densidad de probabilidad continua finita.

Se han desarrollado fórmulas de reestimación para los parámetros μ_{jm} , U_{jm} y c_{jm} de las mezclas Gaussianas (ver [2]).

❑ Reconocimiento de palabra aislada usando HMM

Para cada palabra de un vocabulario de V palabras se quiere diseñar un modelo HMM con N estados. Para cada palabra en el vocabulario se tiene un conjunto de entrenamiento con K instancias de la palabra hablada por uno o más locutores. Cada ocurrencia de la palabra constituye una secuencia de observación. Las observaciones son alguna representación espectral (cepstral) o temporal de la señal de voz. En general, la señal correspondiente a cada palabra se representa con una secuencia temporal de vectores espectrales codificados. Se asume que la codificación se realiza con un libro de código con M palabras código. Cada observación resulta entonces igual al índice del vector en el libro de código que está **más cerca** del vector espectral correspondiente a la señal de voz.

El algoritmo de reconocimiento puede resumirse como:

1. Para cada palabra v en el vocabulario V se debe construir un HMM λ_v , es decir se deben estimar los parámetros del modelo (A_v, B_v, π_v) que optimizan la probabilidad del conjunto de vectores de entrenamiento asociados a la v-ésima palabra.
2. Para cada palabra desconocida que se quiere reconocer, se debe obtener la secuencia de vectores de observación O. Luego calcular la probabilidades de que esa secuencia haya sido generada por cada uno de los modelos posibles $P(O|\lambda_v)$, con $1 \le v \le V$ y luego seleccionar la palabra cuyo modelo tenga la más alta probabilidad, *i.e.*

$$v^* = \underset{1 \le v \le V}{\operatorname{argmax}} \left[P(O \mid \lambda_v) \right]$$

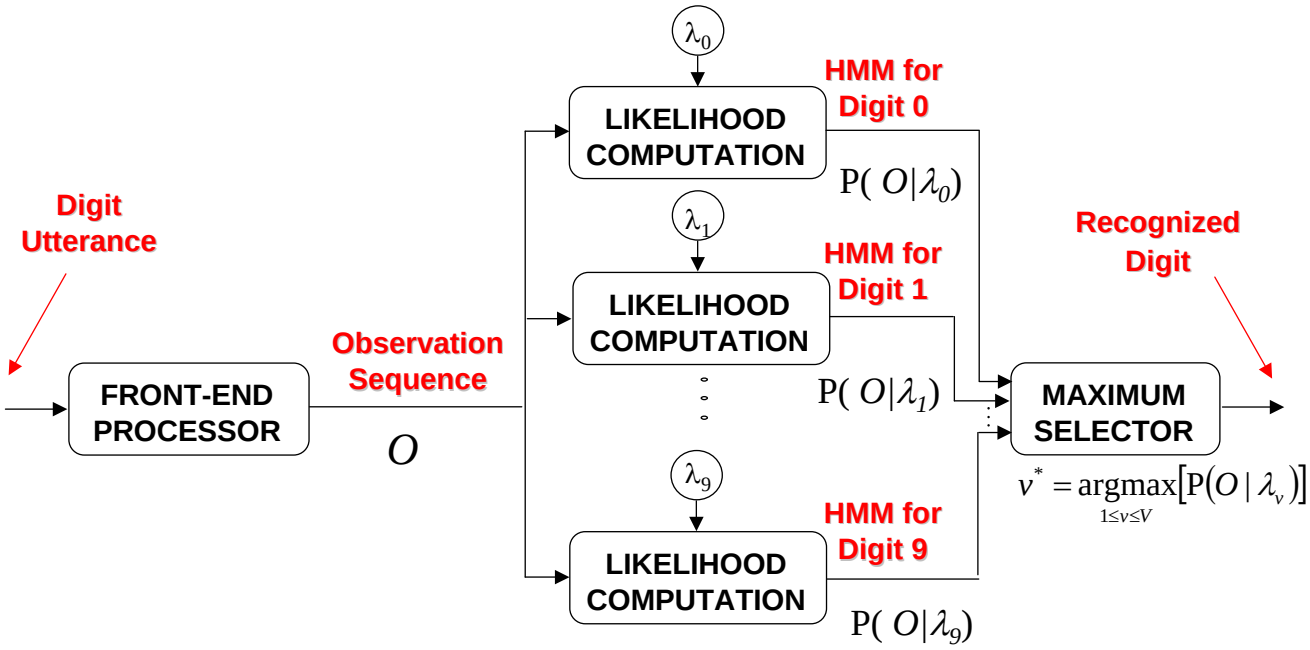

Fig. 12: Diagrama de bloques de un reconocedor de dígitos aislados.